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

Tesi di Laurea

METODI DI APPRENDIMENTO AUTOMATICO PER LA PREDIZIONE DELLA QUALITÀ DI IMMAGINI SAR DESPECKLED

MACHINE LEARNING METHODS FOR THE PREDICTION OF DESPECKLED IMAGE QUALITY

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. I satelliti SAR, grazie a questa loro capacita, trovano applicazione in molteplici contesti. In ambito geologico, 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. Il funzionamento di questo tipo di satellite 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. L'immagine così generata 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. 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 statistica multivarianza o 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 utilizza un meccanismo di 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. Il cross attention, infatti, nasce nel contesto della fusione multimodale di immagini (ad esempio infrarosso e visibile) e mira a combinare i punti di forza di diverse fonti mantenendo l'informazione complementare ed eliminando quella ridondante. La chiave è che, mentre la self-attention tende a enfatizzare le correlazioni interne ad una singola immagine (quindi rischia di rafforzare il rumore residuo), la cross-attention sfrutta l'eterogeneità tra le diverse versioni despeckled per mettere in evidenza le informazioni non correlate, cioè i dettagli che risultano più affidabili in una ricostruzione ma assenti o degradati nelle altre. In questo modo il modello riesce a potenziare le componenti salienti e a ridurre le parti rumorose. Questo

approccio consente di ottenere un'immagine despeckled più accurata e leggibile, perché tiene conto non solo del valore di ogni singolo pixel, ma anche del contesto e della complementarità tra più metodi di denoising, superando i limiti della semplice media pesata.

2

STATE OF THE ART

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. 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. 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. Alcuni di queste tipologie di filtri sono stati illustrati e confrontate nell'articolo *A Tutorial on Speckle Reduction in Synthetic Aperture Radar Images* [1]. Negli ultimi anni, l'attenzione si è spostata ancora più avanti verso i metodi non locali, come i filtri non local means o BM3D adattati per le 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 general-

mente 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. Come riportato nell'articolo *Deep Learning for SAR Images Despeckling* [2], 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 ditribuzione statistica del rumore o una struttura matematica da preservare, ma si affida alla capacità della rete di apprendere automaticamante 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. Questo tipo di approccio non va a reinventare la ruota, cioè non punta a realizzare un nuovo modello con cui è possibile fare denoising, 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. Questa tecnica però non sfrutta a pieno le qualità dei singoli modelli in quanto la stima della qualità è locale e si concentra sul singolo pixel, perdendo informazione contestuale importante. Inoltre la combinazione pesata a livello di pixel non sfrutta la complementarità tra caratteristiche a livello di patch, perciò non valorizza

pienamente i punti di forza di ciascun metodo Un approccio migliore è basato sull'articolo *CrossFuse: A Novel Cross Attention Mecha- nism based Infrared and Visible Image Fusion Approach.* [3] che sfrutta tecniche basate sull'attenzione incrociata. 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 del 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 1, il dataset impiegato è costituito da tre tipologie di immagini: clean, noisy e denoised. 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. 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, invece, assieme a quelle denoised, vengono impiegate per la generazione della mappa di qualità, che costituisce il riferimento necessario per il calcolo della funzione di perdita. 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.

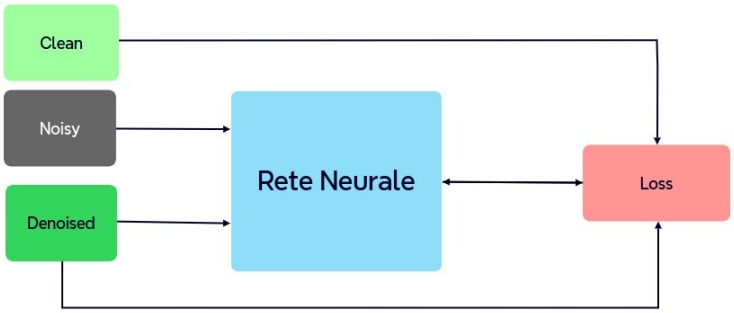

Figure 1: Mock della struttura logica per la previsione della qualità

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. Il compito di questa rete è assegnare a ciascun pixel un valore continuo 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 è composta da due parti principali: encoder e decoder. L'architettura è organizzata in due parti principali: Encoder (contrazione): 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 (espansione): 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. 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.

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

Non è una fusione "cieca" (es. una media semplice). 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 quell'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 l'output dell'algoritmo migliore per una data caratteristica dell'immagine, scartando virtualmente 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, ovvero l'immagine clean [Î].

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

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.

3.5.1 *Tabella di confronto tramite PSNR dei vari modelli*

Bioma	SAR-CAM	FANS	SARBM3D	TESI
Agricultural	24.95	24.58	25.37	19.07
Airplane	26.07	23.91	23.20	23.64
Baseball diamond	28.50	27.20	26.87	21.05
Beach	28.89	25.84	24.50	16.11
Buldings	24.51	22.23	21.52	21.89
Chaparral	22.93	21.49	22.43	18.78
Forest	26.48	25.66	25.97	18.44
Freeway	25.88	24.03	23.89	21.22
Golf course	28.41	27.23	26.97	20.86

Table 1: Confronto dei modelli di despeckling e della loro fusione

I valori sopra, indicano la media del PSNR di 100 immagini rappresenti diversi biomi. Ogni modello per la previsione della qualità, usato in TESI, è stato allenato per 3 epoche con un dataset da 30'000 immagini. Come si puo notare dalla tabella [1] questa tecnica di fusione non valorizza al meglio i singoli modelli portando il risultato finale ad essere peggiore di quello dei singoli modelli presi singolarmente

4

APPROCCIO BASATO SULLA SELF E CROSS ATTENTION

BIBLIOGRAPHY

[1] Argenti, Lapin, Bianchi, and Alparone. A tutorial on speckle reduction in synthetic aperture radar images. *IEEE Xplore*, 1(2):6–35, 2013.

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

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