Scuola di Scienze Matematiche, Fisiche e Naturali
Corso di Laurea in Informatica

Tesi di Laurea

METODI DI APPRENDIMENTO AUTOMATICO PER LA FUSIONE DI IMMAGINI SAR DESPECKLED

MACHINE LEARNING METHODS FOR THE FUSION OF DESPECKLED SAR IMAGES

YOUNESS AHARRAM

Relatore: *Daniele Baracchi*

Correlatori: *Fabrizio Argenti, Tommaso Pecorella*

Anno Accademico 2024-2025

CONTENTS

LIST OF FIGURES

"Insert citation"
— Insert citation's author

1

INTRODUZIONE

I satelliti SAR sono satelliti dotati di un radar ad apertura sintetica che permette loro di acquisire immagini della superficie terrestre indipendentemente dalle condizioni meteorologiche e dalla luce solare. Missioni e piattaforme di riferimento includono Sentinel-1 (ESA) [5], TerraSAR-X (DLR), COSMO-SkyMed (ASI) e RADARSAT. I satelliti SAR, grazie a questa loro capacita, trovano applicazione in molteplici contesti. In ambito geologico [4], sono impiegati per il monitoraggio del suolo e dei processi geomorfologici, consentendo la mappatura di foreste, deserti e aree soggette a trasformazioni ambientali. Inoltre, risultano particolarmente efficaci nell'analisi dei fenomeni di deforestazione attraverso il rilevamento dei cambiamenti nella copertura boschiva. Marittimo, permettono di localizzare navi anche in condizioni meteorologiche avverse e di rilevare sversamenti di petrolio o altre sostanze inquinanti. Infrastrutture e urbanistica, vengono utilizzati per misurare gli spostamenti del terreno e delle aree urbane, oltre che per il controllo di dighe, ponti e ferrovie, e per l'osservazione dello sviluppo delle città. L'immagine generata dal satellite però presenta un particolare tipo di rumore. Quest'ultimo si forma quando un impulso radar colpisce il terreno, questo non riflette semplicemente un segnale uniforme. In realtà, il segnale viene riflesso da moltissimi piccoli scatter presenti sulla superficie come foglie, rocce o edifici. Tutti questi ritorni interferiscono tra di loro, sommando le onde con fasi diverse. Il risultato di questa interferenza prende il nome di Speckle. Questo tipo di rumore non è un errore del satellite o del radar, ma una caratteristica intrinseca del tipo di misura e si presenta con un pattern granuloso che rende l'immagine difficile da interpretare ed analizzare. Il processo di riduzione dello speckle prende il nome di despeckling [13]. Quest'ultimo cerca di smussare o filtrare il rumore granulare senza però perdere le informazioni reali presenti nell'immagine. In letteratura vi sono molteplici approcci: alcui si basano su filtri spaziali che analizzano i pixel vicini, altri usano tecniche più sofisticate come metodi di deep

learning. Ogni approccio ha i suoi punti di forza e le sue lacune sulla base del tipo di ambiente rappresentato nell'immagine. Lo scopo di questa tesi è cercare di unire i punti di forza di alcuni modelli in modo da ottenere l'immagine con il despeckling più accurato possibile. Un primo approccio per ottenere ciò consiste nell' utlizzare tecniche di machine learning per predire la qualità di un immagine denoised attraverso una mappa di qualità. Ad ogni modello, è associata una mappa che indica, pixel per pixel dove il modello ha funzionato meglio e dove invece peggio. La fusione avviene tramite la media pesata dove i relativi pesi sono le mappe di qualità. Questo approccio però non sfrutta al massimo i punti di forza di ogni singola immagine despeckled portando ad un risulato finale non soddisfacente, in quanto la qualità del denoising viene stimata concentrandosi sul singolo pixel senza guardare i vicini. Un secondo approccio più efficiente è basato sull'attenzione. Invece di utilizzare mappe di qualità che determinano la bontà del denoising di un singolo pixel, si utilizzano meccanismi basati sulla self e cross attention. Questo consente di andare oltre la valutazione locale pixel per pixel, mettendo in relazione l'informazione proveniente da più immagini despeckled e valorizzando i dettagli complementari.

2

STATE OF THE ART

Il funzionamento di questo tipo di satellite come spiagato nel sito della *N.A.S.A.* [3] si basa sull'uso di onde radar che vengono inviate verso la Terra. Questi impulsi elettromagnetici rimbalzano sul terreno e sugli oggetti come edifici o vegetazione e tornano al satellite. Quest'ultimo analizzando il segnale di ritorno riesce ad ottenere informazioni sia sull'intensità del riflesso sia sul tempo impiegato dal segnale per tornare, dati fondamentali per ricostruire l'immagine del territorio. Il punto di forza del SAR è l'apertura sintetica. Poichè il satellite si muove lungo la sua orbita, i segnali raccolti in posizioni diverse vengono combinati insieme. Questo processo permette di simulare un'antenna molto più grande di quella reale, ottenendo così immagini ad altissima risoluzione, molto piu dettagliate di quelle che un radar di dimensioni fisiche limitate potrebbe generare da solo. In pratica, il movimento del satellite trasforma un radar relativamente piccolo in uno strumento potentissimo per osservare il pianeta. A seguito della cattura della scena di interesse, l'immagine ottenuta risulta disallineata, in quanto si genera un angolo θ tra l'asse del satellite e la superficie terrestre. Per correggere questo effetto, si applica una tecnica di allineamento dell'immagine [6]. L'immagine generata dal satellite però presenta un rumore che prende il nome di speckle. Il fenomeno dello speckle è una caratteristica intrinseca delle immagini acquisite da sensori coerenti, come i radar ad apertura sintetica (SAR), i sistemi laser o gli interferometri ottici. Dal punto di vista fisico, lo speckle nasce dall'interferenza coerente tra le onde elettromagnetiche riflesse da molteplici scatterer presenti all'interno di una singola cella di risoluzione del sensore. Ciascuno di questi scatterer contribuisce con un segnale complesso avente ampiezza e fase proprie; la somma coerente di tali contributi produce una risultante la cui ampiezza varia casualmente nel tempo e nello spazio.

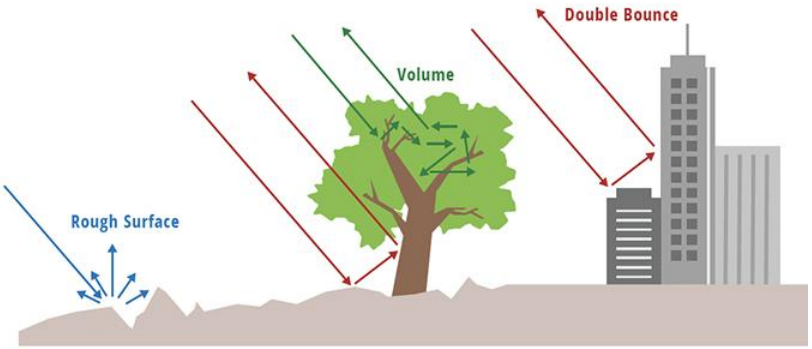

Figure 1: Impulsi radar inviati dal satellite SAR verso la Terra e riflessi indietro.

Questo effetto di interferenza, costruttiva o distruttiva, genera un pattern granulare nell'immagine osservata, noto appunto come speckle. Quest'ultimo degrada la qualità visiva e radiometrica dell'immagine, rendendo più complessa l'analisi e l'interpretazione dei dati. Le immagini ottenute da sistemi radar ad apertura sintetica (SAR) sono caratterizzate da un disturbo granulare denominato speckle. Questo fenomeno è intrinseco a tutti i sistemi di imaging coerente e rappresenta un rumore di natura moltiplicativa, dipendente dal segnale, che degrada l'aspetto visivo e le prestazioni dei processi automatici di analisi dell'immagine, come classificazione o rilevamento di cambiamenti.

2.1 ORIGINE FISICA DELLO SPECKLE

Il sensore SAR emette un'onda elettromagnetica e riceve il segnale retrodiffuso da una piccola area della scena, detta *cella di risoluzione*. Tale cella contiene numerosi scatterer elementari, ciascuno caratterizzato da una propria ampiezza A_i e fase ϕ_i. Il segnale complesso ricevuto può essere modellato come somma coerente dei contributi di tutti gli scatterer presenti nella cella [1]:

$$z = \sum_{i=1}^{N} A_i e^{j\phi_i}. \tag{1}$$

Le fasi ϕ_i variano casualmente a causa delle differenze di cammino ottico e, per un numero sufficientemente grande di scatterer indipendenti, il teorema del limite centrale implica che le componenti reale e immaginaria di z sono variabili gaussiane a media nulla e varianza $\sigma^2/2$. In tal caso, si parla di *speckle pienamente sviluppato* (*fully developed speckle*).

2.2 MODELLO STATISTICO

Il modulo del segnale complesso $A = |z|$ segue una distribuzione di Rayleigh:

$$p_A(A) = \frac{A}{\sigma^2} e^{-A^2/(2\sigma^2)} \tag{2}$$

mentre l'intensità $I = A^2$ segue una distribuzione esponenziale:

$$p_I(I) = \frac{1}{\sigma^2} e^{-I/\sigma^2} \tag{3}$$

Il valore medio dell'intensità è $\mathbb{E}[I] = \sigma^2$, che rappresenta la riflettività media del bersaglio o Radar Cross Section (RCS). Si ottiene così il cosiddetto modello moltiplicativo del rumore:

$$I = \mu u \tag{4}$$

dove μ è la riflettività media (informazione utile) e u è una variabile casuale esponenziale a media unitaria che rappresenta il rumore speckle. La varianza di I è proporzionale al quadrato della media, quindi i pixel più luminosi sono affetti da un disturbo più forte. Per ridurre la varianza, si può effettuare una media su L osservazioni indipendenti (tecnica detta *multilooking*), ottenendo una distribuzione $\mathcal{C}$ per l'intensità media I_L, con varianza ridotta a σ^2/L.

2.3 EFFETTI E PROBLEMATICHE

Lo speckle riduce il rapporto segnale rumore (SNR) e ostacola l'estrazione di informazioni affidabili da immagini SAR. Inoltre, poiché la sua intensità è proporzionale al segnale stesso, il rumore è più evidente nelle aree ad alta riflettività. La sua natura coerente e moltiplicativa richiede quindi modelli e tecniche di riduzione specifiche, diverse dai filtri per rumore additivo.

2.4 TECNICHE DI DESPECKLING

Negli ultimi trant'anni sono stati proposti numerosi metodi per la riduzione dello speckle nelle immagini SAR. I primi approcci sfruttano filti spaziali

come Lee, Frost e Kuan [9]. Questi operavano direttamente nel dominio dell'immagine, cioè sui pixel, sfruttando finestre locali per stimare statisticamente il rumore e ridurlo. Erano strumenti semplici, poco costosi dal punto di vista computazionale ed efficaci ma soffrivano di un limite strutturale. Per attenuare lo speckle tendevano a smussare anche i dettagli fini, specialmente lungo i bordi o nelle aree eterogenee. Con lo sviluppo della teoria delle trasformate multisensoriale negli anni Novanta, si passò ad un approccio diverso. Invece di agire direttamente sul'immagine, si inziò a trasformarla in un dominio in cui il segnale e il rumore potessero essere seprati. Nascono così i metodi basati su trasformata, come quelli che usano wavelet [2]. Questi strumenti rappresentano un'evoluzione concettuale dei filtri spaziali, perchè superano alcune loro debolezze: riescono a distinguere meglio il rumore dalle strutture significative, ad adattarsi a diverse scale ed a preservare in maniera più accurata bordi, texture e linee sottili. Tuttavia, portano con sè una maggiore complessità computazionale e la possibilità di introdurre artefatti se non calibrati con attenzione. Infine dato che lo speckle è un rumore moltiplicativo e non semplicemente additivo, se non viene trasformato prima, la wavelet può non essere del tutto efficace [1]. Negli ultimi anni, l'attenzione si è spostata verso metodi non locali, come il filtro Non-Local Means e l'algoritmo SARBM3D, adattati specificamente per il despeckling di immagini SAR. Qui l'idea è radicalmente diversa, ovvero non ci si limita più a gurdare in un introno locale del pixel, ma si cercano nel resto dell'immagine regioni simili e si usano queste corrispondenze per ridurre il rumore. In questo modo lo speckle viene attenuato in maniera molto efficace, mentre i dettagli strutturali si preservao quasi intatti. La qualità delle immagini risultanti è generalmente superiore a quella ottenuta con filtri locali o multirisoluzione, ciò comporta però un costo computazionale elevato e la necessità di algoritmi sofisticati per gestire le similitudini tra regioni. Negli ultimi dieci anni si è aperta una nuova fase, spinta dall'esplosione del deep learning [7]. L'idea è che le reti neurali, in particolare convoluzionalil o basate su autoencoder, possano imparare direttamente dai dati le caratteristiche dello speckle e il modo migliore per ridurlo. Questo approccio non si basa più nell'assumere una distribuzione statistica del rumore o una struttura matematica da preservare, ma si affida alla capacità della rete di apprendere autonomamente dalle coppie di immagini rumorose e pulite. I risultati hanno portato ad una qualità visiva migliore e un eccellente preservazione dei dettagli. D'altro canto, le reti neurali hanno bisogno di grandi quantità di dati ben calibrati per l'addestramento e possono soffrire di scarsa generalizzazione se

applicate a scenari diversi da quelli su cui sono state addestrate oltre che ad un costo computazionale molto elevato. Le performance dei modelli di despeckling non è uniforme per tutti i tipi di scenari. La loro efficacia può variare in base alle caratteristiche statistiche del bioma come contesti di vegetazione, aree rocciose e urbane, poichè la distribuzione del rumore e le strutture da preservare differiscono sensibilmente. Un'immagine SAR potrebbe comprendere due o più tipi di biomi, ciò implica che utlizzando un unico modello di despeckling, indipendetemente da quale esso sia, l'immagine risultante avrà aree in cui è stata ripulita meglio e aree in cui è stata ripulita peggio a seconda di dove il modello per come è stato realizzato ha più facilità ad operare. L'idea da cui nasce questa tesi è quello di unire le caratteristiche migliori di determinate tecniche di despeckling, in modo tale che l'immagine risultante rispecchi il più possibile la realtà di interesse. Questo tipo di approccio non va a reinventare la ruota, cioè non punta a realizzare un nuovo modello con cui è possibile fare despeckling, ma è mirato a sfruttare i punti di forza di modelli già esistenti. Inizialmente è stata usata una tecnica naive che prevede un architettura Unet per addestare tanti modelli quante sono le tecniche di despeckling della quale si vuole imparare a prevedere la qualità del denoising generando così mappe di qualità che determinano quanto dell'immagine denoised di un modello prendere in relazione alla bontà del denoising. Tuttavia, questa tecnica non riesce a sfruttare appieno le potenzialità dei singoli modelli, poiché la stima della qualità è effettuata a livello locale e si concentra sul singolo pixel, trascurando così informazioni contestuali di più ampia scala. Inoltre, la combinazione pesata pixel-per-pixel non permette di cogliere la complementarità delle caratteristiche tra i diversi metodi a livello di patch, limitando la capacità del sistema di valorizzare i punti di forza specifici di ciascun modello. L'approccio proposto in [8], invece, supera tali limiti introducendo meccanismi di attenzione incrociata, che consentono una fusione più coerente e informata tra le diverse rappresentazioni. Gli approcci basati su questa tecnica propongono esattamente questa linea di azione: invece di pesare singoli pixel, si estraggono rappresentazioni (feature) dai diversi output despeckled e si usa un modulo di attenzione per selezionare, a livello di feature e di contesto, quali informazioni preservare e quali attenuare. Questo approccio permette di superare i limiti dell' approccio pixel per pixel, in quanto una fusione basata su patch consente di catturare informazioni contestuali e di valorizzare le relazioni strutturali presenti nell'immagine. Operando su blocchi di feature e non su singoli pixel isolati, il modello è in grado di preservare meglio i bordi e le discontinuità, la coerenza

spaziale della patch riduce il rischio di smussare i contorni netti, tipico delle fusioni locali.

3

APPROCCIO BASATO SU RETE UNET

Il primo approccio divide il problema in due parti. La prima è relativa alla previsine della qualità tramite tecniche di machine learning. La seconda parte invece riguarda la fusione delle immagini denoised di ogni modello.

3.1 DATASET

Come illustrato in Figura 2, il dataset impiegato è costituito da tre tipologie di immagini: clean, noisy e denoised tutte e tre ad un canale, ovvero in scala di grigi. Queste immagini sono state realizzate tramite uno strumento ottico di un satellite per poi essere sporcate con speckle artificiale e su cui infine è stato fatto denoising. Questa scelta è stata fatta in quanto risulta difficile reperire un dataset contenente immagini SAR abbastanza grande da poter essere utlizzato per addestrare una rete neurale. Le immagini su cui viene fatto l'addestramento sono relative ad un singolo modello di despeckling ed ad un determinato look. Quest'ultimo è una metrica che indica l'intensità dello speckle artificiale, in quanto a più look, corrispondono più catture di quella che è la realtà di interesse e quindi si ha una maggior precisione e uno speckle ridotto. Le immagini prima di essere passate alla rete vengono normalizzate nell'intervallo $[0, 1]$. In fase di addestramento, le immagini noisy e denoised vengono concatenate in un unico tensore a due dimensioni e utilizzate come input per la rete neurale. Le immagini clean $[\hat{I}]$, invece, assieme a quelle denoised $[I]$, vengono impiegate per la generazione della mappa di qualità $[QM]$, che costituisce il riferimento necessario per il calcolo della funzione di perdita. Questa mappa viene generata facendo la diffrenze in valore assoluto delle due immagini:

$$QM = |\hat{I} - I| \tag{5}$$

Questo permette al modello di imparare a prevedere la qualità del denoising relative ad un dato modello di despeckling e ad una determinata intensità dello speckle.

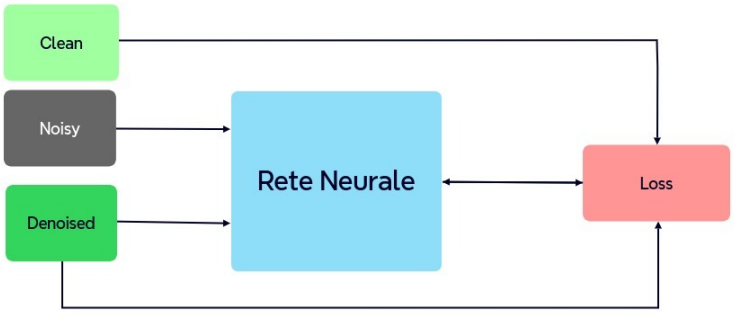

Figure 2: Mock della struttura logica per la previsione della qualità

3.2 ARCHITETTURA RETE NEURALE

Come architettura della rete neurale è stata utilizzata la Unet. Questo perchè risulta particolarmente adatto per la generazione di una mappa di qualità che valuti l'affidabilità di un'immagine sottoposta a denoising [12]. che trattano dell'argomento. Il compito di questa rete è assegnare a ciascun pixel un valore compreso nell'intervallo $[0, 1]$ che ne rappresenti la qualità locale. Questo output per la sua natura richiede un'architettura in grado di produrre una mappa di densità della stessa risoluzione spaziale dell'input, preservando la localizzazione precise delle informazioni. L'architettura [10], è composta da due parti principali: encoder e decoder. Encoder, ovvero una sequenza di blocchi convoluzionali e operazioni di pooling che estraggono feature gerarchiche via via più astratte, consentendo alla rete di catturare il contenuto semantico dell'immagine, distinguere texture, bordi e regioni omogenee, e modellare la natura statistica del rumore. Decoder, una fase simmetrica che ricostruisce progressivamente la risoluzione spaziale mediante up-sampling e convoluzioni, trasformando le feature ad alto livello in una mappa di qualità densa e dettagliata. Elemento distintivo della U-Net sono le skip connections, che collegano i livelli dell'encoder ai corrispondenti livelli del decoder. Questi collegamenti trasferiscono mappe di feature ad alta risoluzione direttamente al decoder, preservando l'informazione spaziale fine indispensabile per localizzare correttamente dettagli critici come bordi sottili

e piccole texture. Senza tali connessioni, il decoder produrrebbe output più sfocati, perdendo la capacità di discriminare le aree più sensibili agli artefatti di oversmoothing o agli errori di ricostruzione.

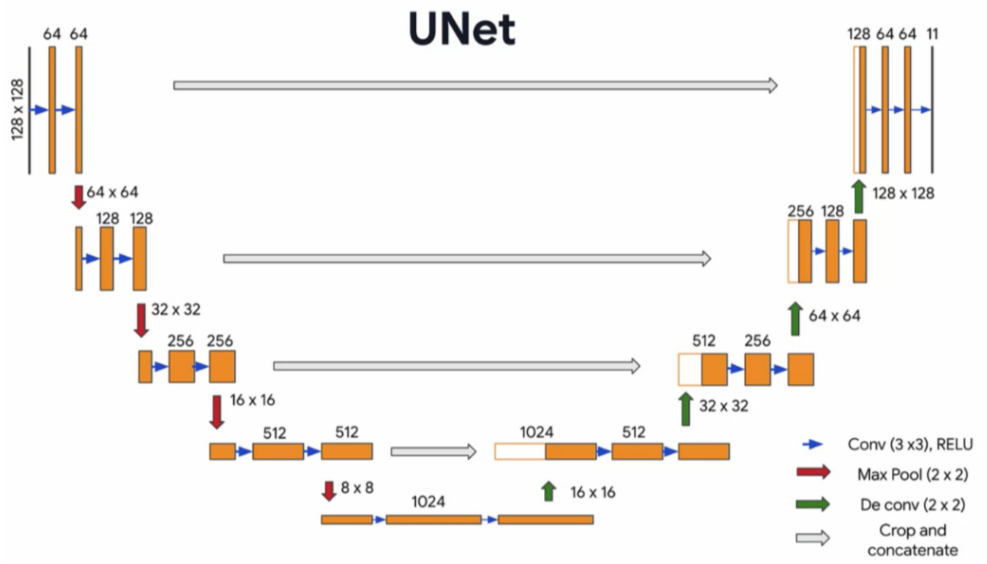

Figure 3: Architettura Unet

L'aspetto più importante di questo modello è la sua capacità di andare oltre una semplice misura di errore pixel-wise. La rete apprende a sviluppare una vera e propria percezione semantica dell'errore, cioè a valutare la gravità di una discrepanza in funzione del contesto visivo in cui compare. Ad esempio, un errore di grande entità in una regione uniforme (come un cielo sereno) è visivamente più disturbante di un errore di pari entità in una regione altamente texturizzata (come del fogliame), dove può essere facilmente mascherato. Analogamente, un piccolo errore su un bordo netto è percettivamente rilevante. L'encoder, stratificando feature di complessità crescente, apprende questa gerarchia di contenuti visivi e fornisce al decoder le informazioni necessarie per produrre una mappa di qualità che pondera ogni errore in base alla sua importanza percettiva.

3.3 FUSIONE DEI MODELLI

La fusione delle immagini denoised [I] prodotte tramite i rispettivi modelli di despeckling vengono fusi attraverso una media pesata sfuttando le mappa di qualità [M] come pesi per le singole immagini.

$$\frac{\sum_{i=1}^{n} I_i M_i}{\sum_{i=1}^{n} M_i} \tag{6}$$

È un processo che si comporta in modo diverso per ogni singolo pixel dell'immagine. Se il Modello A ha una qualità stimata molto alta in una regione e il Modello B molto bassa, la fusione privilegerà maggiormente il Modello A in quella determinata area. Mentre alcuni sono eccellenti nel preservare i bordi netti ma possono lasciare del rumore residuo nelle aree lisce, altri sono ottimi nell' eliminare il rumore dalle superfici lisce ma tendono a sfocare (oversmooth) i bordi e le texture. La fusione pesata permette di prendere il meglio di ogni modello usando di più l'output dell'algoritmo migliore per una data caratteristica dell'immagine, e meno i suoi contributi peggiori.

3.4 FUNZIONE DI LOSS

Come funzione di perdità è stata utilizzata la mean square error [MSE] tra l'output della fusione [I] e il ground truth [$\hat{I}$], ovvero l'immagine clean
.

$$\text{MSE} = \left(I - \hat{I}\right)^2 \tag{7}$$

Questa funzione di perdità è stata scelta in quanto è molto facile da calcolare eccellenti non presenta particolari complicazioni in fase di ottimizzazzione. Inoltre dato che PSNR è funzione del MSE, ottimizzare la MSE porta anche ad avere valori più alti di PSNR, che spesso viene usato come metrica di valutazione

3.5 LIMITI DEI QUESTA TECNICA

Sebbene la media pesata presenti il vantaggio di combinare in maniera coerente i contributi dei vari modelli di despeckling, essa introduce alcune criticità che ne riducono l'efficacia in scenari complessi. In primo luogo, la media pesata tende a diluire i dettagli sottili. Se una delle immagini denoised contiene strutture fini o bordi ben preservati che altre non hanno, questi dettagli possono risultare attenuati o addirittura persi, poiché la media li combina con le versioni più lisce prodotte dagli altri modelli. Ciò porta a un effetto di oversmoothing, che riduce il livello di dettaglio complessivo dell'immagine fusa. Un'ulteriore limitazione deriva dal fatto che la media pesata opera pixel per pixel, senza considerare la correlazione spaziale tra pixel adiacenti. Pattern, texture e strutture complesse che sono distribuite su più pixel non vengono trattati in maniera coerente. Questo può portare a artefatti locali, soprattutto ai bordi o in aree con pattern ripetitivi, dove una decisione puramente puntuale può introdurre discontinuità. Inoltre, se il rumore è altamente non stazionario o presenta componenti strutturate, la rete che genera le mappe di qualità può faticare a distinguere tra rumore residuo e dettaglio fine. In queste situazioni la mappa di qualità può risultare inaccurata, assegnando un peso elevato a regioni che in realtà presentano artefatti o penalizzando aree visivamente corrette. Questo porta a una fusione non ottimale, che può accentuare il rumore residuo o degradare regioni già ben restaurate. Va considerato anche il problema della sensibilità agli errori di predizione della mappa di qualità. Poiché i pesi influenzano direttamente il risultato finale, eventuali errori nella stima della qualità si traducono in artefatti amplificati nell'immagine fusa, specialmente se un singolo modello viene sovrastimato in regioni dove la sua qualità non è affidabile.

4

APPROCCIO BASATO SULLA SELF E CROSS ATTENTION

Usare un approccio basato sulla self e cross attention permette di effettuare la fusione non basandosi su una strategia pixel per pixel ma su patch, ovvero una piccola regione localizzata o un sottoinsieme di pixel all'interno di un'immagine più grande, in genere rettangolare o quadrata. L'obiettivo del meccanismo di fusione basato su self e cross attention è combinare in modo intelligente più immagini despeckled provenienti da diverse acquisizioni o canali per ottenere un'immagine finale più informativa, pulita e coerente. Per affrontare questa problematica, l'approccio proposto in [8] introduce un meccanismo di fusione basato su self e cross attention, ispirato ai modelli transformer, in grado di individuare e valorizzare le componenti informative complementari tra le immagini despeckled, riducendo al contempo la ridondanza residua. Il paper CrossFuse: A Novel Cross Attention Mechanism based Infrared and Visible Image Fusion Approach mostra che una fusione che usa blocchi di self-attention per rinforzare le caratteristiche intra-modalità e una cross-attention progettata per esaltare le informazioni non correlate tra le modalità, produce immagini fuse con più dettaglio e con meno artefatti, migliorando le strutture rispetto a metodi più semplici. Il paper sottolinea inoltre che, in multimodal fusion, la cosa cruciale è valorizzare l'uncorrelation (cioè la complementarità) tra le modalità, cosa che la cross-attention è progettata per fare. Questo permette di combinare informazioni provenienti da regioni diverse e di differenti modalità, non solo di sommare pixel con pesi locali. Il metodo si fonda su una architettura ibrida composta da due blocchi principali.

4.0.1 *Self-Attention*

La self-attention è un meccanismo introdotto originariamente nei Transformer per consentire a una rete di mettere in relazione diverse parti dello

stesso input tra loro, pesandole in base alla loro importanza reciproca. Questo è diverso dalle convoluzioni (CNN) tradizionali, che analizzano solo piccole regioni locali, la self-attention, invece, cattura dipendenze globali, anche tra punti molto distanti dell'immagine.

Funzionamento della Self-Attention

Sia un'immagine despeckled $I \in \mathbb{R}^{H \times W \times C}$, dove ogni patch può essere considerato come un *token*. Per ogni posizione i nell'immagine:

- **Query** Q_i: rappresenta ciò che la posizione i sta cercando negli altri patch della stessa immagine.
- **Key** K_j: rappresenta il contenuto informativo della posizione j rispetto agli altri patch.
- **Value** V_j: è l'informazione effettiva che la posizione j può trasmettere a i.

Il pixel i può "guardare" altri pixel dell'immagine e decidere quali informazioni (ad esempio strutture, bordi, texture) prendere per migliorare la propria rappresentazione despeckled.

4.0.2 *Calcolo della Self-Attention*

Il primo passo consiste [11] nel calcolare la similarità tra ogni Query della target e ogni Key della source, mediante un prodotto scalare:

$$Q \cdot K^T$$

Questo produce una matrice di dimensione $n x m$ in cui n è il numero di elementi della target e m quello della source. Ogni entry misura quanto un elemento della target "presta attenzione" a un elemento della source. Per stabilizzare la scala dei valori e prevenire problemi numerici, si divide per $\sqrt{d_k}$, ottenendo:

$$\frac{QK^T}{\sqrt{d_k}}$$

Successivamente, applichiamo la funzione softmax riga per riga, trasformando i punteggi in probabilità normalizzate. In questo modo, per ogni

elemento della target otteniamo un insieme di pesi che indicano quanto ciascun elemento della source contribuisce alla rappresentazione finale:

$$\alpha = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

Infine, la matrice dei pesi α viene moltiplicata per la matrice dei Value V, combinando le informazioni della source in modo ponderato e producendo i nuovi vettori rappresentativi per la target:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{8}$$

4.0.3 *Cross-Attention*

La cross-attention [11] è un meccanismo di attenzione che permette a una sequenza di query provenienti da una sorgente (ad esempio un decoder) di pesare gli elementi di un'altra sequenza, da cui provengono le key e le value (ad esempio l'encoder). In altre parole, essa modella le dipendenze incrociate tra due insiemi distinti di rappresentazioni. L'attenzione viene calcolata come visto in 8 con la differenza che nella cross-attention le matrici Q, K e V provengono da sorgenti diverse.

- **Query** Q: proiezione lineare delle rappresentazioni del decoder
- **Key** K, **Value** V: proiezioni lineari delle rappresentazioni del encoder

Nel paper CrossFuse: A Novel Cross Attention Mechanism based Infrared and Visible Image Fusion Approach [8], la cross attention è il cuore del metodo di fusione. Il suo scopo principale non è, come nei transformer classici, massimizzare la correlazione tra due insiemi di feature, ma enfatizzare le informazioni complementari (cioè non correlate) tra immagini visibile e infrarossa.

4.0.4 *Architettura del modello*

Due encoder (con la stessa struttura ma parametri diversi) estraggono le feature rispettivamente dall'immagine infrarossa (IR) e visibile (VI), servono per catturare le caratteristiche specifiche di ciascuna modalità. Le feature di ciascun encoder passano prima attraverso blocchi di

self-attention (SA) per migliorare la coerenza intra-modale (dettagli e struttura interna all'immagine). Poi interviene il Cross-Attention Mechanism (CAM), dove avviene la fusione vera e propria tra le due modalità. Il decoder ricostruisce l'immagine fusa a partire dalle feature integrate dal CAM, con skip connection per preservare dettagli e salienza.

Cross-Attention Mechanism

Il CAM combina self-attention e cross-attention in modo da: potenziare le intra-feature di ciascuna modalità; evidenziare le inter-feature complementari tra IR e VI. Ogni modalità entra in una catena di blocchi:

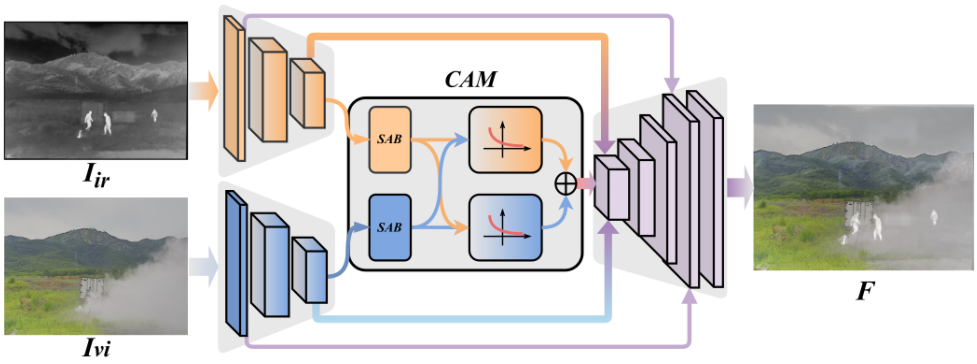

Figure 4: I due Encoder hanno la stessa architettura ma parametri differenti. Il meccanismo di cross-attention (CAM) viene utilizzato per fondere le caratteristiche multimodali. "SAB" indica il blocco di self-attention. L'immagine fusa può essere ottenuta tramite il Decoder, che include una connessione lunga proveniente dagli encoder.

I due blocchi di Self-Attention (SA) servono a rafforzare le caratteristiche interne. L'operazione di Shift/Unshift sposta e ripristina le feature, per aumentare la copertura spaziale globale. Il Cross-Attention (CA) è il passo cruciale, che integra le due modalità.

Reversed Softmax

Il cross-attention è calcolato come nel transformer standard, ma con una differenza fondamentale:

$$\text{re-softmax}(X) = \text{softmax}(-X)$$

Questo significa che: invece di dare peso alto alle feature simili (correlate) come fa il softmax classico, la re-softmax dà peso alto alle feature dissimili

(non correlate). CAM enfatizza ciò che una modalità ha e l'altra no, ossia l'informazione complementare, che è essenziale per la fusione

4.0.5 *Fusione dei modelli*

Nel caso della fusione si hanno quattro autoencoder, ciascuno dedicato a una diversa rappresentazione dell'immagine SAR (SAR-CAM, FANS, SARBM3D, noisy). Ciò estende il principio di CrossFuse da una fusione bimodale a una fusione multimodale. L'idea alla base rimane la stessa: combinare più sorgenti che condividono la stessa struttura spaziale ma presentano contenuti informativi parzialmente diversi o complementari, al fine di produrre un'immagine finale più completa, bilanciata e informativa. Il punto di partenza del sistema è costituito dai quattro encoder, ognuno dei quali apprende a rappresentare la propria modalità secondo il suo dominio specifico. Le versioni despeckled prodotte con SARCAM, FANS e BM3D, pur avendo eliminato il rumore di tipo speckle, differiscono nel modo in cui trattano i dettagli fini e le strutture deboli: alcune tendono a privilegiare la continuità tonale, altre la preservazione dei bordi. L'immagine noisy, invece, pur essendo la più "sporca", conserva informazione strutturale che spesso viene attenuata durante il despeckling. In questo contesto, gli encoder agiscono come estrattori di caratteristiche complementari: ciascuno mappa la propria immagine in uno spazio di rappresentazione latente dove si preservano sia i pattern condivisi (ad esempio, i contorni principali) sia le particolarità proprie del metodo di despeckling. Una volta estratte le feature dalle quattro reti, la fusione non può limitarsi a un semplice concatenamento o media ponderata. È qui che entra in gioco la cross-attention. Invece di gestire due rami (come nel caso infrarosso-visibile), il modello deve essere capace di trattare interazioni multiple, valutando quanto ogni rappresentazione contribuisca alla formazione di un contenuto informativo unico. Ogni coppia di modalità può essere posta in relazione attraverso una cross-attention, dove la re-softmax viene applicata per enfatizzare la dissimilarità: in questo modo le componenti ridondanti, cioè le parti in cui le varie versioni coincidono, vengono attenuate, mentre le parti complementari, presenti solo in uno dei canali, vengono esaltate. Il passo successivo consiste nel combinare queste diverse mappe di attenzione in una rappresentazione fusa. Questo può essere fatto in modo gerarchico, ad esempio con una attention aggregation layer, che raccoglie i risultati delle cross-attention tra le varie coppie di encoder e li integra progressivamente. In questo modo si costruisce un'unica mappa di caratteristiche

latenti che racchiude i dettagli fini dei vari modelli, e riduce al minimo la ridondanza informativa. Il decoder, a questo punto, ha il compito di ricostruire l'immagine finale partendo da questa rappresentazione fusa. Dal punto di vista concettuale, questa architettura si comporta come un "mediatore intelligente" tra diverse visioni dello stesso segnale: non sceglie a priori quale metodo di despeckling sia migliore, ma apprende a ponderare dinamicamente l'informazione proveniente da ciascuno in base alla sua unicità. La cross-attention, con la re-softmax, garantisce che il modello privilegi la diversità informativa invece della somiglianza, rendendo possibile una fusione realmente complementare e non una semplice media delle soluzioni.

4.1 SEMPLIFICAZIONE DEL MODELLO CROSSFUSE

Il modello CrossFuse nasce con l'obiettivo di affrontare un problema tipico della fusione multimodale tra immagini provenienti da domini diversi, in particolare ottico e infrarosso. In questi casi, i due sensori catturano informazioni complementari: l'immagine visibile contiene ricchezza di texture e colore, mentre quella infrarossa evidenzia le sorgenti di calore e le strutture non visibili a occhio nudo. Il compito della rete è quindi combinare tali informazioni massimizzando la complementarità e minimizzando la ridondanza tra le due modalità. Durante le spierimentazioni della tesi pero si è notato che, tuttavia, se consideriamo il caso di immagini provenienti dallo stesso dominio, come due immagini ottiche acquisite in condizioni diverse o due versioni degradate di uno stesso contenuto, la situazione cambia profondamente. In questo scenario, le informazioni di partenza non sono complementari ma ridondanti: la sfida non è più quella di mettere in evidenza le differenze tra domini, ma piuttosto di fondere in modo coerente informazioni simili, eliminando il rumore e preservando i dettagli condivisi. Di conseguenza, molti dei meccanismi complessi introdotti in CrossFuse, come la doppia attenzione incrociata, il re-softmax o la doppia fase di addestramento, possono risultare superflui o addirittura controproducenti.

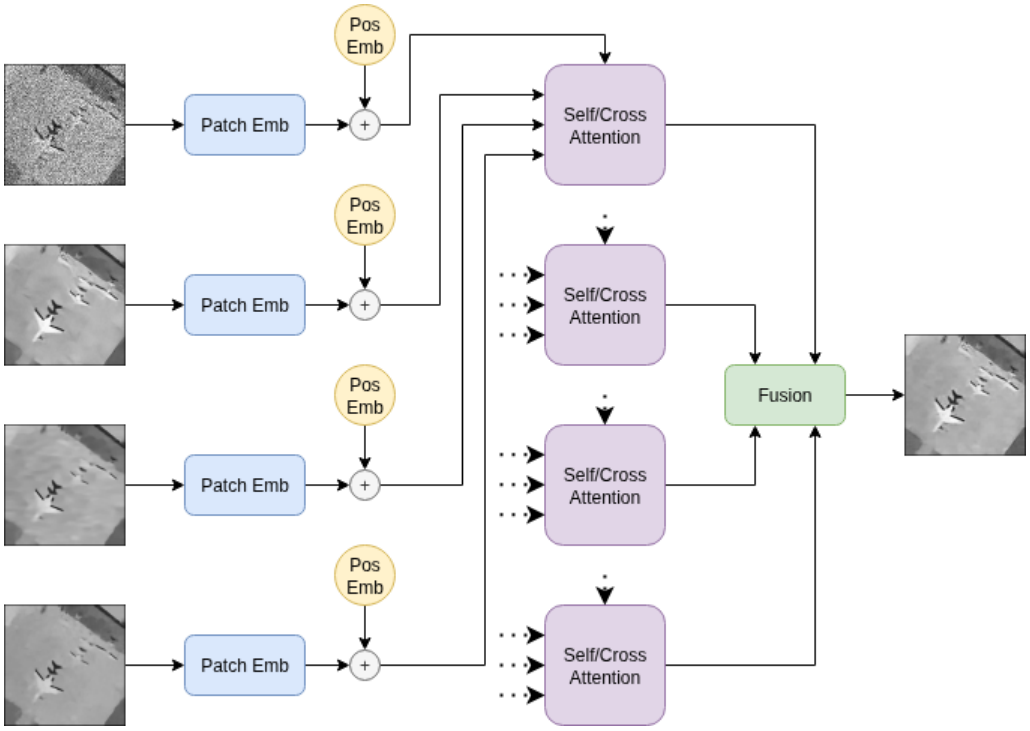

Figure 5: Schema semplificato.

Si parte da più immagini di input che appartengono allo stesso dominio. Ogni immagine viene divisa in patch e passata attraverso un Patch Embedding , poi si aggiunge un Positional Embedding per mantenere l'informazione spaziale. Le rappresentazioni ottenute passano attraverso una serie di blocchi di attenzione, indicati come "Self/Cross Attention" che possono modellare sia le relazioni interne all'immagine (self-attention) sia tra immagini (cross-attention). Infine, le feature elaborate vengono combinate in un modulo di Fusion, che produce l'immagine fusa finale. CrossFuse cerca l'informazione mancante nell'altro dominio, la versione semplificata cerca l'informazione comune e più stabile tra immagini simili. Per questo motivo, eliminare meccanismi come il re-softmax o la doppia fase di addestramento, nati per gestire la non-correlazione può non solo semplificare l'architettura, ma anche migliorare le prestazioni nei contesti monomodali.

5

CONFRONTO DEI METODI DI DESPECKLING

Per ogni approccio utilizzato è stato calcolato sia il PSNR che SSIM, in modo da capire quanto bene è stato eseguito il despeckling e come è stata mantenuta la struttura dell'immagine in relazione anche ai modelli di base che sono stati fusi.

5.1 METRICHE PER LA VALUTAZIONE DELLE IMMAGINI DESPECKLED

5.1.1 *Peak Signal-to-Noise Ratio*

Per la valutazione della qualità delle immagini despeckled è stata utilizzata la metrica PSNR come indicatore. Il PSNR è una metrica usata per misurare la qualità di un'immagine ricostruita o compressa rispetto a un'immagine di riferimento (ground truth). Si basa sull'errore quadratico medio (MSE, Mean Squared Error) tra i pixel dell'immagine originale e quelli dell'immagine degradate/ricostruita. La formuala è:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{MAX_I^2}{\text{MSE}} \right) \tag{9}$$

Dove MAX_I è il valore massimo possibile per un pixel. Più alto è il PSNR, migliore è la qualità dell'immagine ricostruita.

5.1.2 *Structural Similarity Index*

Un'altra metrica utilizzata per la valutazione della qualità delle immagini despeckled è lo SSIM. SSIM è progettato per valutare la qualità percepita

di un'immagine rispetto a un'immagine di riferimento, tenendo conto delle caratteristiche strutturali dell'immagine. La formula è:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{10}$$

Dove:

- μ_x e μ_y sono le medie delle immagini x e y.
- σ_x^2 e σ_y^2 sono le varianze delle immagini x e y.
- σ_{xy} è la covarianza tra le immagini x e y.
- C_1 e C_2 sono costanti per stabilizzare la divisione quando i denominatori sono piccoli.

5.1.3 *Tabella di confronto tramite PSNR della media e media pesata*

Bioma	SAR-CAM	FANS	SARBM3D	MEDIA	MEDIA PESATA
Agricultural	24.95	24.58	25.37	25.28	25.29
Airplane	26.07	23.91	23.20	24.85	24.77
Chaparral	22.93	21.49	22.43	22.60	22.59
Harbor	23.17	21.09	20.56	22.13	22.08

Table 1: Confronto dei modelli di despeckling e della loro fusione. I valori sopra, indicano la media del PSNR di 100 immagini rappresentanti diversi biomi. Ogni modello per la previsione della qualità, usato in MEDIA PESATA, è stato allenato per 10 epoche con un dataset da 30'000 immagini.

Dall'analisi dei valori di PSNR riportati in Tabella 1, è possibile osservare che i metodi di fusione MEDIA e MEDIA PESATA si comportano in modo coerente rispetto alle prestazioni dei singoli modelli di despeckling (SAR-CAM, FANS e SARBM3D). In generale, i valori di MEDIA e MEDIA PESATA risultano compresi tra quelli dei tre modelli originari, come atteso da una procedura di fusione. Questo comportamento indica che la combinazione delle uscite tende ad attenuare le debolezze dei singoli modelli, garantendo una maggiore stabilità e robustezza complessiva del risultato. Nella maggior parte dei biomi considerati, la

MEDIA raggiunge valori di PSNR molto vicini al modello con le migliori prestazioni, spesso SARCAM e supera nettamente il modello peggiore (solitamente SARBM3D). Ciò dimostra che la fusione aritmetica costituisce una strategia efficace per ottenere un risultato medio-bilanciato, in grado di avvicinarsi alla qualità del modello più performante senza sacrificare la coerenza tra le varie scene. Il confronto tra MEDIA e MEDIA PESATA mostra differenze minime, generalmente inferiori a 0.05 dB su tutto il dataset. Tale scostamento marginale suggerisce che i pesi adottati nella media pesata non hanno inciso in modo significativo sul risultato finale, probabilmente a causa di una distribuzione dei pesi simile a quella uniforme o di prestazioni già bilanciate tra i tre modelli di base. Nel complesso, la fusione dei risultati anche nella sua forma più semplice permette di ottenere un compromesso ottimale tra i diversi approcci di despeckling, mantenendo un livello di qualità molto vicino al migliore modello individuale e al contempo più stabile rispetto alle variazioni del contenuto dell'immagine. Tuttavia non riesce nella maggior parte dei casi a dare risultati migliori rispetto al miglior modello.

5.1.4 *Tabella di confronto tramite SSIM della media e della media pesata*

Bioma	SAR-CAM	FANS	SARBM3D	MEDIA	MEDIA PESATA
Agricultural	0.55	0.47	0.59	0.55	0.55
Airplane	0.71	0.67	0.68	0.70	0.70
Chaparral	0.62	0.46	0.56	0.57	0.57
Harbor	0.81	0.74	0.74	0.78	0.78

Table 2: Confronto dei modelli di despeckling e della loro fusione. I valori sopra, indicano la media del SSIM di 100 immagini rappresentanti diversi biomi. Ogni modello per la previsione della qualità, usato in MEDIA PESATA, è stato allenato per 10 epoche con un dataset da 30'000 immagini.

Anche dai dati riportati nella tabella 4 si osserva come sia la MEDIA sia la MEDIA PESATA tendano a riflettere le prestazioni del modello migliore in termini di SSIM. La qualità strutturale dell'immagine non risulta mai inferiore o pari a quella del modello meno performante, confermando che, sebbene questo approccio non rappresenti la soluzione ottimale, costituisce comunque un buon compromesso tra i diversi modelli.

Tabella di confronto tramite PSNR del modello crossfuse

Bioma	SAR-CAM	FANS	SARBM3D	CROSSFUSE	CROSSFUSE LIGHT
Agricultural	24.95	24.58	25.37		23.73
Airplane	26.07	23.91	23.20		26.48
Chaparral	22.93	21.49	22.43		21.90
Harbor	23.17	21.09	20.56		22.28

Table 3: Confronto dei modelli di despeckling e della loro fusione tramite Cross-Fuse tramite PSNR.

Tabella di confronto tramite SSIM del modello crossfuse

Bioma	SAR-CAM	FANS	SARBM3D	CROSSFUSE	CROSSFUSE LIGHT
Agricultural	0.55	0.47	0.59	0.38	0.54
Airplane	0.71	0.67	0.68	0.45	0.70
Chaparral	0.62	0.46	0.56	0.48	0.62
Harbor	0.81	0.74	0.74	0.64	0.78

Table 4: Confronto dei modelli di despeckling e della loro fusione tramite Cross-Fuse tramite SSIM.

QUI METTERE LE CONSIDERAZIONI UNA VOLTA VISTI I RISULTATI

5.1.5 *Confronto visivo dei vari modelli*

Media pesata

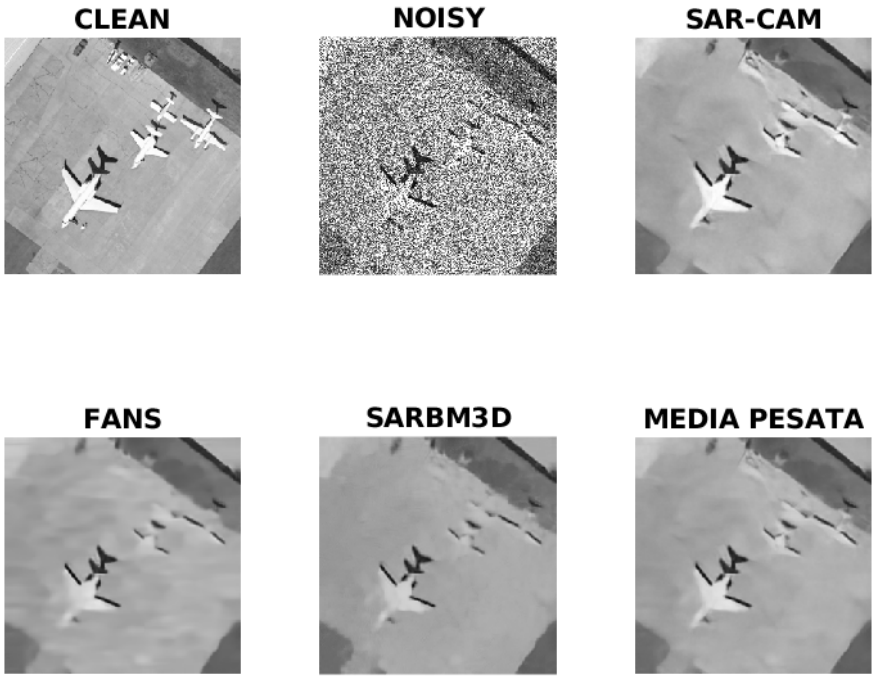

Figure 6

Confrontando i singoli modelli di despeckling con il ground truth si puo notare come sia SARBM3D che FANS abbiano perso dettagli sematici importanti durante il denoising come ad esempio l'aero più piccolo e il corpo principale degli aerei accanto sono spariti dall'immagine despeckled. SARCAM invece preserva meglio le informazioni semantiche eseguendo un despeckling che si avvicina di più all'immagine clean. Nell'immagine con i tre modelli fusi tramite la media pesata notiamo che sono stati riportati gli aerei che si erano persi in FANS e SARBM3D. Cio torna con l'analisi fatta con le metriche come PSNR e SSIM in cui l'immagine fusa tende sempre ad allinearsi al modello migliore senza fare mai peggio del modello meno efficace.

media aritmetica

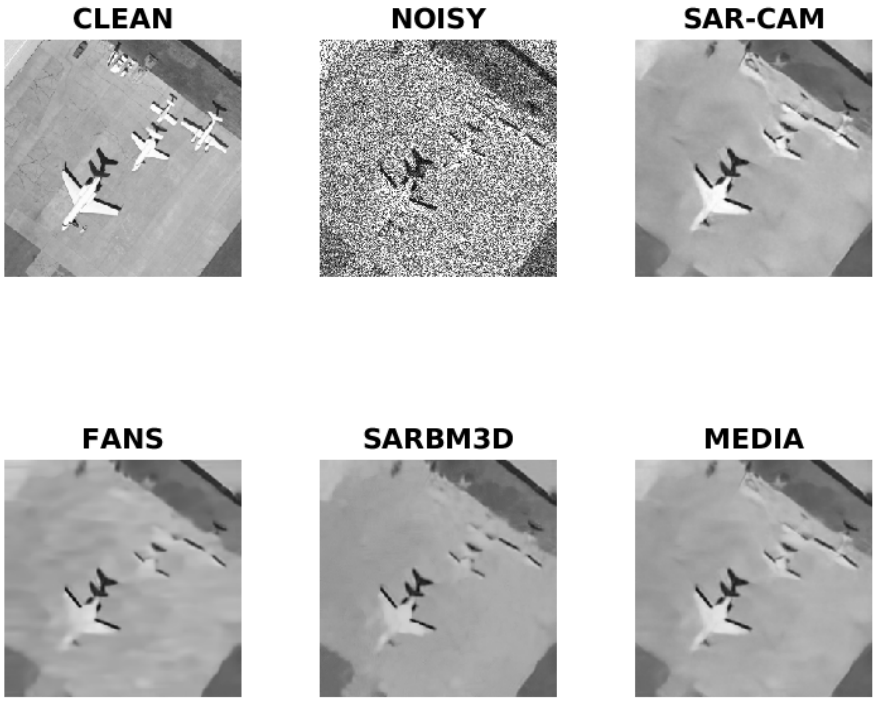

Figure 7

Queste immagini ci danno la prova visiva di quanto già osservato attraverso le metriche PSNR e SSIM, ovvero che la media aritmetica si comporta in modo molto simile alla media pesata. In entrambi i casi, la fusione tende a privilegiare le regioni in cui i singoli modelli hanno performato meglio, combinando in maniera efficace le loro rispettive capacità di riduzione del rumore e preservazione del dettaglio.

6

CONCLUSIONS AND FUTURE WORK

In questa tesi sono stati esplorati differenti approcci per la fusione di immagini SAR despeckled. Tecniche naive come la media e la media pesata hanno mostrato comunque di avere un comportamento stabile e robusto, d'altro canto hanno il limite di avvicinarsi sempre al modello migliore ma senza mai superarlo. Non consiglierei sviluppi futuri riguardandi questo approccio in quanto anche nel migliore dei casi non si arriva alle prestazione degli altri metodi di fusione più avanzati.

l'approccio basato sull'articolo CrossFuse è sicuramente più interessante in quanto presenta meccanismi più avanzati con la quale è più efficace capire dove il despeckling è avvenuto con maggiore successo. Consiglio di esplorare ulteriormente questo approccio in quanto sfruttare meccanismi di attention per la fusione di immagini despeckled è sicuramente un campo interessante e con ampi margini di miglioramento. Un possibile sviluppo futuro potrebbe essere quello di utlizzare altri modelli di despeckling come input, che abbiano tra loro caratteristiche più diverse in modo da avere una fusione più efficace.

BIBLIOGRAPHY

[1] Fabrizio Argenti, Alessandro Lapini, Tiziano Bianchi, and Luciano Alparone. A tutorial on speckle reduction in synthetic aperture radar images. *IEEE Geoscience and Remote Sensing Magazine*, 1(3):6–35, 2013.

[2] Fabrizio Argenti and Gionatan Torricelli. Speckle suppression in ultrasonic images based on undecimated wavelets. *EURASIP Journal on Advances in Signal Processing*, (5):379638, 2003.

[3] NASA Earthdata. Synthetic aperture radar (sar), 2025.

[4] G. Esposito, I. Marchesini, A. C. Mondini, P. Reichenbach, M. Rossi, and S. Sterlacchini. A spaceborne sar-based procedure to support the detection of landslides. *Natural Hazards and Earth System Sciences*, 20(9):2379–2395, 2020.

[5] European Space Agency (ESA). Sentinel-1: Radar vision for copernicus, 2025.

[6] Lloyd Haydn Hughes, Diego Marcos, Sylvain Lobry, Devis Tuia, and Michael Schmitt. A deep learning framework for matching of sar and optical imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:166–179, 2020.

[7] Lattari, Leon, Asaro, Rucci, Prati, and Matteucci. Deep learning for sar image despeckling. *remote sensing*, 1(1):20, 2019.

[8] Hui Li and Xiao-Jun Wu. CrossFuse: A Novel Cross Attention Mechanism based Infrared and Visible Image Fusion Approach. *Information Fusion*, 103:102147, 2024.

[9] Sanjjushri Varshini R, Rohith Mahadevan, Bagiya Lakshmi S, Mathivanan Periasamy, Raja CSP Raman, and Lokesh M. Speckle noise analysis for synthetic aperture radar (sar) space data, 2024.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[12] G. V. Vyver, S. E. Måsøy, H. Dalen, B. L. Grenne, E. Holte, S. H. Olaisen, J. Nyberg, A. Østvik, L. Løvstakken, and E. Smistad. Regional image quality scoring for 2-d echocardiography using deep learning. *Ultrasound in Medicine & Biology*, 51(4):638–649, 2025.

[13] Yongjian Yu and S.T. Acton. Speckle reducing anisotropic diffusion. *IEEE Transactions on Image Processing*, 11(11):1260–1270, 2002.