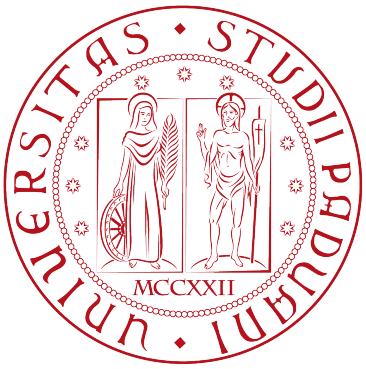


Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Corso di laurea in ingegneria dell'informazione

Stima della salienza per immagini omnidirezionali

Relatore:
Dott.ssa Sara Baldoni

Co-relatore:
Prof.ssa Federica Battisti

Candidate:
Ortisa Poci
2000253

Anno accademico 2022-2023

Sommario

Negli ultimi anni la diffusione delle immagini omni-direzionali ha acquisito sempre più rilevanza. A differenza del caso bidimensionale, questi contenuti emergono poiché offrono all'utente un'esperienza immersiva. Le immagini omni-direzionali, infatti, possono essere fruite utilizzando dei visori che consentono all'utente di essere circondato dal contenuto multimediale e di esplorarlo liberamente. Per questo motivo diventa di particolare interesse comprendere come cambia il meccanismo di attenzione e visione di questi contenuti rispetto al caso bidimensionale. Lo studio di questo fenomeno consente di identificare le aree salienti dell'immagine omni-direzionale al fine di costruire una mappa di salienza.

In letteratura sono stati proposti diversi approcci per la stima delle mappe di salienza. Molti di essi si basano sull'estrazione e combinazione di varie caratteristiche delle immagini (*feature*) per deteriminare la mappa finale. Tuttavia, un aspetto poco investigato è il contributo delle singole feature e delle loro possibili combinazioni sulla stima della salienza.

Per questo motivo questa tesi ha lo scopo di analizzare il contributo di un insieme di *feature* comunemente impiegate nello stato dell'arte. Più nello specifico, è stato valutato il loro impatto quando applicate singolarmente e quando vengono combinate. I risultati mostrano come una buona stima della salienza sia in primis condizionata dalla tipologia di combinazione piuttosto che dal numero di *feature* impiegate. In particolare, è vantaggioso includere un insieme di caratteristiche non ridondanti sia relative al contenuto che relative alle caratteristiche intrinseche delle immagini, quali il colore, l'intensità o il contrasto.

Alla mia famiglia che mi ha sempre sostenuta,

alla mia relatrice e correlatrice per avermi dato la possibilità di svolgere il mio lavoro di tesi.

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Introduzione

Negli ultimi anni stiamo assistendo ad una diffusione sempre maggiore delle immagini a $360°$, caratterizzate da un grado più elevato di immersività e di realismo. A differenza del caso 2D caratterizzato da un campo visivo (Field Of View (FOV)) limitato, i contenuti omni-direzionali si estendono in un intervallo di $360°$ in orrizontale e $180°$ in verticale, offrendo agli utenti la sensazione di essere presenti fisicamente nella scena.

Tutto questo è stato possibile grazie allo sviluppo di nuove tecnologie multimediali di acquisizione e di *rendering* facilmente accessibili. Tra queste, l'Head-Mounted Display (HMD) ha avuto un ruolo significativo per migliorare la qualità percepita dell'esperienza di visualizzazione, permettendo allo spettatore di indirizzare liberamente l'area di visualizzazione sul contenuto desiderato tramite il movimento della testa, proprio come accade nel mondo reale. Nello specifico, il visualizzatore attraverso il movimento della testa dirige l'attenzione verso una specifica area (*viewport*) della scena a $360°$ mentre con il movimento degli occhi si focalizza sugli elementi desiderati all'interno dell'area selezionata.

La diversa natura di questi nuovi contenuti multimediali comporta però alcuni cambiamenti, motivo per cui ci sono vari aspetti che devono essere indagati. Nello specifico, diventa di particolare interesse studiare come gli utenti percepiscono ed esplorano la scena durante la visualizzazione delle immagini a $360°$. La stima della salienza visiva rappresenta un contributo significativo per questo scopo.

I modelli di salienza visiva mirano ad identificare le aree di un immagine che risultano essere più rilevanti per un osservatore umano durante la visualizzazione.

Il sistema visivo umano (Human Visual System (HVS)), infatti, affronta il problema dell'incapacità di elaborare in parallelo la grande quantità di informazioni con cui ci confrontiamo portando il focus dell'attenzione visiva su parti cospicue degli stimoli visivi. Questo com-

portamento è adottato in particolar modo quando un soggetto si relaziona con i contenuti multimediali, siano essi 2D o a $360°$.

Attualmente, in letteratura si parla di due possibili processi che conducono alla salienza visiva: processo *top-down* quando è guidato dal contenuto o processo *bottom-up* quando le regioni sono salienti a causa delle caratteristiche *low-level* delle immagini (quali il colore, il contrasto o l'intensità).

Sulla base di queste teorie, esiste un'ampia varietà di modelli che mirano ad imitare il processo cognitivo dell'attenzione visiva per le immagini omni-direzionali. Inizialmente, sono stati proposti dei modelli basati sull'estrazione e la combinazione delle caratteristiche delle immagini (*feature*), mentre solo negli ultimi anni, con il recente sviluppo del *Deep Learning*, la stima della salienza ha ottenuto notevoli miglioramenti sia grazie ad architetture specifiche che a grandi set di dati per l'addestramento. Come risultato si ottengono le mappe di salienza, ovvero immagini dello stesso formato dell'immagine di ingresso in cui per ciascun pixel viene indicata la rilevanza per l'utente.

Nonostante l'elevato numero di modelli che si basano sulla combinazione di diverse caratteristiche dell'immagine, manca uno studio che possa valutare l'incidenza delle diverse *feature* e delle loro combinazioni alla mappa finale. Questo aspetto deve essere tenuto in considerazione poichè gioca un ruolo importante a favore di un calcolo più accurato della salienza.

Questo elaborato, pertanto, ha lo scopo di eseguire un'analisi approfondita di quali caratteristiche contribuiscono maggiormente alla stima della salienza, esaminando il contributo di ciascuna di esse e valutando l'impatto della loro combinazione.

Nello specifico la tesi è strutturata come segue:

- Capitolo 2: trattazione dello stato dell'arte dei modelli utilizzati per la stima della salienza, dei dataset presenti in letteratura e delle metriche di valutazione comunemente adottate;
- Capitolo 3: presentazione del metodo sviluppato per la stima della salienza;
- Capitolo 4: analisi dei risultati ottenuti;
- Capitolo 5: conclusione.

Capitolo 2

Stato dell'arte

2.1 Nozioni introduttive

Un'immagine omni-direzionale è un nuova tipologia di contenuto multimediale che ritrae una scena in tutte le direzioni attorno ad un unico punto di vista, fornendo agli spettatori la sensazione di essere fisicamente presenti nella scena visualizzata. Più concretamente, il contenuto rappresentato è ottenuto dalla proiezione dell'informazione visuale circostante su una sfera di raggio unitario con al centro lo spettatore. Queste immagini, pertanto, vengono spesso chiamate anche immagini a $360°$ o immagini sferiche.

A causa della loro natura, acquisire, trasmettere e visualizzare video a $360°$ non è un compito facile se confrontato con il caso 2D.

Innanzitutto la creazione di questi contenuti spesso richiede l'integrazione di immagini catturate da più fotocamere che vanno a coprire tutto il campo visivo. Per formare il contenuto sferico sono utilizzati degli algoritmi per allineare le aree sovrapposte tra i campi visivi delle fotocamere e adattare le immagini catturate alla superficie della sfera.

Normalmente il contenuto sferico viene proiettato in una rappresentazione planare in modo da facilitare s*torage* e trasmissione e per essere elaborato dagli strumenti di elaborazione multimediale 2D attualmente disponibili.

La proiezione equirettangolare (Equirectangular Projection (ERP)) e la proiezione cubica (Cube-Map Projection (CMP)) sono le proiezioni sfera-piano più utilizzate.

La proiezione equirettangolare consiste nel campionamento ad angoli costanti della sfera. Si noti come il risultato però consista in una versione distorta dell'immagine sferica (Figura 2.1), problema che deve essere preso in considerazione negli algoritmi di elaborazione delle immagini. La proiezione equirettangolare, infatti, presenta diversi artefatti come distorsioni

nella curvatura delle linee, variazione delle dimensioni degli elementi presenti nel contenuto dell'immagine oppure la rottura della continuità presente nel contenuto sferico.

La proiezione cubica è un'alternativa introdotta per evitare il problema degli artefatti portata dalla proiezione equirettangolare. Come suggerita dal nome, questa consiste nella proiezione dell'immagine sferica su un cubo centrato attorno al centro della sfera. Una volta che il cubo viene aperto otteniamo la proiezione desiderata nel piano 2D (Figura 2.1). Tuttavia anche questa proiezione presenta alcuni antefatti come la presenza di angoli finti o la presenza di oggetti ripetuti su più facce.

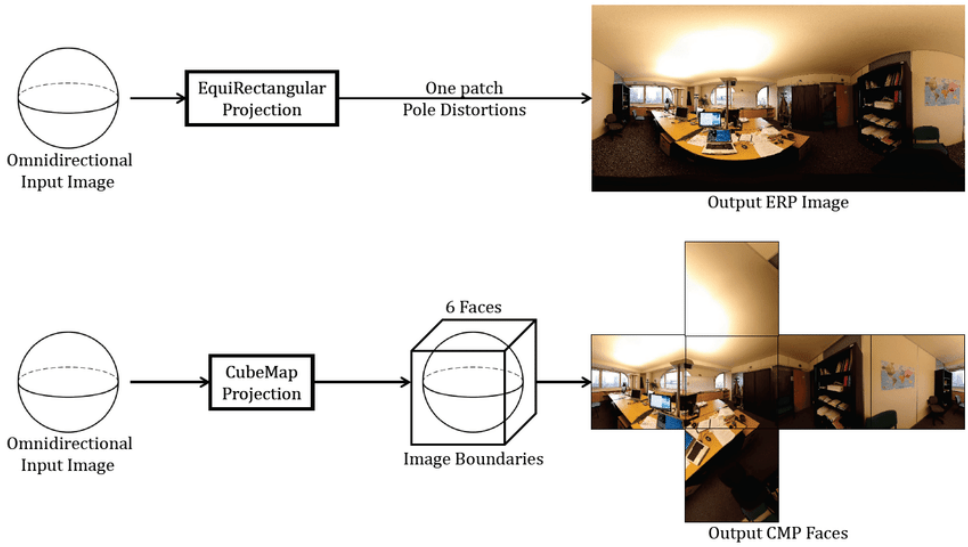
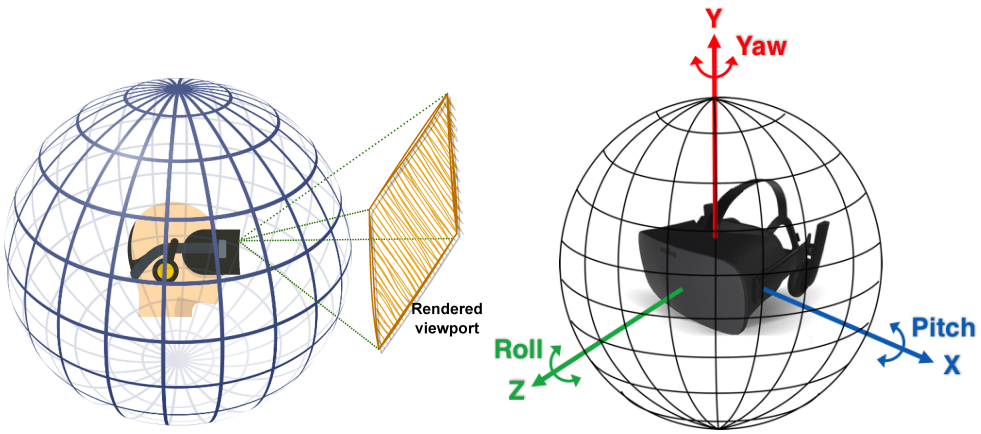


Figura 2.1: Esempio proiezioni ERP e CMP [1].

Per ultimo anche nella fase di *rendering* sono necessarie alcune operazioni per gestire la modalità di visualizzazione di questi contenuti. Mentre le immagini 2D possono essere viste nella loro completezza, ciò non accade con le immagini sferiche per cui lo spettatore si focalizza solo sulla parte del contenuto desiderata (*viewport*) (Figura 2.2a). Infatti la visualizzazione è comunemente effettuata mediante l'utilizzo degli HMD, motivo per cui è richiesto un certo grado di interazione da parte dell'utente durante l'esplorazione del contenuto a $360°$. Nello specifico tramite gli HMD sono possibili 3 gradi di libertà corrispondenti alle tre rotazioni della testa, come mostrato in Figura 2.2b.



(a) Visione della *viewport* [4] (b) Possibili movimenti della testa [5]

Figura 2.2: Visualizzazione dei contenutiti omnidirezionali tramite l'uso degli HMD.

.

Di fronte ad una immagine omni-direzionale che porta una grande quantità di informazioni, l'uomo di natura è portato a focalizzarsi su particolari elementi del contenuto tramite l'attenzione selettiva. L'apparato visivo umano (HVS) consta principalmente di due organi principali: l'occhio e il cervello. Quest'ultimo, in particolare, gioca un ruolo importante nella rielaborazione ed interpretazione delle informazioni provenienti dall'ambiente esterno [6]. Inoltre è proprio tra le operazioni svolte dal cervello di fronte ad uno stimolo visivo che compare il meccanismo di attenzione, largamente studiato da diverse discipline dalla psicologia [7] fino alla *Computer-Vision* [8].

Fondamentalmente, di fronte ad un'immagine i segnali visivi salienti attirano la nostra attenzione.

L'attenzione visiva generalmente subisce una grande suddivisione: si può parlare di attenzione "*bottom-up*" o "*top-down*". Nel primo caso, particolari elementi dell'immagine catturano l'attenzione a causa delle loro proprietà intrinseche *(low-level)* come il colore, il contrasto o l'intensità. Per esemplificare questo concetto si può considerare il caso di un soggetto che di fronte ad un contenuto multimendiale viene attratto all'improvviso da un elemento che si distingue per il colore come un oggetto giallo in mezzo ad altri neri. D'altra parte l'attenzione *top-down* si riferisce alla focalizzazione volontaria dell'attenzione verso particolari oggetti, caratteristiche o aree dell'immagine visualizzata.

Il rilevamento della salienza per le immagini omni-direzionali è il processo che mira a stimare l'attenzione visiva riguardo a questi contenuti multimediali andando a determinare gli oggetti o le regioni che tendono ad attirare l'attenzione dell'uomo.

Ad oggi il modo in cui un utente visualizza la scena può essere facilmente determinato mediante l'utilizzo di nuovi dispositivi dedicati come *head o eye tracker* con lo scopo di rilevare

il movimento della testa e degli occhi.

I risultati ottenuti da questi studi sono presentati come delle mappe di salienza, che includono diversi tipi di rappresentazione.

Una mappa di salienza è un'immagine in scala di grigi in cui la salienza di ciascun pixel è direttamente proporzionale al suo valore. Le varie tipologie di mappe di salienza includono mappe di fissazioni (Eye-Movement (EM)), ovvero mappe che mostrano la probabilità di fissazione in particolari punti della scena visiva; mappe *scan-path* in cui sono riportati informazioni sul percorso di scansione degli oggetti; mappe in cui sono riportati solo le informazioni del movimento della testa (Head-Movement (HM)) o la combinazione del movimento della testa e degli occhi (HM+ EM).

Di seguito viene riportato un esempio di immagine omni-direzionale con la relativa mappa di salienza (Figura 2.3).

Esistono diverse applicazioni che richiedono conoscenze sulla stima della salienza in cui sono incluse il marketing, la compressione delle immagini, il riconoscimento degli oggetti e la robotica.

(a) Immagine omni-direzionale (b) Mappa di salienza HM

Figura 2.3: Esempio di immagine omni-direzionale con la mappa di salienza in cui è stata applicata la *colormap* corrispondente presi dal dataset *"Salient360!"* [2, 3].

.

2.2 Stato dell'arte

Nel corso degli anni si è prestata particolare attenzione allo studio della salienza per le immagini omnidirezionali [9, 10].

Sebbene, come menzionato in precedenza, queste informazioni possano essere ottenute direttamente attraverso test sperimentali in cui vengono utilizzati sistemi di *eye-tracking* e software specifici in grado di rilevare il movimento della testa, questo approccio porta a diversi svantaggi come un accrescimento dei costi e sconvenienza in termini di tempo. Da questo deriva la necessità di progettare modelli di stima della salienza.

In questo capitolo viene presentata una breve rassegna dello stato dell'arte riguardo la stima della salienza di immagini omnidirezionali. Di seguito la trattazione si concentra sui modelli di predizione basati sui movimenti della testa, HM, e degli occhi, EM, che portano alla definizione di una HM/ EM *saliency map*, mentre verranno omesse altre tipologie di modellazione dell'attenzione visiva come gli *scanpath prediction* [11, 12] o *HM/EM prediction* [9, 13].

I modelli presenti in letteratura sono generalmente suddivisi in base alla loro architettura: si parla comunemente di modello *"Hand-crafted"* (spesso chiamato anche modello euristico o classico) e modello *"Data-driven"*.

2.2.1 Modelli *"Hand-crafted"*

I modelli *"Hand-crafted"* si basano sull'estrazione di caratteristiche dell'immagine (*feature*) per stimare le mappe di salienza. In questo approccio si può identificare uno schema comune costituito da tre fasi:

- fase di pre-elaborazione: si effettuano alcune operazioni preliminari prima del calcolo della salienza come, ad esempio, il cambiamento dello spazio dei colori o la proiezione delle immagini dal piano sferico su un piano 2D;
- calcolo della salienza: viene fatto il calcolo delle mappe di salienza;
- post-elaborazione: sono eseguite eventuali operazioni finali come l'integrazione di un *bias* equatoriale alla mappa di salienza o l'operazione di normalizzazione.

I modelli euristici subiscono ulteriori categorizzazioni a seconda delle operazioni effettuate nelle diverse fasi.

Se consideriamo il primo dei tre step sono possibili più proiezioni. Tra queste troviamo la proiezione equirettangolare (ERP), la proiezione cubica (CMP) o la proiezione piramidale fino ad arrivare alla combinazione di più proiezioni. Molti ricercatori hanno anche scelto di lavorare direttamente sul piano sferico per ovviare ad acuni problemi, quali distorisioni o antefatti, che spesso sono derivati dall'operazione di passaggio sul piano 2D. In diversi modelli, invece, la proiezione dell'intera immagine lascia spazio alla proiezione gnomonica sfera-piano per estrarre le *viewport* su cui eseguire le operazioni successive. La mappa di salienza finale viene quindi determinata prendendo in considerazione il contributo di ogni *viewport*. Infine negli anni si è anche considerato di suddividere l'immagine equirettangolare in *superpixel* a diversi livelli per effettuare il calcolo della salienza.

Dal momento che ad oggi sono stati trattati diffusamente i modelli di stima della salienza per le immagini 2D [14], in molte occasioni questi ultimi sono stati usati come base per la progettazione nel caso omni-direzionale. Un primo esempio è la Fused Saliency Maps (FSM) [15], che consiste nell'applicazione diretta della Saliency in Context (SALICON) [16] progettata per le immagini 2D. Adottando la stessa idea di estendere modelli 2D già noti al caso omnidirezionale, in [17] vengono proposti i modelli BMS360 e GBVS360 basati rispettivamente sui modelli Boolean Map based Saliency (BMS) [18] e Graph Based Visual Saliency (GBVS) [19], diffusamente addottati nel caso 2D. Nel primo modello si ottengono mappe booleane richiamando alcuni principi di Gestalt secondo i quali la distinzione tra il primo piano e lo sfondo sia basata anche sull'ambiente circostante. In GBVS360 viene applicato il modello 2D GBVS sulle *viewport* estratte, ottenendo delle mappe di salienza che verranno poi combinate per ottenere la mappa finale.

In letteratura sono stati proposti diversi modelli che si basano sull'estrazione di *feature*. In termini di modellazione i primi lavori sulla previsione della salienza risalgono agli anni Ottanta in cui Treisman *et al.* [20], sulla base delle prime teorie cognitive, presentarono una nuova teoria dell'integrazione delle caratteristiche per l'attenzione visiva. Questa teoria sostiene che le caratteristiche delle immagini vengono rilevate in anticipo, mentre gli oggetti vengono specificati separatamente e solo in una fase successiva, richiedendo un'attenzione localizzata. Da questo Itti *et al.* [8] presentarono il primo modello computazionale di salienza e successivamente si susseguirono molti modelli che si basarono sulla stessa idea di estrarre le *features* dalle immagini e combinarle insieme.

Attualmente esistono una grande varietà di approcci per il calcolo della salienza: alcuni modelli, ad esempio, lavorano anche nel dominio della frequenza per identificare le regioni con variazioni significative [21], altri sono basati sulla teoria dell'informazione per misurare l'unicità delle aree dell'immagine [22], altri ancora sono ispirati alla biologia [23].

In letteratura è comune suddividere i modelli in due possibili categorie sulla base della teoria della salienza visiva: approccio "*bottom-up*" e approccio "*top-down*".

Nel primo, per calcolare le mappe di salienza, sono presi in causa le proprietà *low-level* dell'immagine quali il colore, il contrasto o l'intensità. La mappa di salienza è quindi costituita dalla combinazione pesata di un insieme di mappe, una per ogni *feature*. I modelli "top-down" invece fanno riferimento a caratteristiche *high-level* come il contesto, il significato o la semantica.

Altri modelli combinano le due tipologie di *feature* per produrre mappe di salienza più accurate. Ne è un esempio il modello proposto in [24] che stima la salienza selezionando la

tonalità, la saturazione e la GBVS come *low-level feature*, mentre come *high-level feature* sono considerate il rilevamento della pelle e del viso.

Questo modello rientra inoltre anche tra quelli nei quali viene eseguito un'operazione di pesaggio con l'obiettivo di risaltare maggiormente l'area vicino all'equatore. È stato dimostrato, infatti, che questa zona viene preferita durante l'esprolazione dei contenuti omnidirezionali [2].

2.2.2 Modelli "Data-driven"

Il successo del *Deep Learning* ha portato alla nascita di nuovi modelli di salienza molto più accurati, tanto che quelli precedenti basati sull'approccio euristico sono diventati meno preferibili.

Tuttavia, lo svantaggio di questo approccio riguarda il fatto che per offrire prestazioni elevate sono necessarie una grande potenza di calcolo e un set di dati di addestramento di grandi dimensioni, requisiti spesso mancanti. Per questo motivo questi modelli vengono addestrati in un primo momento su set di dati di immagini di grandi dimensioni prima di essere ottimizzati su set di dati di piccola scala, consentendo il riutilizzo delle informazioni acquisite nelle reti neurali. Infatti il set di dati di piccole dimensioni è specifico per per il task che si vuole svolgere e per il quale non ci sono tante immagini, mentre quello di grandi dimensioni ha un contenuto diverso.

Per ovviare a questo problema recentemente è stato reso disponibile il dataset SALICON [16] che costituisce ad oggi il più grande set di dati per l'addestramento per la stima della salienza.

Grazie al progresso dell'hardware e dell'accesso ai dati, sono state sviluppate e utilizzate nella pratica molte soluzioni di *Deep Learning* per la salienza visiva. Il vantaggio rispetto ai modelli classici riguarda le capacità di catturare i segnali complessi che attirano automaticamente lo sguardo e di estrarre caratteristiche *high-level* nel contenuto omnidirezionale. Tuttavia, alcuni modelli classici hanno fatto fronte a questa mancanza incorporando esplicitamente rilevatori di oggetti e/o rilevatori di volti, motivo per cui nella pratica ci sono casi in cui i modelli classici di salienza prevalgono sui modelli *data-driven*.

Alcuni importanti modelli di salienza basati sui dati includono: eDN, il modello proposto da Vig *et al.* [25] che costituisce il primo tentativo di utilizzare le reti neurali convoluzionali (Convolutional Neural Network (CNN)) per prevedere la salienza dell'immagine; DeepGaze I [26], un modello basato sulla CNN addestrata su un dataset di immagini con corrispondenti dati di *eye-tracking*; DeepFix [27], che è la prima applicazione di reti neurali completamente

convoluzionali (Fully Convolutional Neural Networks (FCNN)) per la previsione della salienza; SalGan [28] un modello che vede l'utilizzo della rete generativa avversaria (Generative Adversarial Networks (GAN)); SalNet360 [29] che dopo la poiezione CMP ancora una volta vengono applicate le CNN.

A seguito di questa trattazione dello stato dell'arte dei modelli di salienza si evincono delle debolezze sia per modelli *hand-crafted* che per i modelli *data-driven*. Tra queste, uno degli aspetti negativi per i modelli *deep learning* risiede nella loro poca spiegabilità, mentre per i modelli *hand-crafted* è poco chiaro quale sia il contributo di ogni *feature* alla salienza finale. Di conseguenza, nasce la necessità di tenere conto di questi apsetti. Per questo motivo questo lavoro di tesi ha come obiettivo studiare come le varie *feature* incidano sulla mappa di salienza, andando ad analizzare il contributo delle diverse *feature* quando prese singolarmente e l'effetto delle loro combinazioni.

2.2.3 Dataset

Per definire un modello di salienza affidabile è fondamentale avere accesso a un set di dati che includono immagini a 360° affiancati da tutte le informazioni che riguardano i punti su cui si concentra l'attenzione visiva per mezzo del movimento degli occhi EM e/o il movimento della testa HM. Questi ultimi in particolare sono necessari dal momento che, diversamente dal caso 2D, i contenuti omnidirezionali sono visualizzati in un intervallo che si estende a 180° in verticale e 360° in orrizontale, causando in tal modo un movimento della testa per selezionare la porzione della scena desiderata (*viewport*). In seguito, all'interno di questa porzione dello spazio, lo sguardo si concentra sulle aree salienti.

Questi dati possono essere direttamente raccolti tramite il contrubuto di numerosi utenti che visualizzano i contenuti omni-direzionali indossando l'HMD. Nello specifico, mentre i dati HM sono ottenuti direttamente tramite un kit di sviluppo software (Software Development Kit (SDK)) che elabora i dati forniti dall'HMD stesso, per acquisire i dati EM è necessario incorporare l'HMD con un *eye-tracker* tramite il quale vengono tracciati i riflessi della pupilla e della cornea.

Come suggerito da Xu *et al.* [10], ci sono alcuni aspetti da valutare quando si prepara un dataset per la stima della salienza a 360°:

- Coerenza tra soggetti: il comportamento di esplorazione tra i diversi soggetti non dovrebbe essere tanto diverso. In tal senso vengono realizzate mappe di salienza HM/EM

proiettate sul piano 2D con lo scopo di studiarne la loro somiglianza;

- Bias dell'equatore e del centro: l'uomo durante la visualizzazione di un contenuto 2D e all'interno di una *viewport* è portato a dirigere l'attenzione al centro della scena. In modo analogo si è visto che nel caso omni-direzionale, i soggetti preferiscono guardare la porzione equatoriale del contenuto. In altre parole esiste un *bias* statistico che vale sia per EM che per HM.

- Impatto del contenuto: come nel caso bidimensionale, anche in questo caso è emerso che il contenuto svolge un ruolo importante sull'attenzione visiva.

- Relazione tra HM e EM: le distribuzioni di HM e EM possono differire in modo significativo. Sono stati portati avanti una serie di studi circa la relazione tra le due distribuzioni [2], rivelando che esse sono simili ma non uguali. Per questo motivo in molti modelli di stima di salienza sono state separate le mappe di HM e EM, per cui alcuni autori hanno trattato solo una tipologia di mappa (studiando solo HM o solo EM) oppure entrambe applicando lo stesso modello o modelli diversi.

Di seguito sono riportati degli esempi di dataset utilizzati in letteratura.

Tabella 2.1: Esempi di dataset proposti.

Dataset	Dimensione dataset	Partecipanti	Risoluzione immagini
De Abreu *et al.* [15]	21	32	4096x2048
Sitzmann *et al.* [30]	22	169	1920x1080
Salient360! [31]	85	63	5376×2688 fino a 18332×9166
MIT300 [32]	300	39	variabile
CAT2000 [33]	4000	24	1920x1080
SALICON [34]	20000	16	640x480
Toronto [35]	120	20	681x511

2.2.4 Metriche di valutazione

Oggigiorno esiste un'ampia varietà di modelli che mirano a stimare le aree salienti dei contenuti omni-direzionali, motivo per cui nasce la necessità di determinare delle metriche per poterli valutare in modo equo e per poterli confrontare.

Disponendo di questo requisito, sono stati creati dei veri e propri *benchmark* in cui vengono riproposti i modelli dello stato dell'arte con tutte le informazioni che riguardano le loro prestazioni [3, 36].

Esistono tre tipologie di metriche a seconda della loro natura: “basate sul valore”, per cui vengono confrontate le ampiezze di ciascun pixel della mappa di salienza stimata e la corrispondente ground-truth; “basate sulla posizione”, che analizzano la corrispondenza spaziale tra le due mappe; “basate sulla distribuzione”, che, come suggerisce il nome, calcola il grado si somiglianza tra le mappe sotto un punto di vista statistico.

Alcune delle metriche utilizzate per valutare i modelli sono:

- Area Under the Curve (AUC): in primis viene estratto uno stesso numero casuale di pixel dalle mappe. Questi, prendendo in considerazione diversi valori di soglia, vengono classificati come “di fissazione” o “di sfondo” . Tra i primi si distinguono i veri positivi se sono effettivamente salienti e i falsi positivi altrimenti. Infine, viene disegnata la curva Receiver Operating Characteristic (ROC) e calcolata l’area al di sotto di questa (AUC).

- Normalized Scanpath Saliency (NSS): calcola i valori normalizzati della mappa di salienza nelle posizioni di fissazione. Il valore restituito è una media fatta su tutte le posizioni. Le due mappe sono simili se il valore è piccolo (una piccola varianza o una piccola differenza tra il valore della mappa e la media), altrimenti il modello non è molto predittivo. Siano ρ una generica posizione nel’immagine, N il numero totale di pixel, $S(\rho)$ la mappa valutata nella posizione specificata e σ_S e μ_S rispettivamente la varianza e la media della mappa di slienza, allora:

$$NSS = \frac{1}{N} \times \sum_{\rho=1}^{N} NSS(\rho), \qquad NSS(\rho) = \frac{S(\rho) - \mu_S}{\sigma_S}. \tag{2.1}$$

- Earth Mover’s Distance (EMD): misura la distanza tra due distribuzioni di probabilità in una regione. Nello specifico, questa è una metrica che quantifica la quantità minima di “lavoro” necessaria per trasformare la distribuzione di probabilità delle mappe di salienza in quella delle fissazioni dell’occhio umano (data in partenza).

- CC: siano S ed F due mappe date, questa metrica calcola la correlazione tra i valori di salienza delle due mappe. Il calcolo è dato dalla seguente formula:

$$CC = \frac{cov(S, F)}{\sigma_S \times \sigma_F}, \tag{2.2}$$

  dove si è usata la funzione covarianza (cov) e σ_S e σ_F sono le varianze delle due mappe. Si ricorda che la correlazione assume un valore nel range [-1,1]: -1 e 1 indicano una perfetta proporzionalità lineare e 0 si riferisce a due mappe incorrelate.

- KLD: è una misura utilizzata per valutare la similitudine tra due tipologie di distribuzioni. Essa quantifica la quantità delle informazioni perse quando la distribuzione di probabilità della mappa di fissazione dell'occhio umano viene approssimata utilizzando la distribuzione di probabilità della mappa di salienza. In particolare si calcola nel seguente modo:

$$KLD = \sum_{\rho=1}^{N} F(\rho) \times log\left(\frac{F(\rho)}{S(\rho) + \varepsilon} + \varepsilon\right), \tag{2.3}$$

dove X è il numero di pixel e ε è una piccola costante. Le distribuzioni di S e F sono entrambe normalizzate:

$$S(\rho) = \frac{S(\rho)}{\sum_{\rho=1}^{N} S(\rho) + \varepsilon}, \qquad F(\rho) = \frac{F(\rho)}{\sum_{\rho=1}^{N} F(\rho) + \varepsilon}. \tag{2.4}$$

Quando le due mappe sono strettamente uguali, il valore di divergenza KL è nullo.

Capitolo 3

Metodo Proposto

Questo capitolo mira a spiegare più approfonditamente il lavoro svolto per identificare le *feature* che risultano avere maggior peso nella salienza visiva dei contenuti a 360°.

Nell'introduzione si è gia accennato che l'utente, nel corso della visualizzazione di una scena a 360°, inevitabilmente si indirizza verso una singola *viewport*. La visualizzazione è quindi composta da una successione di *viewport* visualizzate con un ordine e una durata variabili.

Sulla base di questo comportamento, per il calcolo della salienza dell'intero contenuto bisogna in primis effettuare l'estrazione delle *viewport* a partire dall'immagine equirettangolare. Per ciascuna di queste verrà poi calcolata una mappa di salienza specifica sulla base di determinate *feature* e, infine, queste mappe verranno unite per creare la mappa di salienza omni-direzionale.

Si possono riconoscere 4 stadi fondamentali:

- pre-elaborazione;
- estrazione delle *feature*;
- calcolo della salienza;
- post-elaborazione.

3.1 Dataset utilizzato

Il dataset utilizzato nello studio è "Salient360!: Visual Attention Modelling for 360° content" proposto per la ICME'18 Grand Challenge [2, 3]. Il dataset di *training* "Salient360!" contiene 85 immagini equirettangolari che includono contenuti di varia natura come scene natu-

rali interne/esterne, scene contenenti volti umani, scene sportive. Un esempio è mostrato in Figura 3.1.

(a) Scena al chiuso

(b) Scena all'aperto

(c) Scena contenente volti

Figura 3.1: Esempi di immagini presenti nel dataset.

Le mappe con i dati di tracciamento oculare sono forniti in tre forme possibili:

- Dati *scan-path*: include le immagini con associati i dati *scan-path* ottenuti da 48 osservatori, ciascuno dei quali ha osservato i dati per un totale di 25 secondi. I percorsi di scansione sono costituiti dalle singole fissazioni e raccolgono informazioni sul movimento della testa e degli occhi. Ogni linea contiene un vettore che indica il numero di fissazione, il tempo di fissazione (in secondi), la posizione nell'immagine equirettangolare (indicate in pixel). Il numero di fissazione aumenta gradualmente per un osservatore specifico e si reimposta ad 1 ad ogni osservatore successivo. Ciascun osservatore è differenziato tramite l'utilizzo di diversi colori. Il numeri accanto ai cerchi specificano l'ordine di fissazione, mentre il cerchio stesso indica la posizione fissata. Un esempio è mostrato in Figura 3.2.

Figura 3.2: Mappa scan-path corrispondente all'immagine.

- Mappe di salienza basate sul movimento della testa (HM): comprende le immagini e le rispettive mappe di salienza con i dati di movimento della testa di 48 osservatori che hanno guardato l'immagine per 25 secondi ciascuno. I dati sono organizzati in un file binario contenente per ciascun pixel il rispettivo valore di salienza. Il valore minimo di salienza è 0 e la somma di tutti i valori di salienza dei pixel è pari a uno. Un esempio della mappa di salienza con l'applicazione di una *colormap* è fornito in Figura 3.3. Le regioni rosse più sature indicano le aree dove si focalizza maggiormente l'attenzione mentre le regioni blu più sature non sono molto considerate durante la visione.

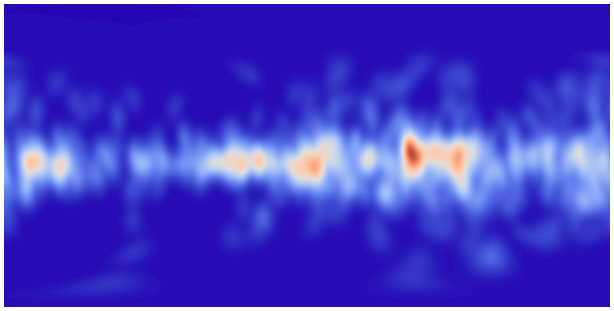


Figura 3.3: Mappa HM corrispondente all'immagine.

- Mappe di salienza basate sul movimento della testa e degli occhi (HM+EM): include le immagini e le rispettive mappe di salienza con i dati di movimento della testa e degli occhi (Yaw, Pitch, Roll e anche X-Gaze, Y-Gaze) di 48 osservatori che hanno guardato l'immagine per 25 secondi ciascuno. Un esempio è fornito in Figura 3.4

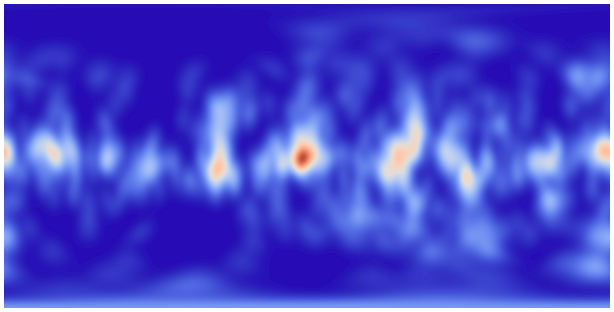


Figura 3.4: Mappa HM+EM corrispondente all'immagine.

3.2 Estrazione delle viewport

Disponendo in input di un'immagine equirettangolare bisogna innanzitutto provvedere all'estrazione delle *viewport*. A questo scopo si è utilizzato l'approccio descritto in [24], nel quale viene fatto uso della proiezione gnomonica sfera-piano.

In questo metodo tutte le *viewport* estratte, V_i dove $i = 1, 2, \ldots, n$, hanno le stesse dimensioni fissate: indicheremo la lunghezza con V_{width} e l'altezza con V_{height}. La proiezione gnomonica può essere ottenuta a partire da una sfera di raggio unitario e un piano tangente a questa in un punto della superficie: essa consiste nell'intersezione tra il piano e una semiretta che ha origine nel centro della sfera e attraversa la superficie nel punto che si desidera proiettare. Pertanto con questo metodo le *viewport* estratte giacciono nel piano tangente alla sfera come mostrato nella Figura 3.5.

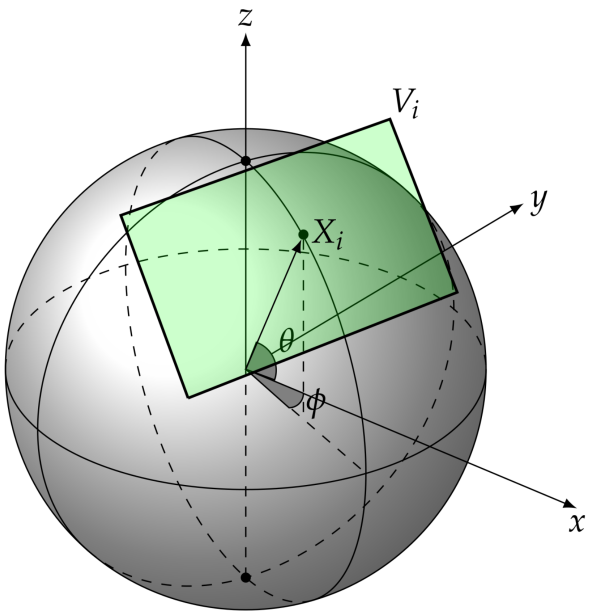


Figura 3.5: Estrazione delle *viewport*.

Nella pratica, l'immagine sferica viene sottoposta ad un campionamento angolare non uniforme. Le velocità di campionamento angolare orizzontale e verticale sono indicati rispettivamente come $\Delta\phi$ e $\Delta\theta$ e il punto campionato viene rappresentato da $X_i(\phi, \theta)$. Siccome il cambiamento della *viewport* è operato tramite un movimento della testa, consideriamo che la coordinata campionata $X_i(\phi, \theta)$ corrisponderà sempre al centro della *viewport* Vi che giace nel piano tangenge alla sfera.

Sia $C_{V_i}(x, y, z)$ un punto della *viewport*, questo è generalmente rappresentato come mostrato in Equazione 3.1:

$$C_{V_i}(x,y,z) = \begin{bmatrix} 1 \\ \left(2 \times \dfrac{tan\left(\dfrac{a}{2} \times \dfrac{\pi}{180}\right)}{V_{width}}\right) \times \left(x - \dfrac{V_{width}}{2}\right) \\ \left(2 \times \dfrac{tan\left(\dfrac{a}{2} \times \dfrac{\pi}{180}\right)}{V_{width}}\right) \times \left(y - \dfrac{V_{height}}{2}\right) \end{bmatrix}, \tag{3.1}$$

dove a rappresenta la dimensione in gradi della viewport e $\left(2 \times \dfrac{tan\left(\dfrac{a}{2} \cdot \dfrac{\pi}{180}\right)}{V_{width}}\right)$ rappresenta la dimensione dei pixel che la costituiscono.

Sia $||C_{V_i}||$ la norma L^2 del vettore C_{V_i}, la proiezione di quest'ultimo sulla sfera è data dall'Equazione 3.2:

$$\mathrm{C}_{V_i}^{sfera} = \frac{C_{V_i}}{||C_{V_i}||}. \tag{3.2}$$

Per ultimo la proiezione equirettangolare è data dall'Equazione 3.3:

$$\mathrm{C}_{V^i}^{equi}(x,y) = \begin{bmatrix} E_{width} \times \left(\dfrac{ang}{2\pi}\right) \\ E_{height} \times \left(\dfrac{\arcsin(C_{V_i}(z))}{pi + 0.5}\right) \end{bmatrix}, \tag{3.3}$$

dove $[E_{width}, E_{height}]$ è la dimensione dell'immagine equirettangolare e *ang* è dato da $ang = tan^{-1}(C_{V_i}(y), C_{V_i}(x))$.

3.3 Estrazione delle *feature*

Tenendo a mente l'obiettivo di questa tesi bisogna ora individuare un sottoinsieme di feature da impiegare per il calcolo della salienza. Attualmente in letteratura si possono trovare una grande varietà di set di *feature* relativi ai diversi modelli proposti. In questo studio, tuttavia, si è scelto di preferire alcune *feature* ampiamente diffuse.

Nello specifico si è scelto di utilizzare:

- Tonalità, saturazione e luminanza (Hue (H), Saturation (S), Luminance (L)): il modello HSL permette di specificare un colore attraverso tre valori numerici.

  La tonalità è rappresentata da un valore di un angolo del cerchio dei colori, e quindi un numero compreso tra 0 e 360. Lo spazio dei colori è attraversato in senso orario, pertanto considerando la Figura 3.6, l'angolo 0° corrisponde al rosso, l'angolo di 90° al verde e

Figura 3.6: Valori della tonalità.
fonte: https://en.wikipedia.org/wiki/File:HueScale.svg

cosi per tutti gli altri colori.

Il secondo valore corrisponde alla saturazione che rappresenta l'intensità della tonalità: più semplicemente un valore di saturazione più alto è legato ad un colore più acceso, in caso contrario il colore risulta essere una sfumatura di grigio. La saturazione è indicata tramite un valore percentuale.

Anche il terzo valore, ovvero la luminosità, è rappresentata come percentuale. Questa indica il grado di brillantezza dell'immagine: un valore prossimo al 100% porterà qualsiasi tonalità sul bianco mentre un valore di luminosità vicino a zero porterà qualsiasi colore al nero.

- Graph-Based Visual Saliency (GBVS): è un modello di salienza *bottom-up* classico per immagini 2D [19], spesso incluso anche in modelli di stima della salienza per immagini a 360°.

  Questo modello è costituito da tre fasi: i) creazione di *feature map*; ii) attivazione delle mappe tramite l'approccio delle catene di Markov; iii) normalizzazione e combinazione delle mappe ottenute. Tra le *feature map* calcolate nella prima fase troviamo mappe di orientamento calcolate usando i filtri di Gabor, mappe di contrasto calcolate utilizzando la varianza di luminanza e mappe di luminanza.

- Rilevamento dei bordi (Edge Detection (ED)): mira a identificare e mettere in risalto i contorni come la presenza di angoli, linee o curve che sono visivamente importanti e che quindi attirano l'attenzione umana [37]. Sebbene ci siano molte tecniche per riconoscere i contorni, la maggior parte di esse rientra in una delle due categorie principali: metodi basati sul calcolo della derivata del primo ordine dell'intensità dell'immagine o quelli in cui si cercano i punti in cui si annulla la derivata del secondo ordine.

  In questo studio è stato utilizzato l'operatore Prewitt [38] che rientra tra gli operatori del primo ordine. Questo calcola in ciascun punto un'approssimazione del gradiente dell'intensità dell'immagine, e quindi la direzione della sua massima variazione. Nella pratica, come riportato nell'Equazione 3.4 l'operatore agisce tramite la convoluzione

dell'immagine I con due kernel 3×3, uno per le variazioni orizzontali e uno per quelle verticali. Il risultato di questa operazione è dato da due immagini P_x e P_y.

$$P_x = \begin{bmatrix} +1 & 0 & -1 \\ +1 & 0 & -1 \\ +1 & 0 & -1 \end{bmatrix} \times I, \quad P_y = \begin{bmatrix} +1 & +1 & +1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix} \times I. \tag{3.4}$$

- Entropia (Entropy (E)): il concetto di entropia è largamente utilizzata nella teoria dell'informazione. Per definizione l'entropia è una misura dell'incertezza associata a una variabile aleatoria, nello specifico essa quantifica il valore atteso delle informazioni contenute in un messaggio. La relazione per il calcolo dell'entropia è riportata nell'Equazione 3.5:

$$H(X) = H(P_1, \ldots, P_n) = -\sum_{i=1}^{n} P_i \log_2 P_i, \tag{3.5}$$

dove P_i denota la probabilità che $X = x_i$ con x_i che indica l'i-esimo possibile valore di X su n simboli.

Nel contesto delle immagini digitali l'entropia risulta uno strumento vantaggioso per il calcolo della complessità del contenuto, caratteristica che può incidere sulla salienza [39]. Le regioni dell'immagine con un'elevata entropia presentano una complessità maggiore e pertanto sono quelle più salienti, al contrario un'entropia ridotta corrisponde ad un'area omogenea. Si considera quindi l'entropia per una *viewport* come:

$$E_{V_i} = -\sum_{r} p_r \log_2 p_r, \tag{3.6}$$

dove p_r è la probabilità che un pixel *r* sia presente nella viewport V_i.

- Contenuto in primo piano e sullo sfondo (Foreground-Background Content (FB)): i risultati ottenuti dallo studio condotto in [40] hanno confermato teorie precedenti risalenti a Gestalt secondo le quali esiste un'asimmetria attenzionale tra l'area di un immagine in primo piano e lo sfondo. Nello specifico, quando un soggetto visualizza un'immagine è portato a focalizzarsi maggiormente sul contenuto che si trova in primo piano, trascurando il gli elementi sullo sfondo. Questo fatto può rivelarsi determinante nel calcolo della salienza.

  A questo scopo è necessario sviluppare un metodo di estrazione di tale regione saliente dall'immagine. In seguito verrà utilizzato il modello proposto in [41]. Questo approccio

è basato sui grafi e permette di individuare tali regioni basandosi sul calcolo delle loro distanze dal bordo dell'immagine e sul calcolo delle distanze di un vertice del grafo da quelli adiacenti.

- Presenza di oggetti (Presence of Object (PO)): L'attenzione umana è influenzata anche da aree contenenti oggetti [42]. In questo studio per il rilevamento di oggetti è stato utilizzato il modello *YOLO* ("You Only Look Once") [43], addestrato sul dataset *COCO* [44]. *YOLO*, proposto da J. Redmon *et al.* nel 2015, è un modello basato sulle CNN. Per la prima volta, il problema di localizzazione e classificazione degli oggetti in un'immagine è stato risolto utilizzando solo la regressione. *YOLO* a differenza dei modelli precedenti utilizza un approccio a singolo stadio: divide l'immagine in regioni, predice le *bounding-box* e, per ciascuna di esse, determina le probabilità di appartenere ad una certa classe, il tutto utilizzando un'unica rete.
  Il dataset *COCO* è composta da oltre 330.000 immagini suddivise in 80 categorie di classi come ad esempio persone, auto e tavoli. Per ogni *viewport* riconosciamo un insieme di oggetti in cui vengono specificati il nome, che è legato alla classe di appartenenza, l'accuratezza del riconoscimento e le dimensioni della *bounding-box*.
  Va osservato che in una scena non tutti gli oggetti presenti sono salienti: è esemplificativo il caso in cui la nostra attenzione è catturata da una persona se questa si trova in mezzo a tante macchine. Per selezionare correttamente tali oggetti si assegna un livello di salienza ad ogni oggetto calcolato a partire da tre parametri: dimensione della bounding box, accuratezza di riconoscimento e occorrenza dell'oggetto nella *viewport*. Per il calcolo della salienza si scelgono i primi K oggetti riconosciuti, in questo studio si pone K pari a 3.

- Presenza di persone (Presence of People (PP)): questo elemento spesso risulta essere determinante nel guidare il processo di attenzione, pertanto la sua inclusione può migliorare la qualità della stima della salienza [39].
  A tal fine per rilevare i volti verrà utilizzato l'algoritmo *TinyFace* presentato in [45] poiché è efficace sia per le immagini con molti visi piccoli che per le immagini con visi grandi. Si basa su un modello a risoluzione ibrida multistrato per identificare volti in una scena a grande e piccola risoluzione. Il modello utilizza l'architettura CNN *ResNet101* addestrato sul set di dati *WIDER FACE* [46].

3.4 Calcolo delle mappe di salienza

Una volta stabilite le *features* da impiegare, è possibile procedere al calcolo delle mappe di salienza una per ogni *feature* e per ciascuna immagine del dataset adottato. Questo passaggio si traduce nella pratica nel calcolo delle mappe di salienza delle singole *viewport* che costituiscono un'immagine omnidirezionale.

In questo studio, inoltre, viene effetuata un'integrazione delle mappe che fanno riferimento alle caratteristiche intrinseche dell'immagine (*low-level feature*) e caratterische che rigurardano il contenuto (*high-level feature*) ottenendo per ciascuna un'unica mappa.

Per ogni immagine si calcolano le seguenti mappe:

- Mappa Low-Level (LL): questa mappa è calcolata combinando le tre mappe basate sulle *feature low-level* : la luminanza, la saturazione e tonalità.

$$LL = \alpha H + \beta S + \gamma L \tag{3.7}$$

- Mappa High-Level (HL): questa mappa è calcolata a partire dalla mappa presenza di persone e presenza di oggetti, che sono per l'appunto *feature high-level*

$$HL = \lambda PP + \delta PO \tag{3.8}$$

- Entropia (E)
- *Foreground/Background* (FB)
- GBVS (GBVS)
- *Edge Detection* (ED)

Tutte le mappe ottenute vengono poi combinate per ottenere alla fine una singola mappa di salienza omni-direzionale riferita all'immagine di partenza.

A tal fine è stato scelto di considerare tutti i casi possibili calcolando tutte le combinazioni di due, tre, quattro, cinque e sei *feature*.

A questo punto serve definire il valore dei parametri che verranno utilizzati nelle varie combinazioni. In tal senso, si è deciso di proseguire in due fasi.

Inizialmente, una volta analizzata l'incidenza che hanno le diverse caratteristiche sulla mappa finale (4), le *features* sono state combiante assegnando ai parametri il valore unitario, senza perciò eseguire una somma pesata.

In un secondo momento sono stati calcolati dei valori opportuni da dare ai parametri con l'obiettivo di dare maggior peso alle *feature* con un valore di correlazione più alto con la *ground-truth* e minor peso alle *feature* con prestazioni peggiori.

Siano n il numero di *feature* che si vogliono combinare, p_i con $i = 1, \ldots, n$ i parametri usati nella somma pesata, CC_i il valore della correlazione per la *feature* i-esima e CC_m il minimo tra le correlazioni, allora i valori dei pesi sono stati calcolati nel seguente modo:

$$\begin{aligned} 1 &= p_1 + p_2 + ...p_n \\ 1 &= \frac{CC_1}{CC_m} \times p_m + \frac{CC_2}{CC_m} \times p_m + ...\frac{CC_n}{CC_m} \times p_m, \end{aligned} \tag{3.9}$$

da cui

$$p_i = \frac{CC_i}{CC_m} \times p_m \tag{3.10}$$

dove

$$p_m = \frac{1}{\sum_{i=1}^{N} \frac{CC_i}{CC_m}}. \tag{3.11}$$

Infine bisogna proiettare le mappe di salienza 2D delle singole viewport sul piano equirettangolare. A tal proposito si utilizza l'approccio proposto in [24].

3.5 Post-elaborazione

Per ottenere la mappa di salienza definitiva vengono eseguite due ulteriori operazioni: l'applicazione del *bias* equatoriale e normalizzazione.

Si è detto precedentemente che durante l'esplorazione di una scena a 360° l'utente ha una finesatra di visualizzazione (FOV) limitata pertanto, per estendere l'area da esplorare, sono necessari il movimento della testa o del corpo. Questi movimenti intrinsecamente influiscono sulla rilevanza di una scena nella misura in cui viene richiesta una quantità di sforzo fisico diversa per le diverse porzioni dell'immagine. Nel concreto il soggetto è portato a impiegare uno sforzo minore per portare lo sguardo lungo l'equatore in confronto ad altre aree dell'immagine, per cui è necessaria una combinazione del movimento della testa, occhio e tronco. Pertanto la posizione della *viewport* è un elemento che deve essere preso in considerazione durante il processo di stima della salienza.

A tal proposito si è applicata una finestra di ponderazione seguendo l'approccio in [24]: questa consiste in una funzione che cresce proporzionalmente con la distanza dall'equatore ed ha valori compresi nell'intervallo $[1, 4]$ (figura 3.7). Quindi all'area vicino all'equatore viene

fatto corrispondere il valore 1 mentre all'area periferica, più lontana, è assegnato il valore 4. Il calcolo della mappa di salienza infine consiste la divisione per la funzione di costo determinata, dando maggior peso ai pixel che si trovano nella regione equatoriale.

Per ultimo la normalizzazione viene utilizzata per assicurarsi che i valori della mappa di salienza rientrino in un intervallo coerente e che l'importanza relativa delle diverse regioni dell'immagine sia rappresentata in modo accurato.

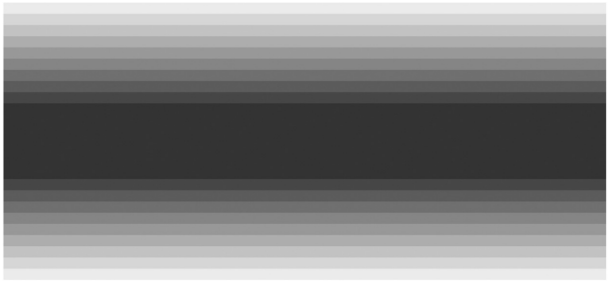


Figura 3.7: Finestra di ponderazione per tener conto del *bias* equatoriale.

Capitolo 4

Risultati sperimentali

Questo capitolo mira a presentare i risultati ottenuti a seguito della valutazione delle prestazioni mediante l'utilizzo del dataset di riferimento *Salient360!*.

Sono state selezionate due metriche standard tra quelle disponibili: il coefficiente di correlazione (CC) e della divergenza Kullback-Leibler (KLD).

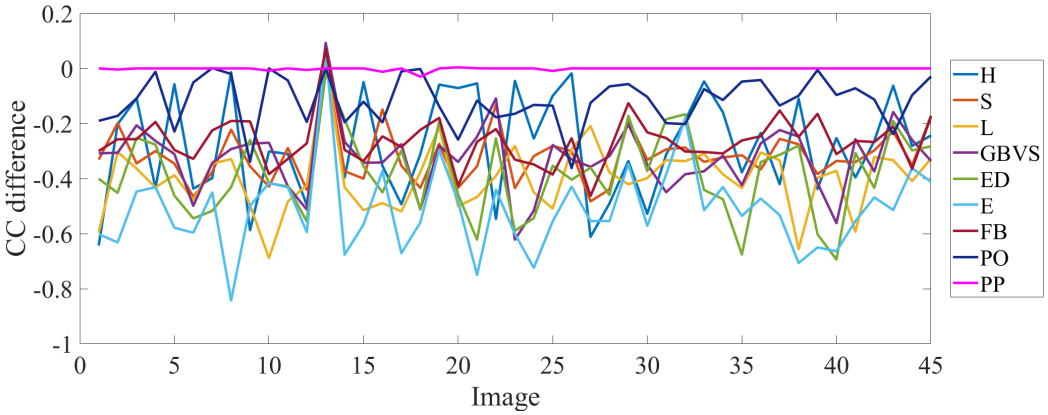
Si ricorda brevemente che la metrica CC è rappresentata da un valore nel range $[-1, 1]$ ed indica la relazione statistica tra la mappa di salienza stimata e quella data (*ground-truth*). Pertanto si ha che tanto maggiore è il valore restituito in valore assoluto quanto migliore è la stima della salienza. D'altra parte, la metrica KLD quantifica il grado di dissomiglianza tra le due mappe: un valore basso indica un buon approccio alla stima della salienza.

4.1 Valutazione del *bias*

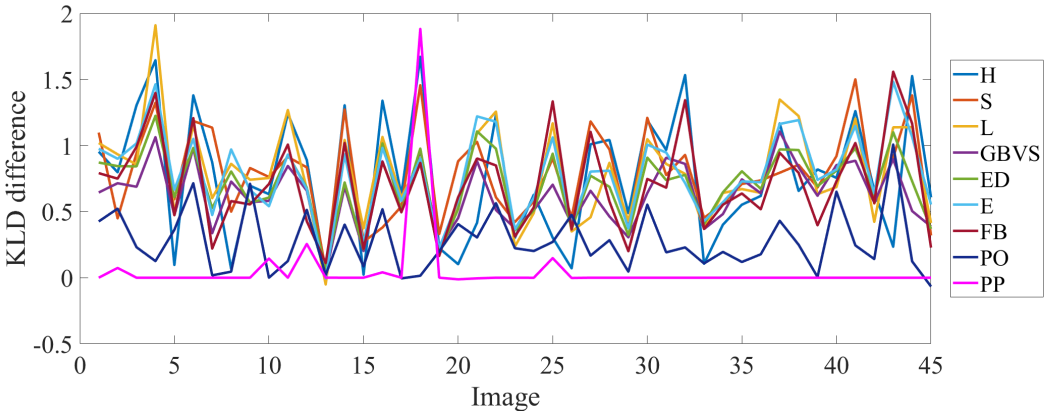
Come prima cosa è stato studiato l'impatto del pesaggio a favore di un *bias* equatoriale fatto nella fase di post-elaborazione.

A tale proposito per il confronto si è scelto di determinare per ciascuna *feature* la differenza tra i valori di CC e KLD ottenuti senza e con pesaggio. La Figura 4.1 rappresenta un grafico che mostra i risultati ottenuti per ogni *feature* e per ogni immagine.

I valori delle differenze sono generalmente tutti negativi per CC, ad eccezione di un'immagine, e positivi per KLD dimostrando cosi l'efficacia di questa operazione. Inoltre si osserva come il *bias* non abbia effetti notevoli sulla mappa di salienza che tiene conto solo delle *feature high-level*. Si nota infatti come i valori più piccoli corrispondano in primis alla presenza di persone e poi alla presenza di oggetti.



(a) CC



(b) KLD

Figura 4.1: Differenze tra i valori di CC and KLD senza e con *bias*.

4.2 Valutazione delle *feature*

Stando ai risultati ottenuti, per lo studio seguente verranno considerate le mappe di salienza che includono l'aggiunta del *bias* equatoriale.

Nella Tabella 4.1 sono riportate le medie dei valori di KLD e CC ottenute per le mappe che utilizzano una singola *feature*, mentre nella Figura 4.2 sono forniti i *boxplot* relativi alle due metriche.

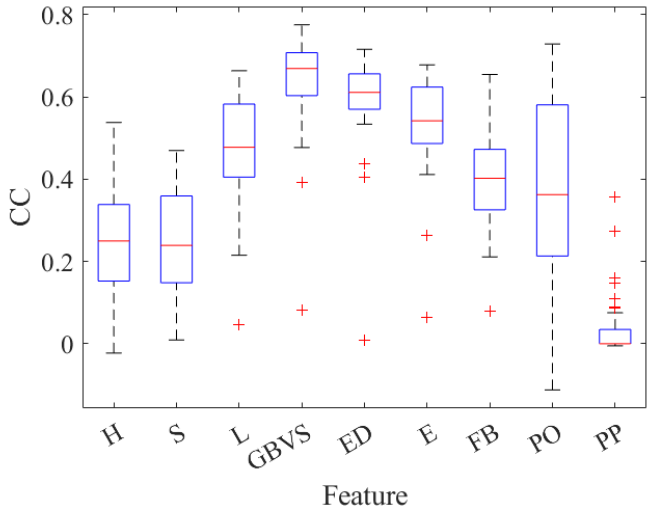
Bisogna specificare che nel caso di immagini mancanti di soggetti umani il valore di CC è stato impostato a 0 in quanto la mappa di salienza calcolata e quella vera sono incorrelate.

Tabella 4.1: Prestazioni dellle singole *feature*.

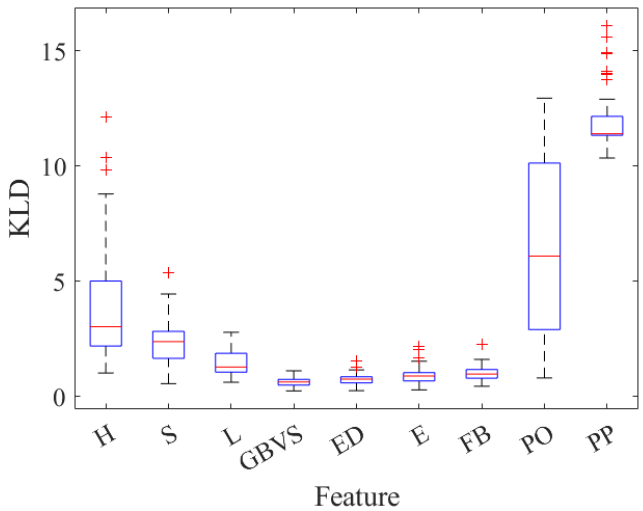
	H	S	L	GBVS	ED	E	FB	PO	PP
CC	0.24	0.25	0.47	0.64	0.60	0.54	0.40	0.37	0.03
KLD	4.03	2.36	1.46	0.66	0.76	0.94	1.02	6.48	12.11

Se ci riferiamo ai dati della tabella si osserva come GBVS abbia prestazioni migliori, seguito da ED ed E, mentre la PP si comporta peggio rispetto tutte le altre *feature*. Inoltre se

consideriamo anche i *boxplot* associati, si può notare come PP e PO abbiano nel complesso prestazioni peggiori rispetto alle altre *feature*, tra le quali dominano in performance GBVS, ED ed E. Questo fenomeno può essere spiegato con il fatto che le *feature high-level* si basano su informazioni troppo specifiche, in assenza delle quali manca l'utilità delle funzionalità di alto livello: ne sono di esempio alcune immagini in cui non sono presenti persone o oggetti che possano essere riconosciuti dagli algoritmi utilizzati per rilevarli. Al contratio, le altre *feature* fanno riferimento a delle caratteristiche globali dell'immagine riuscendo così a raggiungere prestazioni migliori.



(a) CC

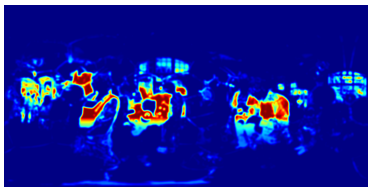


(b) KLD

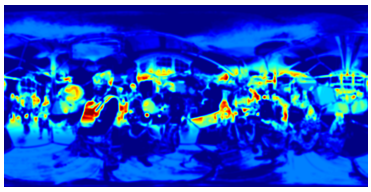
Figura 4.2: Boxplot delle metriche CC e KLD tra la mappa di salienza stimata basata su una singola caratteristica e la mappa *ground-truth*.

Nella Figura 4.3 sono riportate le mappe di salienza ottenute utilizzando le *feature* singolarmente.

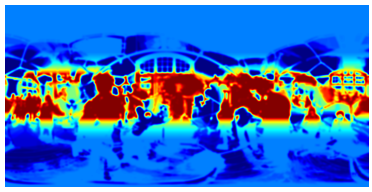
(a) Immagine di partenza



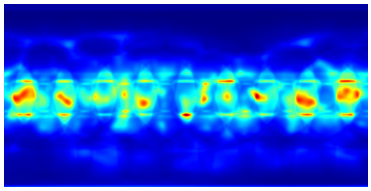
(b) H



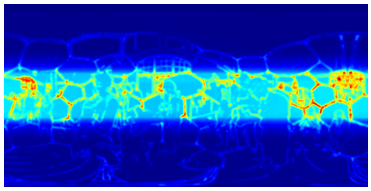
(c) S



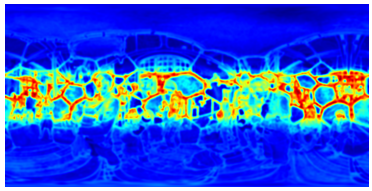
(d) L



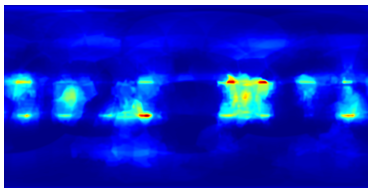
(e) GBVS



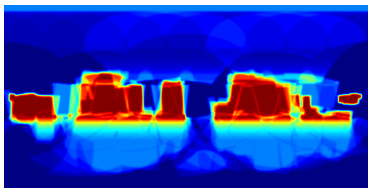
(f) ED



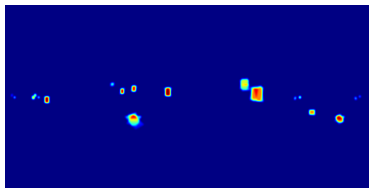
(g) E



(h) FB



(i) PO



(j) PP

Figura 4.3: Esempio di mappe di salienza delle singole *feature*.

4.3 Valutazione delle combinazioni

A questo punto è interessante analizzare le performance delle combinazioni. A tal proposito lo studio è suddiviso in due fasi: nella prima si sono considerate le *feature* da combinare tutte con la stessa rilevanza, mentre nella seconda si tiene conto dell'impatto che queste possono avere sulla mappa finale.

4.3.1 Valutazione considerando parametri unitari

Questa fase si traduce nel concreto con l'attribuire il valore unitario a tutti i parametri nelle varie combinazioni effettuate.

In primo luogo sono state calcolate le mappe LL e HL e in seguito i valori di CC e KLD corrispondenti (Tabella 4.2). Dai risultati ottenuti si evince che ancora una volta si ottengono valori migliori nel caso di stime della salienza fatte tramite l'utilizzo delle *low-level feature*. Per la mappa HL i valori delle metriche coincidono con quelli di PO, meglio performante rispetto a

Tabella 4.2: Prestazioni LL e HL.

	LL	HL
CC	0.47	0.37
KLD	1.11	6.48

PP. Per quanto riguarda il caso di LL il comportamento non è analogo: il valore di correlazione coincide con quello di L, migliore rispetto ai valori delle altre mappe, mentre la KLD è diversa e di poco migliore rispetto ai valori di divergenza delle tre mappe H, S, L.

A questo punto è possibile eseguire l'analisi delle prestazioni delle varie combinazioni. A tale scopo si è deciso di avere un quadro completo di come queste vadano ad incidere sulla mappa di salienza valutando tutti i possibili insiemi di due, tre, quattro, cinque e sei *feature*.

Le seguenti tabelle riportano i risultati ottenuti. Nella Tabella 4.3 sono riportati i risultati relativi alla combinazione di due *feature*, nella Tabella 4.4 sono presentate le prestazioni ottenute utilizzando tre *feature*, nella Tabella 4.5 sono mostrati i risultati relativi alla combinazione di quattro *feature*, e la Tabella 4.6 riporta le prestazioni associate alla combinazione di cinque e sei *feature*.

A prima vista si nota che in queste tabelle i valori ottenuti sono migliori rispetto al caso in cui le *feature* sono considerate singolarmente. Nel caso di due *feature*, la combinazione tra GBVS ed ED porta ad un piccolo miglioramento in CC rispetto che considerare singolarmente la GBVS, che costituisce la mappa di salienza a singola *feature* meglio performante. Globalmente la combinazione composta da GBVS, HL ed ED presenta i risultati migliori, mentre le combinazioni che comprendono un numero maggiore di *feature* portano a risultati che saturano o addiritura peggiorano.

Questi risultati dimostrano che considerare un elevato numero di *feature* per determinare la mappa di salienza non comporta necessariamente un miglioramento delle prestazioni ma, al contrario, è importante scegliere attentamente le tipologie di *feature* da includere nelle combinazioni. Infatti si nota come all'interno della stessa tabella i risultati cambiano a seconda delle combinazioni effettuate e per ottenere buoni risultati è bene selezionare un insieme di *feature* che apportino un contenuto informativo diverso e non ridondante. Questo fenomeno può essere verificato direttamente se si considerano i risultati ottenuti nelle Tabelle 4.5 e 4.6: nella prima le prestazioni hanno un andamento variabile a seconda delle *feature* combinate, senza però ottenere un miglioramento e nella seconda essi subiscono nel complesso un leggero peggioramento.

Quando vengono combinate quattro *feature* la combinazione migliore vede l'aggiunta di

Tabella 4.3: Combinazione di due feature.

Feature	GBVS,LL	GBVS,HL	GBVS,ED	GBVS,E	GBVS,FB
CC	0.54	0.63	0.65	0.63	0.60
KLD	0.89	0.60	0.69	0.76	0.70
Feature	LL,HL	LL,ED	LL,E	LL,FB	HL,ED
CC	0.52	0.51	0.52	0.50	0.60
KLD	0.97	0.93	0.98	0.95	0.66
Feature	HL,E	HL,FB	ED,E	ED,FB	E,FB
CC	0.59	0.51	0.58	0.56	0.56
KLD	0.78	0.80	0.82	0.77	0.84

Tabella 4.4: Combinazione di tre feature.

Feature	GBVS,LL,HL	GBVS,LL,ED	GBVS,LL,E	GBVS,LL,FB	GBVS,HL,ED
CC	0.57	0.56	0.56	0.56	0.66
KLD	0.83	0.86	0.87	0.87	0.62
Feature	GBVS,HL,E	GBVS,HL,FB	GBVS,ED,E	GBVS,ED,FB	GBVS,E,FB
CC	0.65	0.63	0.63	0.63	0.62
KLD	0.69	063	0.75	0.70	0.75
Feature	LL,HL,ED	LL,HL,E	LL,HL,FB	LL,ED,E	LL,ED,FB
CC	0.55	0.55	0.55	0.54	0.53
KLD	0.86	0.89	0.88	0.90	0.90
Feature	LL,E,FB	HL,ED,E	HL,ED,FB	HL,E,FB	ED,E,FB
CC	0.54	0.62	0.61	0.61	0.59
KLD	0.90	0.74	0.67	0.74	0.80

Tabella 4.5: Combinazione di quattro *feature*.

Feature	GBVS,LL,HL,ED	GBVS,LL,HL,E	GBVS,LL,HL,FB	GBVS,LL,ED,E	GBVS,LL,ED,FB
CC	0.59	0.59	0.65	0.58	0.58
KLD	0.81	0.82	0.82	0.85	0.84
Feature	GBVS,LL,E,FB	GBVS,HL,ED,E	GBVS,HL,ED,FB	GBVS,HL,E,FB	GBVS,ED,E,FB
CC	0.57	0.65	0.65	0.65	0.63
KLD	0.85	0.84	0.85	0.88	0.73
Feature	LL,HL,ED,E	LL,HL,ED,FB	LL,HL,E,FB	LL,ED,E,FB	HL,ED,E,FB
CC	0.57	0.57	0.57	0.55	0.63
KLD	0.78	0.80	0.82	0.77	0.84

Tabella 4.6: Combinazione di cinque e sei *feature*.

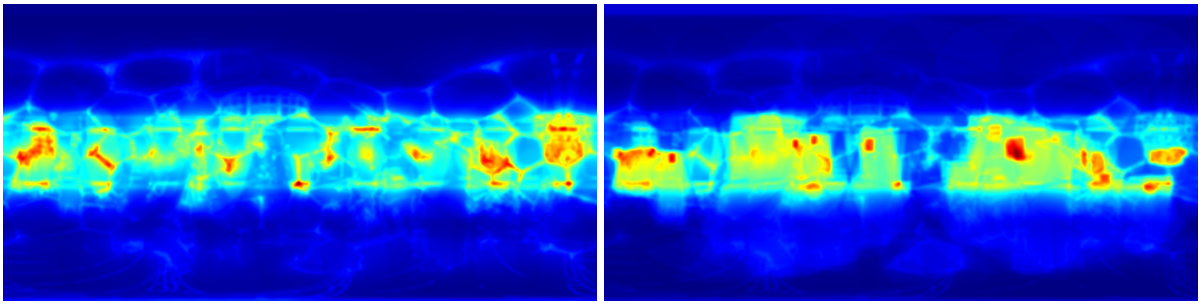
Feature	GBVS,LL,HL,ED,E	GBVS,LL,HL,ED,FB	GBVS,LL,HL,E,FB	GBVS,LL,ED,E,FB
CC	0.60	0.60	0.60	0.59
KLD	0.81	0.80	0.81	0.84
Feature	GBVS,HL,ED,E,FB	LL,HL,ED,E,FB	GBVS,LL,HL,ED,E,FB	
CC	0.65	0.58	0.61	
KLD	0.70	0.83	0.80	

FB a GBVS,HL,ED ottendendo però un piccolo peggioramento del valore di CC. Nell'ultima tabella, invece, l'ulteriore aggiunta di E a questa combinazione porta un piccolo perggioramento in KLD.

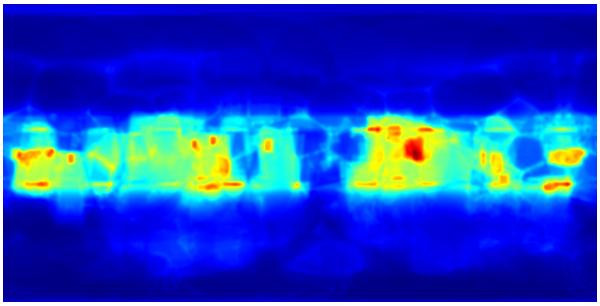
Infine, mentre da un lato era abbastanza prevedibile che le *feature* meglio prestanti comparissero nelle combinazioni con le prestazioni migliori, dall'altro è interessante notare la presenza di HL nelle combinazioni più performanti. Questo dimostra che le *feature high-level*, nonostante non riescano ad ottenere buoni risultati quando considerate singolarmente, rappresentano un valore aggiunto nelle combinazioni in quanto portano informazioni necessarie per ottenere una stima della salienza più accurata. In Figura 4.5 sono riportate le mappe delle tre combinazioni con le prestazioni migliori, corrispondenti all'immagine mostrata in Figura 4.4a. La *ground-truth* corrispondente è fornita in Figura 4.4b.

(a) Immagine originale (b) Mappa *ground-truth*

Figura 4.4: Esempio di immagine del dataset e mappa *ground-truth* corrispondente.



(a) GBVS,ED (b) GBVS,HL,ED



(c) GBVS,HL,ED,FB

Figura 4.5: Esempi delle mappe di salienza più performanti riferiti a 4.4a.

4.3.2 Valutazione a seguito del calcolo dei parametri

In un secondo momento è stato analizzato l'effetto in termini di prestazioni di un pesaggio non unitario delle *feature*. A questo scopo, è stato fatto il calcolo dei pesi secondo la modalità discussa nella Sezione 3.4, con l'obiettivo di fornire maggior rilevanza alle *feature* che hanno mostrato risultati migliori in termini di CC quando valutate singolarmente. I valori dei parametri ottenuti sono riportati nella Tabella 4.7.

Tabella 4.7: Valore dei paramtetri di pesaggio per ciascuna *feature*.

	H	S	L	GBVS	ED	E	FB	PO	PP
CC	0.24	0.25	0.47	0.64	0.60	0.54	0.40	0.37	0.03
Parametro	0.0678	0.0706	0.1328	0.1808	0.1695	0.1525	0.1130	0.1045	0.0085

Si nota innanzitutto come i parametri, in modo conforme al valore di CC relativo alle singole *feature*, diano maggiore rilevaza in primis a GBVS, ED ed E, assumendo tuttavia valori che non differiscono di tanto; d'altra parte il peso attribuito a PP risulta essere molto più piccolo. Le altre *feature* infine assumono valori intermedi che non differiscono di tanto: le coppie H ed S e PO e FB assumono una rilevanza comparabile, mentre L raggiunge un valore più alto. Una volta ottenuti i parametri è possibile procedere con il calcolo delle combinazioni, in modo analogo a quanto già effettuato in precedenza. Nella Tabella 4.8 sono riportati i risultati ottenuti per la mappa LL e HL.

Tabella 4.8: Prestazioni LL e HL dopo il calcolo dei pesi.

	LL	HL
CC	0.47	0.37
KLD	1.08	6.48

Si nota comunque che i valori riportati coincidano con quelli calcolati senza il pesaggio, ad eccezione di KLD per la mappa LL che subisce un piccolo miglioramento.

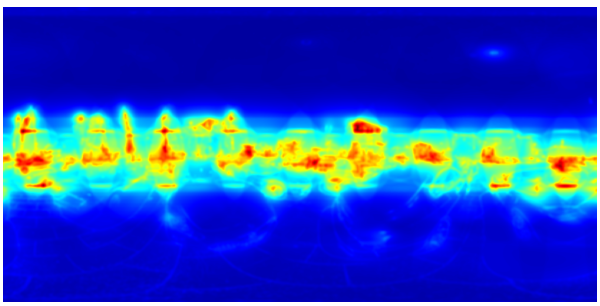
Da questo nasce la necessità di investigare ulteriormente il comportamento nelle altre combinazioni. Nelle Tabelle 4.9, 4.10, 4.11 e 4.12 sono riportati i risultati ottenuti, mentre nelle Figure 4.7 e 4.8 sono riportati i grafici in cui è possibile vedere più chiaramente il cambiamento in termini di prestazioni se confrontato con il caso in cui i pesi sono tutti unitari. In particolare, per ciascuna combinazione è stata calcolata la differenza tra il valore di CC ottenuto dopo il calcolo dei parametri e quello ottenuto considerando il pesaggio con valori unitari.

A prima vista si nota che i valori delle ordinate sono per lo più positivi per CC e negativi per KLD, stando ad indicare un miglioramento delle performance per entrambe le metriche.

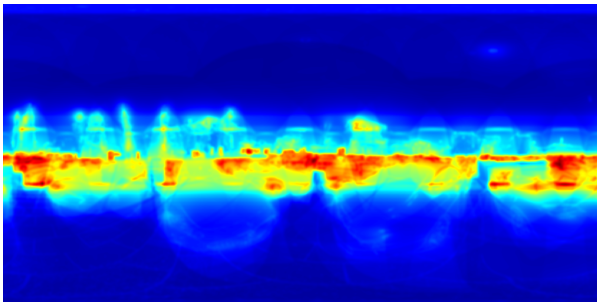
Tuttavia si può osservare che i valori ottenuti risultano essere molto piccoli e in alcuni casi nulli. Infine, per alcune combinazioni si ottiene un peggioramento delle prestazioni.

Questi comportamenti sono dovuti a diversi fattori. Innanzitutto, con il pesaggio le *performance* finali sono influenzate maggiormente dal contributo delle *feature* che singolarmente sono meglio prestanti. Tuttavia, si è visto che, quando considerate singolarmente, queste *feature* sono caratterizzate da un valore di correlazione comparabile per cui nelle combinazioni con pesaggio ottengono la stessa rilevanza per il calcolo della mappa finale. Questo effetto è analogo al caso in cui le *feature* sono considerate i pesi unitari e per questo motivo si ottengono valori simili di CC e KLD. Inoltre, procedendo in questo modo si dà poco peso ad alcune *feature* solo sulla base delle loro singole *performance*. Tuttavia, si è visto che, sebbene una *feature* possa essere poco rilevante quando considerata singolarmente, è possibile che la stessa *feature* rappresenti un valore aggiunto in una combinazione, come accade per HL.

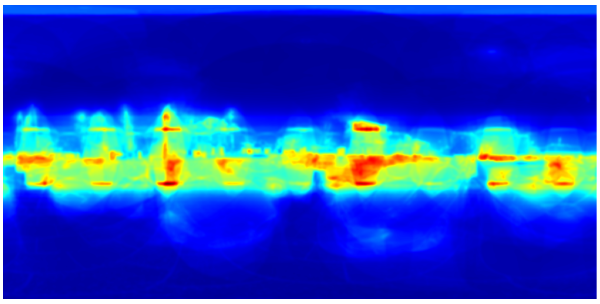
Di seguito, nella Figura 4.6 sono riportate le combinazioni con pesaggio più performanti.



(a) GBVS,ED



(b) GBVS,HL,ED



(c) GBVS,HL,ED,FB

Figura 4.6: Esempi delle mappe di salienza più performanti dopo il pesagio riferite a 4.4a.

Tabella 4.9: Combinazione di due *feature*.

Feature	GBVS,LL	GBVS,HL	GBVS,ED	GBVS,E	GBVS,FB
CC	0.58	0.65	0.65	0.63	0.62
KLD	0.83	0.60	0.68	0.75	0.68
Feature	LL,HL	LL,ED	LL,E	LL,FB	HL,ED
CC	0.53	0.54	0.54	0.52	0.62
KLD	0.95	0.88	0.95	0.92	0.67
Feature	HL,E	HL,FB	ED,E	ED,FB	E,FB
CC	0.59	0.51	0.58	0.59	0.56
KLD	0.80	0.80	0.82	0.76	0.84

Tabella 4.10: Combinazione di tre *feature*.

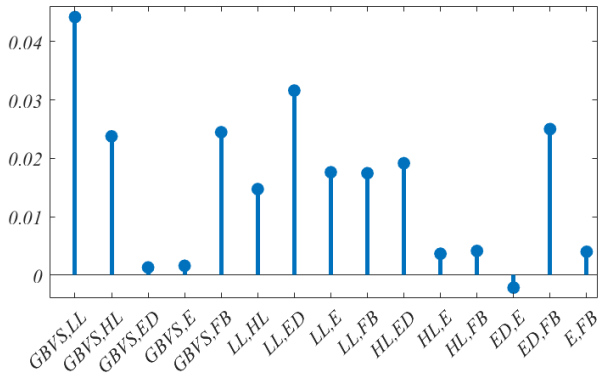
Feature	GBVS,LL,HL	GBVS,LL,ED	GBVS,LL,E	GBVS,LL,FB	GBVS,HL,ED
CC	0.61	0.60	0.60	0.59	0.67
KLD	0.78	0.80	0.83	0.81	0.63
Feature	GBVS,HL,E	GBVS,HL,FB	GBVS,ED,E	GBVS,ED,FB	GBVS,E,FB
CC	0.66	0.65	0.63	0.64	0.63
KLD	0.70	0.62	0.75	0.69	0.74
Feature	LL,HL,ED	LL,HL,E	LL,HL,FB	LL,ED,E	LL,ED,FB
CC	0.58	0.57	0.56	0.56	0.56
KLD	0.83	0.87	0.85	0.87	0.86
Feature	LL,E,FB	HL,ED,E	HL,ED,FB	HL,E,FB	ED,E,FB
CC	0.55	0.62	0.62	0.61	0.59
KLD	0.88	0.75	0.68	0.76	0.80

Tabella 4.11: Combinazione di quattro *feature*.

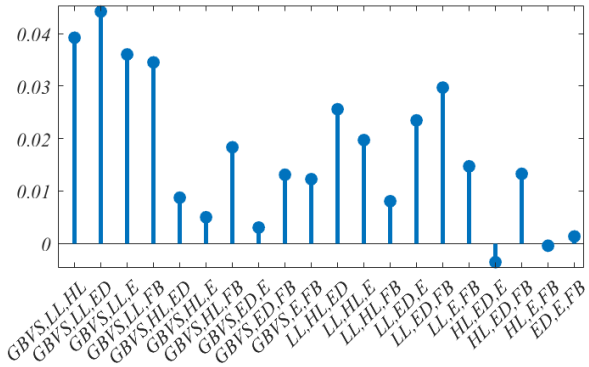
Feature	GBVS,LL,HL,ED	GBVS,LL,HL,E	GBVS,LL,HL,FB	GBVS,LL,ED,E	GBVS,LL,ED,FB
CC	0.62	0.61	0.62	0.61	0.61
KLD	0.77	0.79	0.77	0.81	0.79
Feature	GBVS,LL,E,FB	GBVS,HL,ED,E	GBVS,HL,ED,FB	GBVS,HL,E,FB	GBVS,ED,E,FB
CC	0.60	0.65	0.66	0.66	0.63
KLD	0.81	0.71	0.65	0.70	0.74
Feature	LL,HL,ED,E	LL,HL,ED,FB	LL,HL,E,FB	LL,ED,E,FB	HL,ED,E,FB
CC	0.59	0.59	0.58	0.58	0.63
KLD	0.83	0.81	0.84	0.85	0.74

Tabella 4.12: Combinazione di cinque e sei *feature*.

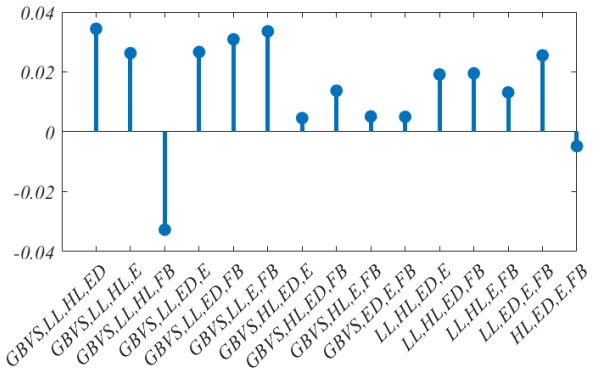
Feature	GBVS,LL,HL,ED,E	GBVS,LL,HL,ED,FB	GBVS,LL,HL,E,FB	GBVS,LL,ED,E,FB
CC	0.62	0.63	0.62	0.61
KLD	0.78	0.76	0.78	0.80
Feature	GBVS,HL,ED,E,FB	LL,HL,ED,E,FB	GBVS,LL,HL,ED,E,FB	
CC	0.66	0.60	0.63	
KLD	0.70	0.81	0.77	



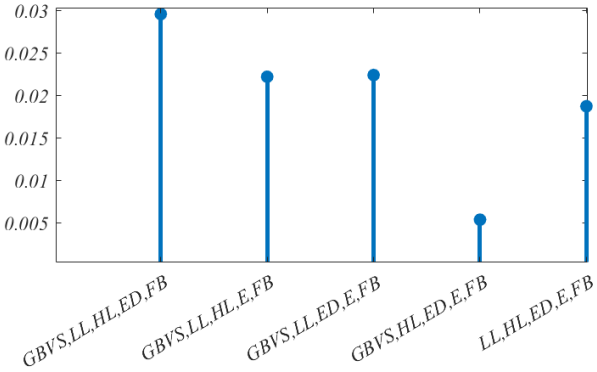
(a) Differenza CC per le combinazioni di 2 *feature*



(b) Differenza CC per le combinazioni di 3 *feature*

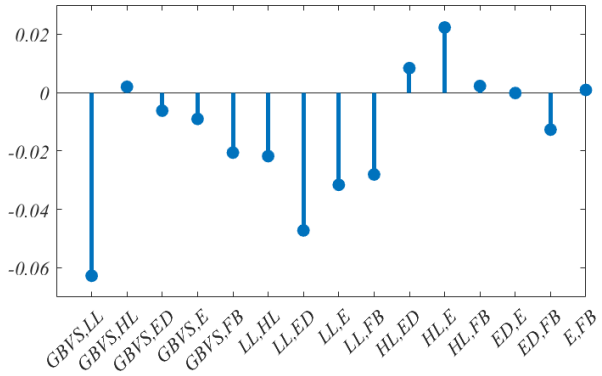


(c) Differenza CC per le combinazioni di 4 *feature*

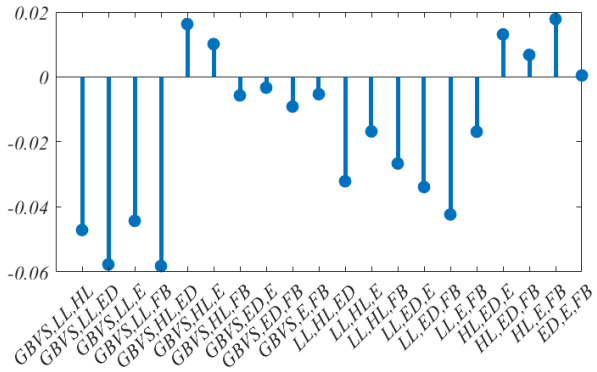


(d) Differenza CC per le combinazioni di 5 *feature*

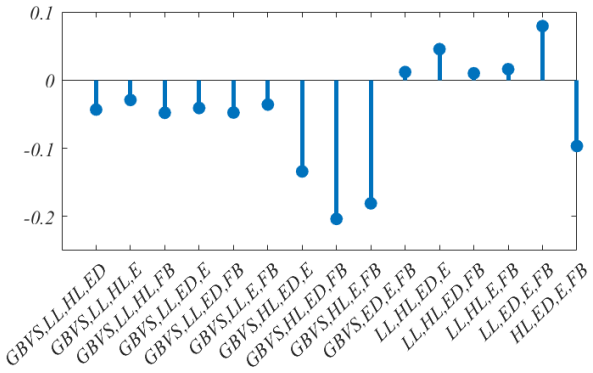
Figura 4.7: Grafici dei valori delle differenze per i valori di CC con e senza somma pesata.



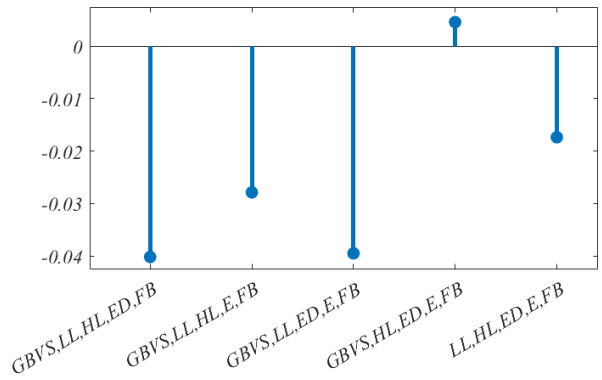
(a) Differenza KLD per le combinazioni di 2 *feature*



(b) Differenza KLD per le combinazioni di 3 *feature*



(c) Differenza KLD per le combinazioni di 4 *feature*



(d) Differenza KLD per le combinazioni di 5 *feature*

Figura 4.8: Grafici dei valori delle differenze per i valori di KLD con e senza somma pesata.

Capitolo 5

Conclusioni

Il lavoro di tesi svolto rappresenta un contributo allo studio della salienza visiva per le immagini omni-direzionali. Infatti, sebbene in letteratura ci siano numerosi modelli per la stima della salienza basati sulla combinazione di diverse *feature*, è stato poco investigato in quale misura queste ultime incidano sul calcolo complessivo.

A questo scopo sono state scelte un insieme di *feature* ampiamente utilizzate nei modelli presenti in letteratura. In un primo momento sono state calcolate le mappe di salienza, una per ogni *feature*, in modo tale da esaminare il loro impatto quando vengono prese singolarmente. In seguito è stata verificata l'efficacia dell'aggiunta di un *bias* equatoriale andando a confrontare le *performance* tra le mappe di salienza con e senza pesaggio. Successivamente si è scelto di calcolare tutti i possibili insiemi di due, tre, quattro,cinque e sei *feature* per studiare come le diverse combinazioni incidessero sulla mappa finale. Per la valutazione delle prestazioni si è scelto di adottare CC e KLD, due metriche ben consolidate per stimare il grado di somiglianza e divergenza con le mappe della *ground-truth*.

Le analisi hanno prodotto alcuni risultati interessanti. Innanzitutto i confronti delle metriche ottenute tra le mappe con e senza *bias* equatoriale hanno confermato che questa operazione migliora notevolmente le la stima delle mappe di salienza. Nel complesso si può affermare che la scelta di *feature* da combinare opportune condiziona notevolmente l'ottenimento di buone prestazioni, al contrario considerare un numero elevato di *feature* non sempre porta a buoni risultati. Questo è dovuto dal fatto che per una buona stima della salienza è opportuno considerare caratteristiche dell'immagine che forniscono un contenuto informativo diverso.

Sulla base delle *feature* analizzate, la combinazione composta da GBVS, HL,ED consente di ottenere i risultati migliori. Non stupisce la presenza delle prime due feature GBVS ed ED, meglio prestanti quando considerate singolarmente, mentre il contributo di HL dimostra

l'importanza delle caratteristiche ad alto livello alla stima della salienza.

Infine, considerare una combinazione pesata in modo da attribuire maggior rilevanza alle *feature* meglio prestanti consente di ottenere un lieve miglioramento delle prestazioni.

Acronimi

AUC	Area Under the Curve
BMS	Boolean Map based Saliency
CC	Correlation Coefficient
CMP	Cube-Map Projection
CNN	Convolutional Neural Network
E	Entropy
ED	Edge Detection
EM	Eye-Movement
EMD	Earth Mover's Distance
ERP	Equirectangular Projection
FB	Foreground-Background Content
FCNN	Fully Convolutional Neural Networks
FOV	Field Of View
FSM	Fused Saliency Maps
GBVS	Graph Based Visual Saliency
GAN	Generative Adversarial Networks
H	Hue
HL	High-Level
HM	Head-Movement
HMD	Head-Mounted Display
HVS	Human Visual System
KLD	Kullback-Leibler Divergence

L	Luminance
LL	Low-Level
NSS	Normalized Scanpath Saliency
PO	Presence of Object
PP	Presence of People
ROC	Receiver Operating Characteristic
S	Saturation
SALICON	Saliency in Context
SDK	Software Development Kit

Bibliografia

[1] I. Djemai, S. A. Fezza, W. Hamidouche, and O. Déforges, "Extending 2D Saliency Models for Head Movement Prediction in 360-Degree Images using CNN-Based Fusion," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2020, pp. 1–5.

[2] Y. Rai, P. Le Callet, and P. Guillotel, "Which saliency weighting for omni directional image quality assessment?" in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, 2017, pp. 1–6.

[3] J. Gutiérrez, E. David, Y. Rai, and P. L. Callet, "Toolbox and dataset for the development of saliency and scanpath models for omnidirectional/360°still images," *Signal Processing: Image Communication*, vol. 69, pp. 35–42, 2018, salient360: Visual attention modeling for 360° Images. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0923596518304594

[4] A. Sendjasni and C. Larabi, "PW-360IQA: Perceptually-Weighted Multichannel CNN for Blind 360-Degree Image Quality Assessment," *Sensors (Basel, Switzerland)*, vol. 23, 04 2023.

[5] T. El-Ganainy, "Spatiotemporal rate adaptive tiled scheme for 360 sports events," *arXiv preprint arXiv:1705.04911*, 2017.

[6] M. Corbetta and G. Shulman, "Control of Goal-Directed and Stimulus-Driven Attention in the Brain," *Nature reviews. Neuroscience*, vol. 3, pp. 201–15, 04 2002.

[7] G. W. Lindsay, "Attention in psychology, neuroscience, and machine learning," *Frontiers in Computational Neuroscience*, vol. 14, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:215769557

[8] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[9] M. Xu, Y. Song, J. Wang, M. Qiao, L. Huo, and Z. Wang, "Predicting Head Movement in Panoramic Video: A Deep Reinforcement Learning Approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 11, pp. 2693–2708, 2019.

[10] M. Xu, C. Li, S. Zhang, and P. L. Callet, "State-of-the-Art in 360° Video/Image Processing: Perception, Assessment and Compression," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 5–26, 2020.

[11] A. D. Aladagli, E. Ekmekcioglu, D. Jarnikov, and A. Kondoz, "Predicting head trajectories in 360° virtual reality videos," in *2017 International Conference on 3D Immersion (IC3D)*, 2017, pp. 1–6.

[12] M. Assens, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor, "SaltiNet: Scan-Path Prediction on 360 Degree Images Using Saliency Volumes," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 2331–2338.

[13] Y. Xu, Y. Dong, J. Wu, Z. Sun, Z. Shi, J. Yu, and S. Gao, "Gaze Prediction in Dynamic 360° Immersive Videos," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5333–5342.

[14] A. Borji and L. Itti, "State-of-the-Art in Visual Attention Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[15] A. De Abreu, C. Ozcinar, and A. Smolic, "Look around you: Saliency maps for omnidirectional images in VR applications," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, 2017, pp. 1–6.

[16] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 262–270.

[17] P. Lebreton and A. Raake, "GBVS360, BMS360, ProSal: Extending existing saliency prediction models from 2D to omnidirectional images," *Signal Processing: Image Communication*, vol. 69, pp. 69–78, 2018, salient360: Visual attention modeling for 360° Images. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0923596518302406

[18] J. Zhang and S. Sclaroff, "Saliency Detection: A Boolean Map Approach," in *2013 IEEE International Conference on Computer Vision*, 2013, pp. 153–160.

[19] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency," *Adv. Neural Inform. Process. Syst*, vol. 19, pp. 545–552, 01 2006.

[20] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0010028580900055

[21] Y. Zhu, G. Zhai, X. Min, and J. Zhou, "The Prediction of Saliency Map for Head and Eye Movements in 360 Degree Images," *IEEE Transactions on Multimedia*, vol. 22, no. 9, pp. 2331–2344, 2020.

[22] W. Luo, H. Li, G. Liu, and K. Ngi Ngan, "Global salient information maximization for saliency detection," *Signal Processing: Image Communication*, vol. 27, no. 3, pp. 238–248, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0923596511001160

[23] B. Han, X. Li, X. Gao, and D. Tao, "A biological inspired features based saliency map," in *2012 International Conference on Computing, Networking and Communications (ICNC)*, 2012, pp. 371–375.

[24] F. Battisti, S. Baldoni, M. Brizzi, and M. Carli, "A feature-based approach for saliency estimation of omni-directional images," *Signal Processing: Image Communication*, vol. 69, pp. 53–59, 2018, salient360: Visual attention modeling for 360° Images. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092359651830242X

[25] E. Vig, M. Dorr, and D. Cox, "Large-Scale Optimization of Hierarchical Features for Saliency Prediction in Natural Images," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2798–2805.

[26] M. Kümmerer, L. Theis, and M. Bethge, "Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet," 2015.

[27] S. S. S. Kruthiventi, K. Ayush, and R. V. Babu, "DeepFix: A Fully Convolutional Neural Network for Predicting Human Eye Fixations," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4446–4456, 2017.

[28] J. Pan, C. C. Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. G. i Nieto, "SalGAN: Visual Saliency Prediction with Generative Adversarial Networks," 2018.

[29] R. Monroy, S. Lutz, T. Chalasani, and A. Smolic, "SalNet360: Saliency maps for omni-directional images with CNN," *Signal Processing: Image Communication*, vol. 69, pp. 26–34, 2018, salient360: Visual attention modeling for 360° Images. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0923596518304685

[30] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein, "Saliency in VR: How Do People Explore Virtual Environments?" *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 4, pp. 1633–1642, 2018.

[31] Y. Rai, J. Gutiérrez, and P. Le Callet, "A Dataset of Head and Eye Movements for 360 Degree Images," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, ser. MMSys'17. New York, NY, USA: Association for Computing Machinery, 2017, p. 205–210. [Online]. Available: https://doi.org/10.1145/3083187.3083218

[32] T. Judd, F. Durand, and A. Torralba, "A Benchmark of Computational Models of Saliency to Predict Human Fixations," MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY, Tech. Rep., 01 2012.

[33] A. Borji and L. Itti, "CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research," 2015.

[34] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in Context," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[35] N. Bruce and J. Tsotsos, "Saliency Based on Information Maximization," in *Advances in Neural Information Processing Systems*, Y. Weiss, B. Schölkopf, and J. Platt, Eds., vol. 18. MIT Press, 2005. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2005/file/0738069b244a1c43c83112b735140a16-Paper.pdf

[36] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "Saliency Benchmarking Made Easy: Separating Models, Maps and Metrics," in *Computer Vision – ECCV 2018*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Springer International Publishing, 2018, pp. 798–814.

[37] Q. Xu, F. Wang, Y. Gong, Z. Wang, K. Zeng, Q. Li, and X. Luo, "A novel edge-oriented framework for saliency detection enhancement," *Image and Vision Computing*, vol. 87, pp. 1–12, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885619300460

[38] D. Ziou, S. Tabbone *et al.*, "Edge detection techniques-an overview," *Pattern Recognition and Image Analysis C/C of Raspoznavaniye Obrazov I Analiz Izobrazhenii*, vol. 8, pp. 537–559, 1998.

[39] P. Mazumdar, G. Arru, M. Carli, and F. Battisti, "Face-aware Saliency Estimation Model for 360° Images," in *2019 27th European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[40] V. Mazza, M. Turatto, and C. Umiltà, "Foreground?background segmentation and attention: A change blindness study," *Psychological research*, vol. 69, pp. 201–10, 02 2005.

[41] P. Mazumdar, K. Lamichhane, M. Carli, and F. Battisti, "A Feature Integrated Saliency Estimation Model for Omnidirectional Immersive Images," *Electronics*, vol. 8, no. 12, 2019. [Online]. Available: https://www.mdpi.com/2079-9292/8/12/1538

[42] G. W. Lindsay, "Attention in psychology, neuroscience, and machine learning," *Frontiers in computational neuroscience*, vol. 14, p. 29, 2020.

[43] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[44] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common Objects in Context," 2014, cite arxiv:1405.0312Comment: 1) updated annotation pipeline description and figures; 2) added new section describing datasets splits; 3) updated author list. [Online]. Available: http://arxiv.org/abs/1405.0312

[45] P. Hu and D. Ramanan, "Finding Tiny Faces," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1522–1530.

[46] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A Face Detection Benchmark," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5525–5533.