DIPARTIMENTO DI INGEGNERIA INFORMATICA, MODELLISTICA,
ELETTRONICA E SISTEMISTICA

Corso di Laurea Magistrale in Ingegneria Informatica

Machine & Deep Learning

Elaborato Finale

Re-Engineering AE-XAD with Vision Transformers for Explainable Anomaly Detection

Docenti:

Prof. **Fabrizio Angiulli**
Ing. **Francesco De Luca**

Studente:
Presta Vincenzo
matr. 252290

ANNO ACCADEMICO 2024/2025

Indice generale

1 Introduzione

L'Anomaly Detection in ambito industriale rappresenta un problema di grande rilevanza applicativa, poiché consente l'individuazione automatica di difetti e anomalie su superfici e componenti prodotti in serie. In molti scenari reali, le anomalie risultano rare, eterogenee e difficilmente annotabili in modo esaustivo, rendendo complessa l'applicazione di approcci supervisionati tradizionali. Per questo motivo, si sono affermati metodi basati su Autoencoder, capaci di apprendere una rappresentazione delle sole istanze normali e di identificare le anomalie come deviazioni rispetto al comportamento appreso.

All'interno di questo paradigma, i metodi di anomaly detection basati su ricostruzione hanno dimostrato particolare efficacia in contesti industriali, soprattutto quando è richiesta una localizzazione spaziale dei difetti. In tale direzione si colloca il framework AE-XAD, che introduce una pipeline strutturata per la rilevazione e la localizzazione delle anomalie attraverso l'analisi dell'errore di ricostruzione [1]. Il metodo combina un encoder convoluzionale, un decoder progettato per enfatizzare le discrepanze rispetto alla normalità e un meccanismo di decisione basato su statistiche globali pixel-wise, ottenendo risultati competitivi sia a livello di immagine che a livello di localizzazione.

Negli ultimi anni, l'evoluzione delle architetture di visione ha portato all'emergere di modelli basati su meccanismi di attenzione globale, come i Vision Transformer, che hanno mostrato elevate capacità di rappresentazione in numerosi compiti di visione artificiale. Questo progresso solleva naturalmente l'interrogativo se tali architetture possano sostituire efficacemente le reti convoluzionali anche all'interno di pipeline di anomaly detection basate su ricostruzione.

Tuttavia, l'integrazione di un Vision Transformer all'interno del framework AE-XAD non è immediata. AE-XAD non è un autoencoder generico, ma un metodo che fa affidamento su specifiche assunzioni strutturali riguardanti la distribuzione spaziale dell'errore di ricostruzione e sulla sua separabilità statistica dal rumore di fondo. In questo contesto, la sostituzione dell'encoder convoluzionale con un'architettura caratterizzata da un diverso tipo di rappresentazione solleva interrogativi fondamentali sulla compatibilità tra il modello di features apprese e il meccanismo decisionale adottato.

L'obiettivo di questo elaborato è quindi analizzare in modo sistematico se, e in quali condizioni, un Vision Transformer possa sostituire l'encoder convoluzionale originale di AE-XAD mantenendo inalterati il decoder, la funzione di loss, le metriche di valutazione e l'intera pipeline di test, al fine di garantire un confronto equo e scientificamente rigoroso.

1.1 Definizione del Problema e Ipotesi di Ricerca

Il framework AE-XAD assume implicitamente che le anomalie producano errori di ricostruzione *spazialmente localizzati e compatti*, tali da poter essere distinti dal rumore di fondo mediante una soglia statistica globale basata su media e deviazione standard ($\mu + \sigma$). Questa assunzione risulta naturalmente coerente

con le proprietà delle architetture convoluzionali, che favoriscono una rappresentazione gerarchica e localmente strutturata delle informazioni spaziali.

Le architetture basate su attenzione globale, come i Vision Transformer, presentano invece un comportamento rappresentazionale differente, orientato alla modellazione di relazioni globali tra regioni dell'immagine. Sebbene tale caratteristica possa risultare vantaggiosa in compiti di natura semantica, non è immediatamente evidente se essa sia compatibile con un paradigma di anomaly detection basato su errori di ricostruzione pixel-wise e su una sogliatura statistica globale.

La domanda di ricerca che guida questo lavoro può pertanto essere formulata come segue:

> *Un Vision Transformer può sostituire efficacemente l'encoder convoluzionale di AE-XAD, mantenendo inalterata la pipeline decisionale, in un contesto di anomaly detection few-shot supervisionato?*

L'ipotesi investigata in questa tesi è che, pur essendo in grado di apprendere rappresentazioni utili per il ranking delle anomalie, i Vision Transformer tendano a produrre errori di ricostruzione più diffusi e meno localizzati rispetto alle architetture convoluzionali. Di conseguenza, il meccanismo di binarizzazione basato su $\mu + \sigma$ risulterebbe intrinsecamente meno efficace, portando a un degrado delle prestazioni di localizzazione in specifiche classi del dataset considerato.

2 Fondamenti teorici

In questa sezione vengono introdotti i concetti teorici necessari a inquadrare il contesto metodologico di questo lavoro e a chiarire le assunzioni implicite alla base del framework utilizzato. In particolare, viene discusso il paradigma di anomaly detection basato su ricostruzione, viene descritto il framework AE-XAD dal punto di vista concettuale e viene analizzato il ruolo dell'inductive bias nelle architetture di visione, in relazione al meccanismo di decisione adottato.

2.1 Anomaly Detection basata su ricostruzione

Gli approcci di anomaly detection basati su ricostruzione si fondano sull'idea di apprendere un modello delle sole istanze normali, in modo tale che le anomalie possano essere identificate come deviazioni rispetto al comportamento appreso. In questo paradigma, un autoencoder viene addestrato a ricostruire immagini prive di difetti, minimizzando un errore di ricostruzione calcolato a livello pixel-wise.

In fase di test, la presenza di anomalie si traduce tipicamente in un aumento dell'errore di ricostruzione nelle regioni difettose, rendendo possibile l'individuazione delle anomalie sia a livello di immagine sia a livello pixel-wise. L'efficacia di tali approcci dipende pertanto non solo dalla capacità del modello

di rappresentare correttamente la normalità, ma anche dalla struttura spaziale dell'errore di ricostruzione prodotto.

2.2 Il framework AE-XAD

Il framework AE-XAD (AutoEncoder for eXplainable Anomaly Detection) rappresenta un'evoluzione degli approcci classici basati su autoencoder, introducendo una pipeline progettata specificamente per la localizzazione delle anomalie in contesti industriali [1]. Il metodo è composto da un encoder convoluzionale, un decoder con una struttura a rami e un meccanismo di decisione basato su statistiche globali dell'errore di ricostruzione.

Un elemento centrale di AE-XAD è la costruzione di una *reconstruction error map*, ottenuta confrontando l'immagine di input con la ricostruzione prodotta dal decoder. Tale mappa rappresenta la distribuzione spaziale dell'errore di ricostruzione ed è utilizzata come base per la localizzazione delle anomalie. Per ridurre il rumore ad alta frequenza, la mappa viene sottoposta a un'operazione di filtraggio, seguita da una binarizzazione mediante una soglia statistica globale definita come $\mu+\sigma$, dove μ e σ indicano rispettivamente la media e la deviazione standard dei valori di errore.

Questo meccanismo di sogliatura implica che le anomalie siano caratterizzate da errori di ricostruzione significativamente superiori al rumore di fondo e concentrati in regioni spazialmente limitate. Di conseguenza, l'efficacia della pipeline AE-XAD dipende in modo critico dalla distribuzione spaziale dell'errore di ricostruzione prodotto dall'encoder–decoder.

2.3 Inductive bias nelle architetture di visione

Con il termine *inductive bias* si intende l'insieme di assunzioni strutturali che un modello incorpora a priori, influenzando il modo in cui generalizza a partire da un numero limitato di esempi. Nel contesto della visione artificiale, l'inductive bias riveste un ruolo particolarmente rilevante in scenari few-shot, dove la quantità di dati disponibili non è sufficiente a guidare completamente l'apprendimento.

Le architetture convoluzionali, come quelle impiegate in AE-XAD, incorporano un forte inductive bias locale, che favorisce la modellazione di pattern spaziali e la produzione di rappresentazioni gerarchiche sensibili alla localizzazione. Tale caratteristica risulta naturalmente coerente con un paradigma di anomaly detection basato su errori di ricostruzione pixel-wise e su una sogliatura statistica globale.

Architetture caratterizzate da un diverso tipo di rappresentazione, orientate alla modellazione di relazioni globali tra regioni dell'immagine, possono invece produrre distribuzioni dell'errore di ricostruzione più diffuse e meno concentrate. In un framework come AE-XAD, in cui la decisione finale dipende dall'applicazione di una soglia globale a una mappa di errore spaziale, tale differenza rappresentazionale può tradursi in un disallineamento tra le assunzioni della pipeline di decisione e le proprietà dell'encoder utilizzato.

3 Il framework AE-XAD

Il framework AE-XAD (AutoEncoder for eXplainable Anomaly Detection) è un metodo di anomaly detection basato su ricostruzione, progettato specificamente per scenari industriali caratterizzati da anomalie rare, eterogenee e difficilmente annotabili in modo esaustivo [1]. A differenza di autoencoder generici, AE-XAD introduce una pipeline strutturata che integra scelte architetturali, una funzione di perdita dedicata e un meccanismo di scoring esplicitamente orientato alla localizzazione delle anomalie.

L'intero framework è concepito per operare in regime few-shot supervisionato, sfruttando un numero limitato di esempi anomali durante l'addestramento senza snaturare la natura reconstruction-based del metodo.

3.1 Architettura del modello

L'architettura di AE-XAD è composta da tre componenti principali: un encoder convoluzionale, un decoder asimmetrico e un modulo di decisione basato sull'analisi statistica dell'errore di ricostruzione.

L'encoder è costituito dai primi blocchi di una Deep CNN pre-addestrata (ResNet), utilizzata come estrattore di feature. L'encoder produce una rappresentazione latente sotto forma di feature map con risoluzione spaziale fissa pari a 28×28 e 64 canali, scelta per garantire la compatibilità con il decoder e per preservare una struttura spaziale sufficientemente dettagliata.

Il decoder AE-XAD presenta una struttura asimmetrica a due rami. Il primo ramo, non addestrabile, esegue un'operazione di upsampling diretto della feature map latente e applica una funzione di attivazione *tanh*, producendo una ricostruzione regolarizzata che cattura la struttura globale dell'immagine. Il secondo ramo, completamente addestrabile, è composto da una sequenza di blocchi convoluzionali e di deconvoluzione con attivazione SELU, ed è progettato per modellare dettagli più fini della ricostruzione.

Le due ricostruzioni vengono fuse mediante una modulazione moltiplicativa, secondo una formulazione del tipo $b_2 + b_1 \cdot b_2$, dove b_1 e b_2 rappresentano rispettivamente l'output del ramo non addestrabile e di quello addestrabile. Questa scelta architetturale consente di enfatizzare le discrepanze locali rispetto alla normalità, rendendo più evidente l'errore di ricostruzione in corrispondenza delle regioni anomale.

Lo strato finale del decoder è costituito da una convoluzione seguita da una funzione di attivazione sigmoide, che produce l'immagine ricostruita nello spazio RGB.

3.2 Funzione di perdita AE-XAD

Un elemento distintivo del framework AE-XAD è la funzione di perdita, progettata per guidare l'apprendimento in presenza di un numero limitato di esempi anomali. La loss è definita a livello pixel-wise e combina il contributo dei pixel normali e dei pixel anomali in modo differenziato.

Indicando con x l'immagine di input, con $\tilde{x}$ la ricostruzione prodotta dal decoder e con $y_j \in \{0, 1\}$ l'etichetta del pixel j (normale o anomalo), la funzione di perdita è definita come:

$$\ell(x, y) = \sum_{j=1}^{D} \left[(1 - y_j) \frac{(x_j - \tilde{x}_j)^2}{(F(x_j) - x_j)^2} + \lambda_y y_j \frac{(F(x_j) - \tilde{x}_j)^2}{(F(x_j) - x_j)^2} \right],$$

dove $F(x_j)$ è una funzione di scaling fissata a 2 e λ_y è un termine di normalizzazione proporzionale al numero di pixel anomali presenti nell'immagine.

Questa formulazione consente di mantenere il focus sull'apprendimento della normalità, limitando al contempo l'influenza dei pochi esempi anomali disponibili durante l'addestramento. In tal modo, AE-XAD preserva il paradigma reconstruction-based, evitando che il modello degeneri in un classificatore supervisionato.

3.3 Pipeline di scoring e localizzazione

In fase di inferenza, l'immagine di input viene ricostruita dal decoder e confrontata con l'originale per ottenere una *reconstruction error map* $M \in \mathbb{R}^{H \times W}$, definita come la distanza pixel-wise tra input e ricostruzione.

Per ridurre il rumore ad alta frequenza e migliorare la coerenza spaziale dell'errore, la mappa M viene sottoposta a un filtraggio gaussiano adattivo. Successivamente, viene applicata una soglia statistica globale definita come:

$$T = \mu + \sigma,$$

dove μ e σ rappresentano rispettivamente la media e la deviazione standard dei valori della mappa di errore filtrata.

La binarizzazione della mappa consente di ottenere una segmentazione delle regioni anomale, mentre uno score globale a livello di immagine viene calcolato aggregando l'errore normalizzato. Questo meccanismo permette di valutare sia la presenza di anomalie nell'immagine sia la loro localizzazione spaziale.

3.4 Assunzioni implicite del framework AE-XAD

Il funzionamento del framework AE-XAD si basa su una serie di assunzioni implicite, che ne determinano l'efficacia nei contesti industriali considerati.

In primo luogo, si assume che le anomalie producano errori di ricostruzione spazialmente localizzati e sufficientemente concentrati, in modo da risultare statisticamente separabili dal rumore di fondo mediante una soglia globale. Questa assunzione è coerente con la presenza di difetti fisici localizzati sulle superfici industriali.

In secondo luogo, il decoder e la pipeline di scoring sono progettati per operare su feature map dotate di una forte struttura spaziale locale, come quelle prodotte da encoder convoluzionali. L'inductive bias locale delle CNN favorisce infatti la preservazione di dettagli e discontinuità spaziali, fondamentali per la localizzazione pixel-wise delle anomalie.

Infine, l'intero framework presuppone una compatibilità strutturale tra encoder, decoder e meccanismo decisionale. Qualsiasi modifica a uno di questi componenti può alterare la distribuzione dell'errore di ricostruzione e compromettere l'efficacia della soglia statistica adottata.

4 Modifica architetturale: integrazione del Vision Transformer

In questa sezione viene descritta la modifica architetturale introdotta in questo lavoro, che consiste nella sostituzione dell'encoder convoluzionale originale di AE-XAD con un Vision Transformer. L'obiettivo è analizzare in modo controllato l'impatto di un diverso inductive bias sulla distribuzione spaziale dell'errore di ricostruzione e, di conseguenza, sulle prestazioni di anomaly detection e localizzazione.

Per garantire un confronto equo e scientificamente rigoroso, tutte le altre componenti del framework AE-XAD vengono mantenute invariate rispetto alla formulazione originale descritta in [1].

4.1 Obiettivo della modifica

La scelta di integrare un Vision Transformer all'interno del framework AE-XAD nasce dall'interesse verso architetture di visione basate su meccanismi di self-attention, che hanno dimostrato elevate capacità rappresentazionali in numerosi compiti di visione artificiale.

Tuttavia, AE-XAD non è un autoencoder generico, bensì un metodo progettato attorno a precise assunzioni sulla struttura dell'errore di ricostruzione e sulla sua separabilità statistica dal rumore di fondo. L'obiettivo di questa modifica non è quindi quello di migliorare direttamente le prestazioni del modello, ma di valutare se un encoder caratterizzato da un inductive bias differente sia compatibile con la pipeline decisionale di AE-XAD, mantenendo invariati decoder, loss e meccanismo di scoring.

4.2 Encoder ViT: struttura e adattamento spaziale

L'encoder convoluzionale originale è stato sostituito con un Vision Transformer di tipo ViT-B/16, pre-addestrato su ImageNet. L'architettura ViT opera suddividendo l'immagine di input in patch non sovrapposte di dimensione 16×16, che vengono proiettate in uno spazio di embedding tramite una convoluzione (*patch embedding*) e successivamente elaborate da una sequenza di blocchi Transformer basati su self-attention.

Poiché il decoder AE-XAD richiede in input una feature map spaziale di dimensione $28 \times 28 \times 64$, è stato necessario introdurre un adattamento architetturale per riconvertire l'output del Vision Transformer in una rappresentazione compatibile. In particolare, i token prodotti dal ViT (ad eccezione del token

di classe) vengono rimappati in una griglia spaziale bidimensionale, successivamente proiettata tramite moduli convoluzionali per ottenere una feature map con la risoluzione e il numero di canali richiesti dal decoder originale.

Questa operazione di riconversione rappresenta una differenza strutturale fondamentale rispetto all'encoder convoluzionale, poiché introduce un passaggio intermedio che non è presente nella formulazione originale di AE-XAD.

Nel processo di estrazione delle feature, il token CLS prodotto dal Vision Transformer non viene utilizzato. Tale scelta è coerente con il framework AE-XAD, che richiede una rappresentazione spaziale densa per la ricostruzione e la localizzazione delle anomalie. Il token CLS, pur veicolando informazione globale, non è associato a una posizione spaziale specifica e risulta pertanto incompatibile con la pipeline di ricostruzione e scoring basata su mappe di errore pixel-wise

4.3 Componenti mantenuti invariati

Al fine di isolare l'effetto della sostituzione dell'encoder, tutte le altre componenti del framework AE-XAD sono state mantenute invariate. In particolare:

- il decoder asimmetrico a due rami è identico a quello descritto nel framework originale;
- la funzione di perdita AE-XAD è utilizzata senza alcuna modifica;
- la pipeline di scoring e localizzazione basata sulla soglia globale $\mu + \sigma$ è mantenuta invariata;
- le metriche di valutazione e pipeline di test adottate sono le stesse previste dal framework originale.

Questa scelta consente di attribuire le variazioni osservate principalmente alle proprietà rappresentazionali dell'encoder utilizzato, a parità di decoder, funzione di perdita e pipeline di scoring.

È tuttavia importante osservare che il decoder di AE-XAD non è architetturalmente neutro rispetto all'encoder. La sua struttura, basata su operazioni convoluzionali e su un ramo di ricostruzione non addestrabile, presuppone una rappresentazione latente caratterizzata da località spaziale, stazionarietà e gerarchia multiscala, tipiche delle architetture CNN. Di conseguenza, la sostituzione dell'encoder convoluzionale con un Vision Transformer non costituisce una modifica simmetrica del modello, ma introduce una potenziale discontinuità tra la natura delle feature estratte e le assunzioni implicite del decoder.

4.4 Regimi di addestramento dell'encoder ViT

Per analizzare in modo più approfondito il ruolo dell'encoder Vision Transformer all'interno del framework AE-XAD, sono stati considerati due distinti regimi di addestramento, che differiscono esclusivamente per la modalità di aggiornamento dei parametri dell'encoder.

4.4.1 Encoder ViT completamente frozen

Nel primo setting sperimentale, l'intero encoder Vision Transformer è mantenuto congelato durante l'addestramento. In questo caso, vengono aggiornati esclusivamente i parametri del decoder AE-XAD.

Questa configurazione riproduce fedelmente l'assunzione adottata nel framework originale, in cui l'encoder convoluzionale pre-addestrato viene utilizzato come estrattore di feature fisso. L'obiettivo di questo setting è valutare se le rappresentazioni apprese dal ViT in fase di pre-addestramento siano direttamente compatibili con il decoder e con il meccanismo di scoring di AE-XAD, senza alcun adattamento al dominio industriale.

4.4.2 Encoder ViT completamente trainable

Nel secondo setting sperimentale, l'encoder Vision Transformer viene reso completamente addestrabile e ottimizzato congiuntamente al decoder AE-XAD. Questo regime consente al ViT di adattare le proprie rappresentazioni al dominio specifico del dataset MVTec AD, caratterizzato da texture ripetitive e difetti locali sottili. L'obiettivo è valutare se un fine-tuning end-to-end dell'encoder sia in grado di produrre feature più compatibili con la pipeline di ricostruzione e di localizzazione di AE-XAD, attenuando il disallineamento introdotto dal diverso inductive bias.

4.5 Differenze strutturali rispetto all'encoder convoluzionale

La sostituzione dell'encoder convoluzionale con un Vision Transformer introduce differenze strutturali rilevanti all'interno del framework AE-XAD. In particolare, il Vision Transformer tende a modellare relazioni globali tra regioni dell'immagine, riducendo l'enfasi sulla località spaziale che caratterizza le architetture convoluzionali.

Questa differenza di inductive bias può influenzare la distribuzione spaziale dell'errore di ricostruzione, producendo mappe di errore più diffuse e meno concentrate. In un framework come AE-XAD, in cui la decisione finale si basa su una soglia statistica globale applicata a una mappa di errore spaziale, tale disallineamento strutturale può compromettere l'efficacia della localizzazione pixel-wise delle anomalie.

L'analisi sperimentale dei due regimi di addestramento consente quindi di distinguere tra limiti intrinseci dell'architettura ViT e limiti dovuti alla mancanza di adattamento al dominio, rendendo possibile una lettura critica dei risultati sperimentali alla luce delle assunzioni implicite del framework AE-XAD.

5 Setup sperimentale

In questa sezione viene descritto il setup sperimentale adottato per valutare l'impatto della sostituzione dell'encoder convoluzionale di AE-XAD con un Vision Transformer. Tutti gli esperimenti sono stati condotti seguendo un proto-

collo controllato, mantenendo invariata la pipeline originale del framework, al fine di attribuire le variazioni osservate esclusivamente alle proprietà rappresentazionali dell'encoder utilizzato.

5.1 Dataset

Gli esperimenti sono stati condotti sul dataset MVTec Anomaly Detection (MVTec AD), ampiamente utilizzato per la valutazione di metodi di anomaly detection in ambito industriale. Il dataset comprende 15 categorie di oggetti e superfici, ciascuna caratterizzata da immagini normali e da un insieme di immagini anomale accompagnate da maschere pixel-wise di ground truth.

Per ogni categoria sono stati utilizzati:

- tutte le immagini normali disponibili nel set di training;
- tutte le immagini normali e anomale del set di test;
- le maschere di ground truth associate alle anomalie di test.

Il dataset MVTec AD presenta una marcata eterogeneità tra le diverse categorie, sia in termini di struttura visiva degli oggetti sia nella tipologia delle anomalie. Alcune classi sono caratterizzate prevalentemente da difetti estesi e strutturati su superfici quasi uniformi (ad esempio texture), mentre altre presentano anomalie localizzate, sottili o di piccole dimensioni, spesso associate a componenti meccanici complessi.

Questa eterogeneità rende MVTec AD particolarmente adatto allo studio di metodi di anomaly detection basati su ricostruzione e localizzazione pixel-wise, poiché permette di valutare la capacità del modello di gestire anomalie con diversa scala spaziale e diverso grado di separabilità dal rumore di fondo.

Seguendo il protocollo di AE-XAD, il dataset è stato utilizzato in regime few-shot supervisionato, rendendo disponibili durante l'addestramento un numero limitato di esempi anomali per ciascuna classe.

5.2 Protocollo few-shot supervisionato

Per ciascuna categoria del dataset MVTec AD sono stati selezionati n_{anom} campioni anomali da includere nel training set, mentre tutte le restanti anomalie sono state utilizzate esclusivamente in fase di test. Tale impostazione replica il regime few-shot supervisionato previsto dal framework AE-XAD, in cui l'obiettivo è guidare l'apprendimento senza disporre di una copertura esaustiva delle possibili anomalie.

La scelta di adottare un regime few-shot supervisionato è particolarmente coerente con il dataset MVTec AD, in cui le anomalie sono per loro natura rare e fortemente sbilanciate rispetto alle istanze normali, rispecchiando scenari industriali realistici.

Le etichette a livello di immagine sono binarie, con valore 0 per le immagini normali e valore 1 per le immagini anomale. Durante il training, alle immagini anomale è associata anche una maschera pixel-wise che identifica le regioni difettose.

5.3 Preprocessing delle immagini

Tutte le immagini sono state preprocessate seguendo una pipeline uniforme. In particolare:

- le immagini sono state ridimensionate a 224×224 pixel tramite interpolazione nearest-neighbor;
- è stata mantenuta la rappresentazione RGB a tre canali;
- non è stata applicata alcuna normalizzazione o trasformazione fotometrica;
- le immagini sono state fornite al modello nella forma $(3, 224, 224)$.

La scelta di non applicare una normalizzazione ImageNet è coerente con l'impostazione originale di AE-XAD e consente di evitare l'introduzione di ulteriori adattamenti specifici dell'encoder Vision Transformer

Seguendo il protocollo standard del dataset MVTec AD, tutti gli esperimenti sono stati condotti addestrando un modello separato per ciascuna categoria. Per ogni classe, il modello è stato addestrato per 200 epoche, utilizzando esclusivamente i dati appartenenti alla categoria considerata.

Questa impostazione consente di evitare interferenze tra distribuzioni visive eterogenee e di valutare l'impatto della sostituzione dell'encoder in modo indipendente per ciascuna classe.

Le scelte fatte sono coerenti sia con i requisiti dell'encoder ViT sia con l'impostazione originale di AE-XAD, che non prevede una normalizzazione standard delle immagini di input.

5.4 Data augmentation

Nel framework AE-XAD Arrays, la strategia di data augmentation è articolata e prevede la replicazione sistematica delle anomalie, l'applicazione di tecniche di cut–paste con trasformazioni geometriche e un meccanismo di oversampling dei batch per garantire una proporzione controllata tra esempi normali e anomali.

In questa sperimentazione, tale strategia non è stata adottata nella sua forma completa. È stata invece utilizzata una pipeline di data augmentation semplificata, con l'obiettivo di isolare l'effetto architetturale della sostituzione dell'encoder convoluzionale con il Vision Transformer.

In particolare:

- non sono state applicate rotazioni, flip o trasformazioni geometriche esplicite;
- non è stato utilizzato alcun meccanismo di oversampling dei batch;

- per ciascun campione anomalo di training è stata generata una singola variante mediante l'aggiunta di un leggero rumore gaussiano;
- opzionalmente, è stata applicata una procedura di copy–paste del difetto su immagini normali, senza distorsioni geometriche e mantenendo la posizione originale dell'anomalia.

La scelta di adottare una versione semplificata della data augmentation è motivata dalla volontà di evitare l'introduzione di ulteriori inductive bias convoluzionali non legati all'architettura dell'encoder. La pipeline di augmentation proposta in AE-XAD Arrays è infatti progettata per rafforzare le proprietà di località e invarianza tipiche delle architetture CNN. L'utilizzo integrale di tale pipeline in combinazione con un encoder Vision Transformer avrebbe reso meno chiara l'attribuzione causale degli effetti osservati, confondendo l'impatto architetturale dell'encoder con quello delle trasformazioni applicate ai dati.

5.5 Funzione di loss

In tutti gli esperimenti è stata utilizzata la funzione di loss AE-XAD originale, implementata fedelmente secondo la formulazione proposta nel paper Arrays. La loss è definita a livello pixel-wise e combina il contributo dei pixel normali e dei pixel anomali mediante un termine di normalizzazione che tiene conto del numero di pixel anomali presenti nell'immagine.

L'uso della loss originale garantisce che il comportamento del modello in fase di training sia direttamente confrontabile con quello descritto nel framework AE-XAD.

5.6 Ottimizzazione e dettagli di training

L'addestramento del modello è stato effettuato utilizzando l'ottimizzatore Adam, con learning rate iniziale pari a 5×10^{-4} e weight decay pari a 1×10^{-5}. È stato adottato uno scheduler di tipo Cosine Annealing Learning Rate, con $T_{max} = 200$ epoche e learning rate minimo pari a 10^{-6}.

È stato applicato gradient clipping con norma ℓ_2 limitata a 1.0. Il batch size utilizzato è pari a 32, e il numero totale di epoche di addestramento è fissato a 200, in accordo con il protocollo sperimentale adottato nel framework AE-XAD originale.

Nel setting con encoder Vision Transformer congelato, sono stati aggiornati esclusivamente i parametri del decoder. Nel setting completamente trainable, tutti i parametri del modello sono stati ottimizzati congiuntamente.

5.7 Pipeline di training e test

Durante il training, ciascun batch fornisce al modello l'immagine di input, la label a livello di immagine e, per i campioni anomali, la maschera pixel-wise di ground truth. La procedura di test segue fedelmente la pipeline ufficiale AE-XAD e prevede:

1. il calcolo dell'errore di ricostruzione normalizzato;
2. l'applicazione di un filtro gaussiano adattivo;
3. la generazione della heatmap binarizzata mediante soglia globale $\mu + \sigma$;
4. il calcolo dello score a livello di immagine;
5. la valutazione delle metriche di localizzazione e rilevazione.

Le maschere pixel-wise fornite dal dataset sono utilizzate esclusivamente in fase di valutazione, al fine di misurare le prestazioni di localizzazione delle anomalie. Durante l'addestramento, esse sono impiegate unicamente per i pochi campioni anomali inclusi nel regime few-shot supervisionato, in accordo con il framework AE-XAD.

Tutti i risultati quantitativi e qualitativi sono stati ottenuti utilizzando questa pipeline invariata.

6 Risultati sperimentali

In questa sezione vengono presentati i risultati sperimentali ottenuti sul dataset MVTec AD utilizzando il framework AE-XAD con encoder Vision Transformer. L'analisi è inizialmente focalizzata sul setting in cui l'encoder ViT è mantenuto completamente frozen durante l'addestramento, al fine di valutare il comportamento del modello in assenza di adattamento al dominio industriale.

Tutti gli esperimenti sono stati condotti secondo il protocollo descritto nella Sezione 5, mantenendo invariati il decoder, la funzione di perdita e la pipeline di scoring del framework AE-XAD. Le prestazioni riportate riflettono pertanto l'impatto delle rappresentazioni fornite da un encoder Transformer pre-addestrato, utilizzato come estrattore di feature fisso.

6.1 Risultati quantitativi per classe (ViT frozen)

La Tabella 1 riporta i risultati ottenuti per ciascuna delle 15 classi del dataset MVTec AD utilizzando un encoder Vision Transformer completamente frozen. Le prestazioni sono valutate mediante metriche image-level (X-AUC) e pixel-level (F1-score e Intersection over Union), calcolate sulle mappe di errore binarizzate tramite soglia statistica globale $\mu + \sigma$.

6.2 Analisi delle prestazioni image-level

Considerando la metrica X-AUC, che misura la capacità del modello di distinguere immagini normali e anomale a livello globale, si osservano valori generalmente elevati per la maggior parte delle classi del dataset. In numerose categorie lo score X-AUC supera 0.90, indicando che il framework AE-XAD, anche con encoder ViT frozen, conserva una buona capacità di ranking globale delle anomalie.

Table 1: Risultati per classe su MVTec AD con encoder ViT completamente frozen.

Classe	**X-AUC**	**IoU$_{\max}$**	**F1$_{\max}$**
bottle	0.877	0.358	0.490
cable	0.898	0.341	0.467
capsule	0.860	0.102	0.166
carpet	0.897	0.332	0.443
grid	0.894	0.199	0.314
hazelnut	0.973	0.449	0.592
leather	0.973	0.366	0.512
metal_nut	0.922	0.359	0.498
pill	0.945	0.246	0.357
screw	0.914	0.063	0.111
tile	0.955	0.642	0.757
toothbrush	0.945	0.160	0.255
transistor	0.744	0.134	0.208
wood	0.963	0.497	0.643
zipper	0.888	0.418	0.566

Tuttavia, tale comportamento non risulta uniforme su tutte le classi. Alcune categorie presentano una riduzione significativa dello score X-AUC, suggerendo che le rappresentazioni fornite dall'encoder Vision Transformer pre-addestrato non siano sempre sufficientemente discriminative a livello di immagine, soprattutto in presenza di anomalie sottili o strutturalmente complesse.

6.3 Analisi delle prestazioni pixel-level

Un comportamento sensibilmente diverso emerge analizzando le metriche di localizzazione pixel-wise. I valori di F1-score e IoU mostrano una marcata variabilità tra le diverse classi del dataset, evidenziando una forte dipendenza dalla natura spaziale delle anomalie.

Le classi caratterizzate da anomalie estese e visivamente coerenti tendono a ottenere valori più elevati di F1-score e IoU, indicando una localizzazione efficace delle regioni difettose. Al contrario, le classi in cui le anomalie sono sottili, filamentose o di dimensioni ridotte mostrano un netto degrado delle prestazioni di localizzazione.

Questo risultato suggerisce che, nel setting frozen, l'encoder Vision Transformer fatichi a produrre errori di ricostruzione spazialmente concentrati, condizione necessaria per il corretto funzionamento del meccanismo di sogliatura globale adottato dal framework AE-XAD.

6.4 Relazione tra rilevazione e localizzazione

Un aspetto rilevante emerso dai risultati è il parziale disaccoppiamento tra le prestazioni di rilevazione a livello di immagine e quelle di localizzazione pixel-wise. In diverse classi, a valori elevati di X-AUC non corrispondono prestazioni altrettanto elevate in termini di F1-score e IoU.

Questo fenomeno indica che la capacità del modello di identificare la presenza di un'anomalia a livello globale non implica necessariamente una corretta localizzazione spaziale della stessa. Nel contesto del framework AE-XAD, tale disaccoppiamento risulta particolarmente critico, poiché la decisione finale e la valutazione della qualità della segmentazione dipendono direttamente dalla distribuzione spaziale dell'errore di ricostruzione.

6.5 Sintesi dei risultati (ViT frozen)

Nel complesso, i risultati ottenuti con encoder Vision Transformer completamente frozen mostrano che il framework AE-XAD conserva una buona capacità di rilevazione delle anomalie a livello di immagine, ma presenta limitazioni significative nella localizzazione pixel-wise per specifiche categorie del dataset MVTec AD.

Queste evidenze indicano che l'utilizzo di un encoder Vision Transformer preaddestrato, senza adattamento al dominio, introduce un disallineamento con le assunzioni implicite del framework AE-XAD, motivate dalla necessità di errori di ricostruzione spazialmente localizzati.

6.6 Risultati quantitativi per classe (ViT trainable)

In questa sottosezione vengono riportati i risultati ottenuti utilizzando un encoder Vision Transformer completamente trainable, ottimizzato congiuntamente al decoder AE-XAD sul dataset MVTec AD. A differenza del setting frozen, in questo caso l'encoder è in grado di adattare le proprie rappresentazioni al dominio industriale durante la fase di addestramento.

La Tabella 2 riporta le prestazioni per ciascuna delle 15 classi del dataset, misurate mediante metriche image-level (X-AUC) e pixel-level (F1-score e Intersection over Union), calcolate secondo la pipeline AE-XAD invariata.

6.7 Confronto tra encoder ViT frozen e ViT trainable

Il confronto tra i due regimi di addestramento del Vision Transformer consente di valutare l'effetto dell'adattamento al dominio industriale rispetto alle limitazioni intrinseche dell'architettura.

La Tabella 3 riporta un confronto riassuntivo delle prestazioni medie ottenute sui 15 oggetti del dataset MVTec AD nei due setting considerati. Le metriche sono calcolate come media per classe delle prestazioni image-level (X-AUC) e pixel-level (F1-score e IoU).

Dai risultati emerge che il fine-tuning end-to-end dell'encoder ViT non comporta un miglioramento sistematico delle prestazioni. In particolare, lo score

Table 2: Risultati per classe su MVTec AD con encoder ViT completamente trainable.

Classe	**X-AUC**	$\mathbf{IoU}_{\max}$	$\mathbf{F1}_{\max}$
bottle	0.874	0.316	0.456
cable	0.902	0.333	0.463
capsule	0.852	0.096	0.160
carpet	0.854	0.281	0.394
grid	0.725	0.055	0.099
hazelnut	0.958	0.416	0.587
leather	0.961	0.354	0.505
metal_nut	0.905	0.337	0.475
pill	0.936	0.236	0.354
screw	0.893	0.061	0.109
tile	0.967	0.639	0.750
toothbrush	0.934	0.122	0.208
transistor	0.633	0.046	0.086
wood	0.875	0.412	0.549
zipper	0.729	0.047	0.087

Table 3: Confronto medio tra encoder ViT frozen e ViT trainable sul dataset MVTec AD.

Setting	**X-AUC (avg)**	$\mathbf{IoU}_{\mathbf{avg}}$	$\mathbf{F1}_{\mathbf{avg}}$
ViT frozen	0.910	0.311	0.425
ViT trainable	0.867	0.250	0.352

X-AUC medio risulta comparabile, ma leggermente inferiore nel setting trainable, indicando che l'adattamento al dominio non migliora la capacità di ranking globale delle anomalie.

Per quanto riguarda la localizzazione pixel-wise, il setting trainable mostra valori medi di F1-score e IoU inferiori rispetto alla configurazione frozen. Ciò indica che il fine-tuning dell'encoder Vision Transformer non è sufficiente a produrre errori di ricostruzione più compatti e meglio separabili dal rumore di fondo.

Nel complesso, questi risultati suggeriscono che il disallineamento osservato tra Vision Transformer e pipeline AE-XAD non sia imputabile esclusivamente alla mancanza di adattamento al dominio, ma rifletta una differenza più profonda nell'inductive bias dell'architettura, che non viene compensata dal training end-to-end.

Il confronto tra i due setting mostra che l'encoder ViT completamente frozen mantiene una discreta capacità di rilevazione a livello di immagine, come evidenziato dai valori medi di X-AUC. Ciò indica che le rappresentazioni pre-addestrate del Vision Transformer risultano in parte compatibili con il paradigma di anomaly detection adottato da AE-XAD.

Al contrario, il fine-tuning end-to-end dell'encoder ViT non porta a un miglioramento delle prestazioni e, in media, comporta un ulteriore degrado sia delle metriche image-level sia di quelle pixel-level. Questo suggerisce che l'adattamento al dominio industriale, nel contesto della pipeline AE-XAD, non riesca a compensare il disallineamento introdotto dall'inductive bias globale del Vision Transformer.

6.8 Confronto finale con AE-XAD originale

Table 4: Confronto medio delle prestazioni sul dataset MVTec AD tra AE-XAD originale (encoder convoluzionale), ViT frozen e ViT trainable.

Metodo	**X-AUC (avg)**	$\mathbf{F1_{avg}}$	$\mathbf{IoU_{avg}}$
AE-XAD (CNN)	0.978	0.556	0.404
ViT frozen	0.910	0.425	0.311
ViT trainable	0.867	0.352	0.250

La Tabella 4 riporta un confronto riassuntivo delle prestazioni medie ottenute sul dataset MVTec AD tra il framework AE-XAD originale, basato su encoder convoluzionale, e le due varianti che impiegano un Vision Transformer come encoder (frozen e trainable).

Dai risultati emerge in modo chiaro che AE-XAD con encoder convoluzionale supera entrambe le configurazioni basate su Vision Transformer, sia in termini di rilevazione a livello di immagine sia, in modo più marcato, per quanto riguarda la localizzazione pixel-wise delle anomalie. In particolare, il gap nelle metriche F1-score e IoU evidenzia una maggiore capacità dell'architettura convoluzionale di produrre errori di ricostruzione spazialmente compatti e ben separabili dal rumore di fondo.

Il confronto tra ViT frozen e ViT trainable mostra inoltre che il fine-tuning end-to-end dell'encoder non è sufficiente a colmare tale divario. Anzi, nel setting trainable si osserva un ulteriore peggioramento medio delle prestazioni, suggerendo che l'adattamento al dominio non compensa il disallineamento tra l'inductive bias globale del Vision Transformer e la pipeline di decisione di AE-XAD.

Nel complesso, questi risultati confermano che l'efficacia del framework AE-XAD dipende in modo critico dall'inductive bias locale dell'encoder convoluzionale, che risulta particolarmente adatto a supportare un paradigma di anomaly detection basato su ricostruzione pixel-wise e sogliatura statistica globale.

7 Analisi qualitativa delle mappe di errore

In questa sezione viene condotta un'analisi qualitativa delle mappe di errore di ricostruzione prodotte dal framework AE-XAD nelle configurazioni basate

su Vision Transformer. L'obiettivo è fornire un'interpretazione visiva dei risultati quantitativi presentati nella Sezione 6, analizzando la struttura spaziale dell'errore di ricostruzione in diverse categorie del dataset MVTec AD.

L'analisi è limitata alle configurazioni con encoder Vision Transformer (frozen e trainable), per le quali sono disponibili le heatmap pixel-wise generate durante la fase di test. Le mappe di errore del framework AE-XAD originale non sono disponibili; tuttavia, tale limitazione non compromette l'interpretazione dei risultati, poiché l'analisi è finalizzata a comprendere il comportamento interno del Vision Transformer all'interno della pipeline AE-XAD.

7.1 Caratteristiche generali delle mappe di errore ViT

Un'osservazione comune a tutte le classi del dataset è che le mappe di errore prodotte dalle configurazioni basate su Vision Transformer presentano una distribuzione spaziale dell'errore tendenzialmente più diffusa rispetto a quanto ci si aspetterebbe in un paradigma di anomaly detection basato su ricostruzione pixel-wise.

In particolare, l'errore di ricostruzione tende a manifestarsi su regioni ampie dell'immagine, anche in presenza di anomalie localizzate. Questo comportamento produce mappe di errore meno contrastate, in cui la separazione tra regioni anomale e sfondo risulta meno netta, rendendo meno efficace l'applicazione di una soglia statistica globale.

7.2 Classi con anomalie estese

Nelle classi caratterizzate da anomalie estese e strutturalmente coerenti, come *tile*, *wood* e *hazelnut*, le mappe di errore prodotte dal Vision Transformer mostrano una maggiore concentrazione dell'errore in corrispondenza delle regioni difettose.

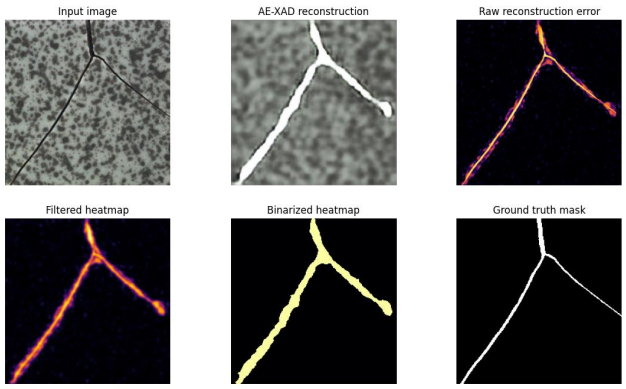


Figure 1: Esempio di mappa di errore per la classe *tile* con encoder ViT frozen. L'anomalia, di natura estesa e strutturata, produce una concentrazione visivamente riconoscibile dell'errore di ricostruzione. Nonostante una certa diffusione del segnale anche nelle regioni circostanti, la struttura spaziale del difetto risulta preservata, in accordo con i buoni valori di F1-score e IoU osservati.

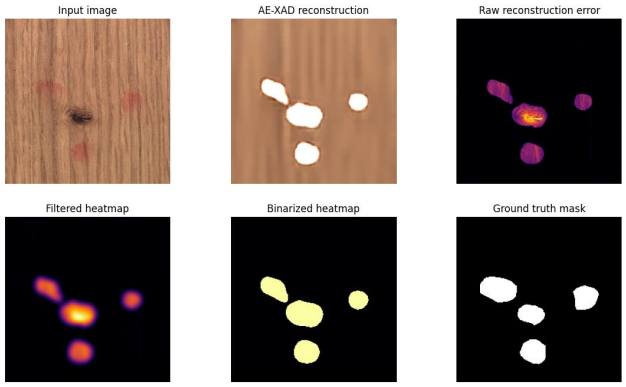


Figure 2: Mappa di errore di ricostruzione per la classe *wood* con encoder ViT frozen. L'errore risulta maggiormente concentrato in corrispondenza delle regioni difettose, sebbene a volte non perfettamente localizzato. Questo comportamento è coerente con le prestazioni quantitative relativamente elevate ottenute per questa classe.

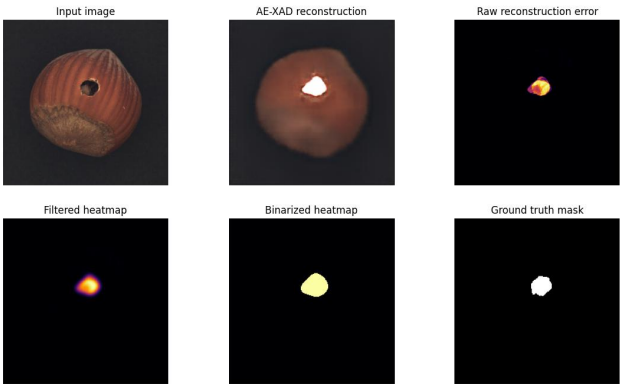


Figure 3: Esempio di heatmap per la classe *hazelnut* con encoder ViT frozen. Le anomalie estese generano una risposta di errore chiaramente distinguibile dallo sfondo, confermando che il Vision Transformer riesce a supportare la localizzazione quando il difetto presenta una struttura spaziale marcata.

In questi casi, nonostante una certa diffusione dell'errore anche nelle regioni circostanti, la struttura spaziale dell'anomalia rimane visivamente riconoscibile. Ciò è coerente con i valori relativamente elevati di F1-score e IoU osservati nelle metriche quantitative, suggerendo che il framework AE-XAD riesca a localizzare correttamente anomalie di grande estensione anche in presenza di un encoder con inductive bias globale.

7.3 Classi con anomalie sottili o localizzate

Un comportamento significativamente diverso emerge nelle classi in cui le anomalie sono sottili, filamentose o di dimensioni ridotte, come *screw*, *capsule*, *transistor* e *toothbrush*.

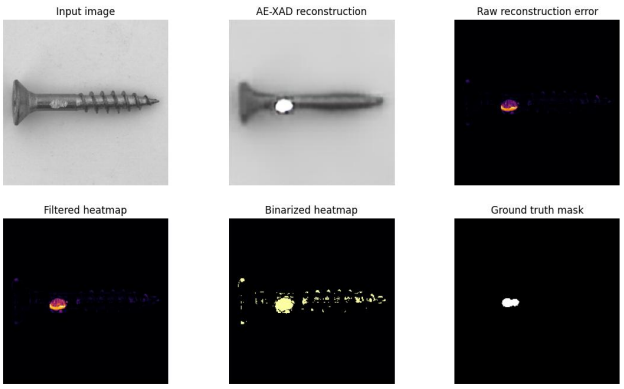


Figure 4: Mappa di errore di ricostruzione per la classe *screw* con encoder ViT frozen. Le anomalie, di dimensioni ridotte e localizzate lungo strutture sottili, non producono una concentrazione spaziale dell'errore. Il segnale risulta diffuso su gran parte dell'immagine, con un contrasto insufficiente tra regioni normali e anomale, rendendo inefficace la sogliatura statistica globale $\mu + \sigma$ e spiegando i bassi valori di F1-score e IoU osservati.

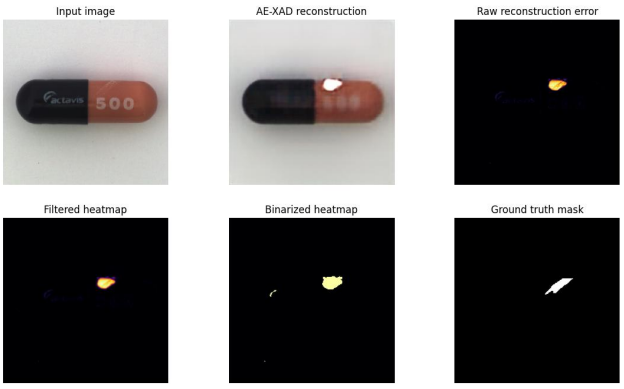


Figure 5: Heatmap di ricostruzione per la classe *capsule* con encoder ViT frozen. L'errore di ricostruzione risulta scarsamente correlato alla posizione dell'anomalia, con una distribuzione spaziale uniforme che ostacola la localizzazione pixel-wise.

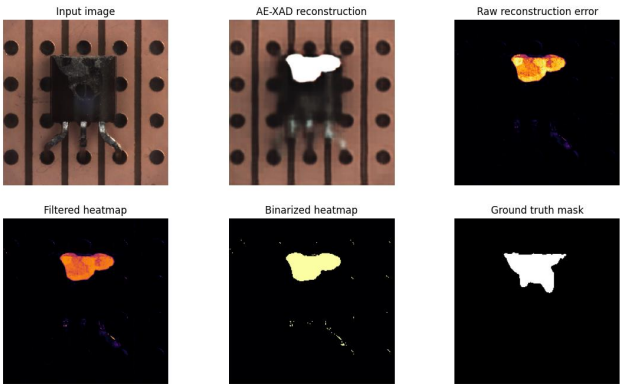


Figure 6: Mappa di errore per la classe *transistor* con encoder ViT frozen. L'anomalia, di dimensioni ridotte e ad alta frequenza spaziale, non produce una concentrazione localizzata dell'errore. Il segnale risulta diffuso su gran parte dell'immagine, rendendo inefficace la sogliatura globale e spiegando i bassi valori di F1-score e IoU osservati.

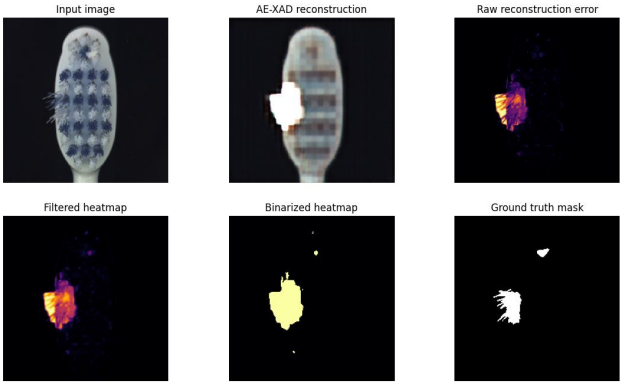


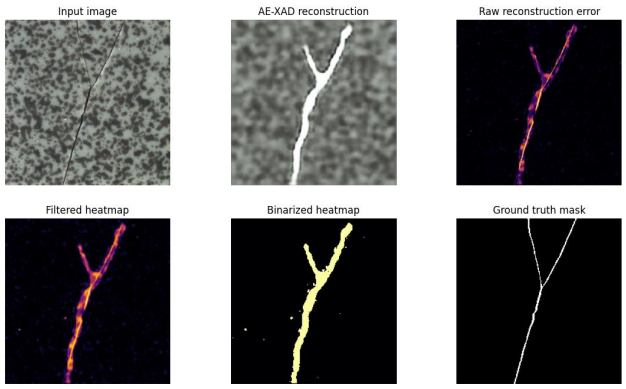
Figure 7: Esempio di mappa di errore per la classe *toothbrush* con encoder ViT frozen. La presenza di anomalie sottili e localizzate non si traduce in una risposta di errore spazialmente concentrata, confermando le difficoltà del Vision Transformer nel supportare la pipeline di localizzazione di AE-XAD in questi scenari.

In questi casi, le mappe di errore risultano fortemente diffuse e prive di una chiara concentrazione spaziale in corrispondenza delle regioni difettose. L'errore

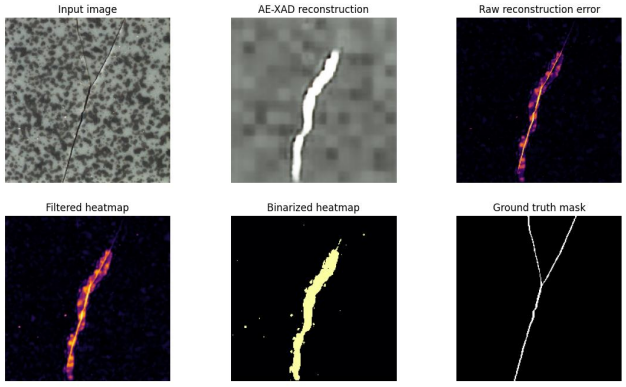
di ricostruzione si distribuisce su gran parte dell'immagine, riducendo il contrasto tra area anomala e sfondo. Di conseguenza, l'applicazione della soglia globale $\mu + \sigma$ tende a produrre segmentazioni frammentate o incomplete, in linea con i bassi valori di F1-score e IoU osservati nella Sezione 6.

7.4 Confronto qualitativo tra ViT frozen e ViT trainable

Il confronto qualitativo tra le mappe prodotte nei setting frozen e trainable evidenzia differenze sottili ma sistematiche. Nel setting frozen, l'errore di ricostruzione mantiene in alcuni casi una struttura più coerente con l'anomalia presente nell'immagine.



(a) ViT frozen

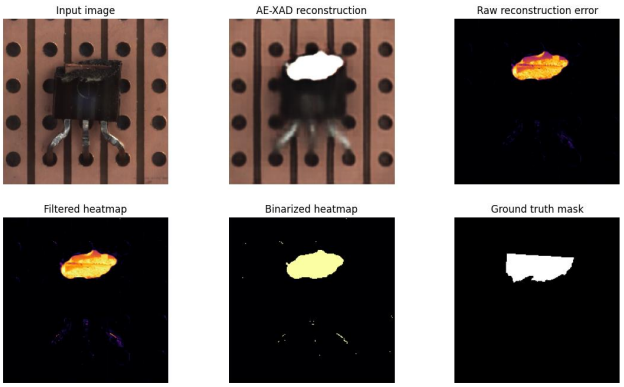


(b) ViT trainable

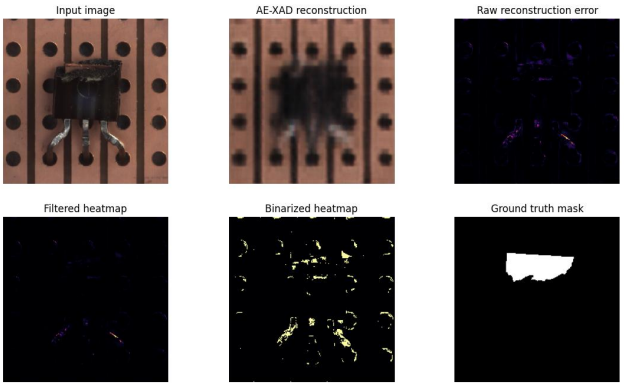
Figure 8: Confronto qualitativo tra encoder ViT frozen e trainable per la classe *tile*. Nel setting trainable si osserva una riduzione del contrasto tra regioni anomale e sfondo, con una mappa di errore più uniforme rispetto alla configurazione frozen.

Nel setting completamente trainable si osserva un comportamento qualitativamente differente. In particolare, il fine-tuning end-to-end dell'encoder Vision Transformer porta in diversi casi a una marcata riduzione, fino alla quasi soppressione, del segnale di errore associato all'anomalia. Le regioni difettose vengono ricostruite in modo più uniforme, producendo mappe di errore scarsamente correlate alla ground truth. Questo effetto è particolarmente evidente nelle classi caratterizzate da anomalie sottili o localizzate, dove l'errore residuo

risulta dominato da rumore spurio, rendendo inefficace la sogliatura statistica globale.



(a) ViT frozen



(b) ViT trainable

Figure 9: Confronto tra mappe di errore per la classe *transistor* con encoder ViT frozen e trainable. Il fine-tuning end-to-end porta a una quasi completa soppressione del segnale di errore associato all'anomalia, rendendo inefficace la localizzazione pixel-wise.

Questa osservazione qualitativa è coerente con il peggioramento medio delle metriche pixel-wise osservato nel setting trainable e suggerisce che il fine-tuning del Vision Transformer non solo non migliori la separabilità spaziale dell'errore di ricostruzione, ma in alcuni casi ne comprometta ulteriormente la correlazione

con la ground truth, rendendo inefficace la pipeline di localizzazione basata su sogliatura statistica globale.

7.5 Sintesi dell'analisi qualitativa

Nel complesso, l'analisi qualitativa delle mappe di errore conferma in modo consistente quanto emerso dai risultati quantitativi. Le configurazioni basate su Vision Transformer producono mappe di errore che risultano spesso diffuse e caratterizzate da un contrasto limitato tra regioni normali e anomale, in contrasto con le assunzioni alla base del framework AE-XAD per una localizzazione pixel-wise efficace.

In particolare, nelle classi con anomalie estese e strutturalmente coerenti, come *tile*, *wood* e *hazelnut*, il segnale di errore rimane visivamente individuabile, sebbene non sempre perfettamente concentrato. Al contrario, nelle classi caratterizzate da anomalie sottili o localizzate, l'errore di ricostruzione risulta scarsamente correlato alla posizione del difetto, rendendo difficile la separazione dal rumore di fondo tramite una soglia statistica globale.

Il confronto tra i setting frozen e trainable evidenzia inoltre che il fine-tuning end-to-end dell'encoder Vision Transformer non migliora la qualità delle mappe di errore e, in diversi casi, porta a una riduzione significativa, fino alla quasi soppressione, del segnale anomalo. Questo comportamento compromette ulteriormente la separabilità spaziale dell'errore di ricostruzione e spiega il peggioramento osservato nelle metriche di localizzazione pixel-wise.

8 Discussione e conclusioni

8.1 Discussione dei risultati

In questo lavoro è stata analizzata la compatibilità tra un encoder basato su Vision Transformer e il framework AE-XAD, mantenendo invariata l'intera pipeline di ricostruzione, scoring e localizzazione. L'obiettivo non era quello di ottimizzare le prestazioni del metodo, bensì di valutare l'impatto di un diverso inductive bias all'interno di un paradigma di anomaly detection basato su errori di ricostruzione pixel-wise e sogliatura statistica globale.

I risultati quantitativi mostrano che l'utilizzo di un Vision Transformer come encoder consente di mantenere prestazioni image-level discrete, in particolare nel setting frozen, dove i valori di X-AUC rimangono relativamente elevati per diverse classi del dataset MVTec AD. Ciò indica che le rappresentazioni preaddestrate del ViT sono in grado di supportare il ranking globale delle anomalie.

Tuttavia, le prestazioni di localizzazione pixel-wise risultano sistematicamente inferiori rispetto al framework AE-XAD originale basato su encoder convoluzionale. Questo divario emerge in modo consistente nelle metriche F1-score e IoU ed è particolarmente marcato nelle classi caratterizzate da anomalie sottili o localizzate, evidenziando un disallineamento strutturale tra le rappresentazioni prodotte dal Vision Transformer e le assunzioni alla base della pipeline

di localizzazione di AE-XAD.

8.2 Ruolo dell'inductive bias

L'analisi qualitativa delle heatmap fornisce una chiave di lettura fondamentale per interpretare i risultati quantitativi osservati. Le architetture convoluzionali incorporano un forte inductive bias locale, che favorisce la modellazione di pattern spaziali e la produzione di errori di ricostruzione compatti e ben localizzati. Tale proprietà risulta intrinsecamente coerente con le assunzioni alla base del framework AE-XAD, in cui la localizzazione delle anomalie si basa sull'applicazione di una soglia statistica globale a una mappa di errore spaziale.

Al contrario, il Vision Transformer è progettato per modellare relazioni globali tra regioni dell'immagine, riducendo l'enfasi sulla località spaziale. Questo diverso inductive bias si riflette in heatmap più diffuse e meno contrastate, che risultano meno compatibili con una pipeline di localizzazione basata su sogliatura globale.

In questo contesto, il limite osservato non è riconducibile a una scarsa capacità rappresentazionale del Vision Transformer, bensì a un disallineamento strutturale tra le proprietà dell'encoder e il meccanismo decisionale adottato da AE-XAD, che risulta fortemente dipendente da una rappresentazione spazialmente localizzata dell'errore di ricostruzione.

8.3 Frozen vs trainable: implicazioni

Il confronto tra i due regimi di addestramento del Vision Transformer fornisce ulteriori indicazioni sul ruolo dell'encoder all'interno del framework AE-XAD. Nel setting frozen, l'encoder ViT mantiene rappresentazioni pre-addestrate che, pur non ottimali per la localizzazione pixel-wise, conservano una certa correlazione con le anomalie presenti nell'immagine.

Nel setting completamente trainable, invece, il fine-tuning end-to-end tende in diversi casi a ridurre ulteriormente il segnale di errore associato alle regioni anomale, fino alla quasi soppressione dello stesso. Questo comportamento suggerisce che l'adattamento al dominio industriale non solo non compensi il disallineamento strutturale introdotto dall'inductive bias globale del Vision Transformer, ma possa accentuarlo all'interno di una pipeline basata su ricostruzione.

Questa evidenza rafforza l'ipotesi che le limitazioni osservate non siano attribuibili a una mancanza di capacità del modello o a un regime di training sub-ottimale, bensì a una incompatibilità più profonda tra la natura delle rappresentazioni apprese dal Vision Transformer e il meccanismo di decisione pixel-wise adottato da AE-XAD.

8.4 Limiti del lavoro

Il presente lavoro presenta alcuni limiti che devono essere esplicitamente riconosciuti. In primo luogo, l'analisi si concentra esclusivamente sulla sostituzione dell'encoder, mantenendo invariata la pipeline AE-XAD originale. Di

conseguenza, non viene esplorata la possibilità di adattare il meccanismo di scoring o la sogliatura statistica per renderli più compatibili con architetture basate su attenzione globale.

Inoltre, l'analisi qualitativa si basa sulle heatmap prodotte dalle configurazioni con Vision Transformer, senza un confronto visivo diretto con le heatmap del framework AE-XAD originale. Tuttavia, tale limitazione non compromette le conclusioni del lavoro, poiché l'obiettivo principale è valutare la compatibilità tra encoder e pipeline decisionale piuttosto che confrontare visivamente due metodi differenti.

8.5 Conclusioni e prospettive future

In conclusione, questo lavoro mostra che la sostituzione dell'encoder convoluzionale di AE-XAD con un Vision Transformer, mantenendo invariata la pipeline di ricostruzione e localizzazione, non risulta efficace per il compito di anomaly detection pixel-wise in ambito industriale. I risultati evidenziano che l'inductive bias locale delle architetture convoluzionali rappresenta un elemento chiave per il successo del framework AE-XAD.

Le architetture basate su attenzione globale possono fornire rappresentazioni utili per il ranking delle anomalie, ma richiedono una riprogettazione del meccanismo di scoring e localizzazione per essere pienamente sfruttate in un contesto di anomaly detection basato su ricostruzione.

Come prospettive future, risulta pertanto di particolare interesse lo studio di pipeline ibride, in cui encoder basati su Vision Transformer siano affiancati da meccanismi di localizzazione adattivi o da strategie di sogliatura non globali, al fine di riallineare le proprietà rappresentazionali del modello con gli obiettivi del task.

References

[1] F. Angiulli, F. Fassetti, L. Ferragina, and S. Nisticò, "Explaining anomalies through semi-supervised autoencoders," *Array*, 2025.