Università degli Studi di Milano Bicocca
Dipartimento di Informatica, Sistemistica e Comunicazione
Corso di Laurea Magistrale in Informatica

Weakly Supervised Video Anomaly Detection using Deep Learning

Relatore: Prof. Paolo Napoletano
Co-relatore: Prof. Raimondo Schettini
Co-relatore: Dott. Flavio Piccoli

Magazzù Gaetano
Matricola 829685

ANNO ACCADEMICO 2022-2023

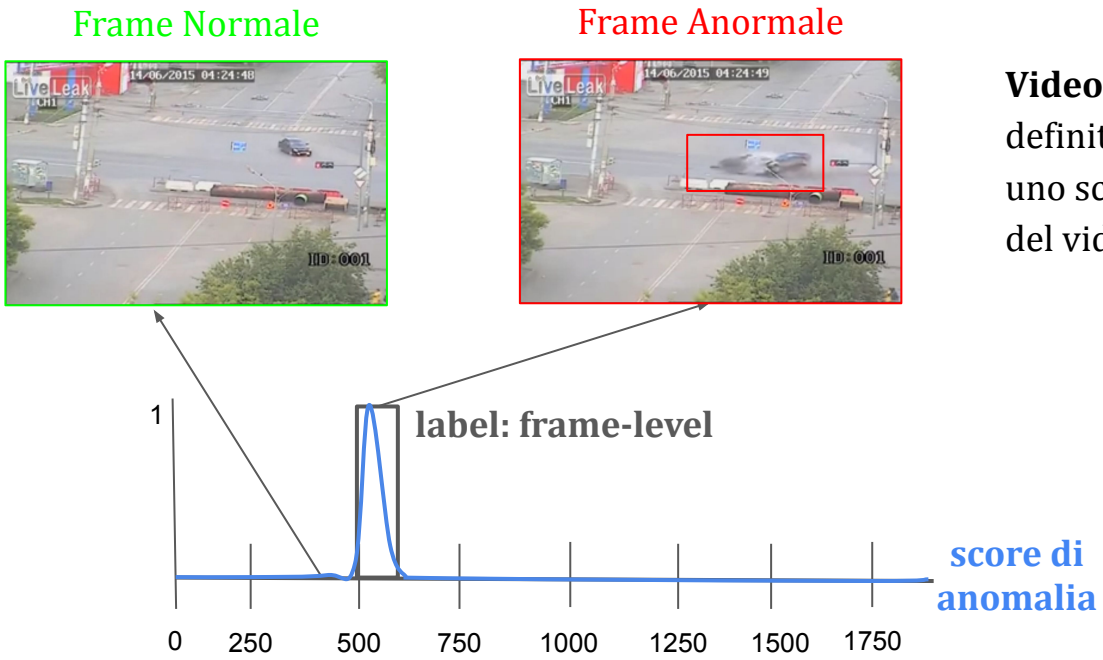
Introduzione

Un'**anomalia** può essere definita come quegli esempi, comportamenti, attività o eventi rari che si discostano significativamente dalla maggior parte dei dati, ossia dai casi normali. (Park et al. 2023)

Possibili Utilizzi e Applicazioni:

- **pubblica sicurezza:** rilevare crimini o incidenti per intervenire prontamente
- **sui social o sul web:** filtrare, segnalare video violenti
- **sicurezza sul lavoro:** rilevare cadute, intrusioni a zone vietate, veicoli non autorizzati o ostacoli presenti vicino ad uscite di emergenza

Definizione del Problema



Video Anomaly Detection: viene definito come task di regressione di uno score di anomalia per ogni frame del video.

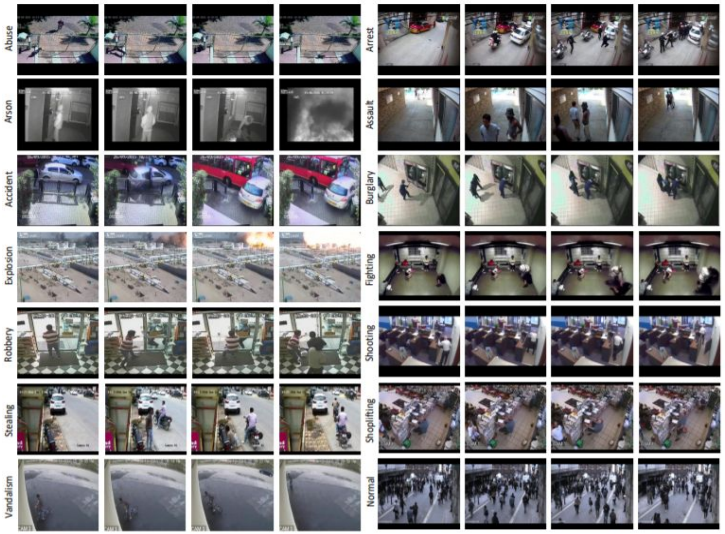
esempio di predizione su un video di un incidente stradale

Principali Dataset di Benchmark

Shanghai-Tech

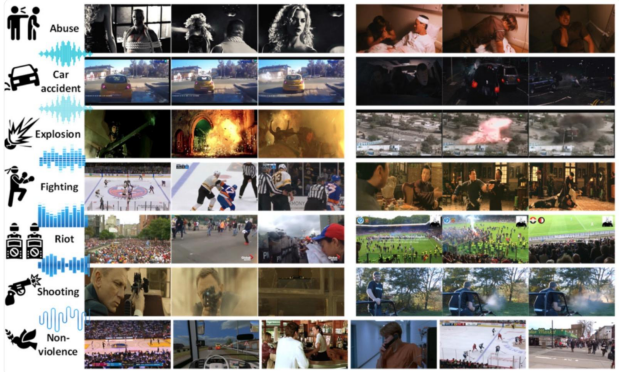
(Deshpande et al. 2022)

UCF-Crime



(Sultani et al. 2018)

XD-Violence



(Wu et al. 2020)

dataset	anno	# video	durata	fps	risoluzione media	# anomalie	sorgente dati
Shanghai-Tech	2017	437	317.398 frame	24	856 × 480	11	telecamere CCVT
UCF-Crime	2018	1900	128 ore	30	320 x 240	13	telecamere CCVT
XD-Violence	2020	4754	217 ore	24	623,49 x 326,83	6	telecamere CCVT, di auto o a mano, sport, giochi e film

Statistiche dei Dataset

Problematiche del Rilevamento di Anomalie nei Video

scene in ambienti “in the wild”

untrimmed video

trimmed video

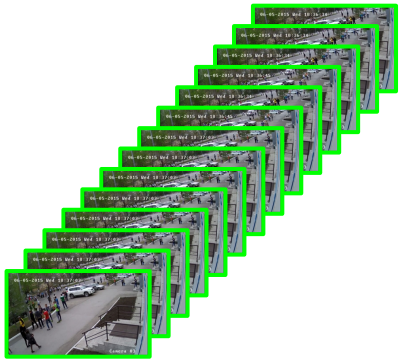
contesto

Anomalia

Normale

Stato dell'Arte

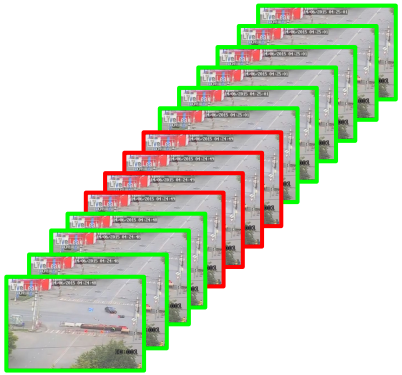
Non Supervisionati
(solo video normali)



Video Normale

- solitamente trattati come metodi one-class
- risultati inferiori

Supervisionati
(etichette frame-level)



Video Anormale

- difficili da mettere in pratica
- teoricamente ottengono i risultati migliori

Debolmente Supervisionati
(etichette video-level)

Video Anormale

- approccio più pratico
- risultati migliori

Stato Dell'Arte - Weakly Supervised

MIL-Based: Quasi tutti i lavori si basano sull' **MIL(Multiple Instance Learning)**, proponendo diversi vincoli e strategie per la selezione delle istanze anomale

Coarse-Grained vs Fine-Grained: i video sono suddivisi in frammenti e molto raramente in clip, oppure tublet (patch di clip)

Encoder-Agnostic vs Encoder-Based: solitamente i metodi usano feature pre-addestrate per il riconoscimento di attività, ed in pochi addestrano anche le feature

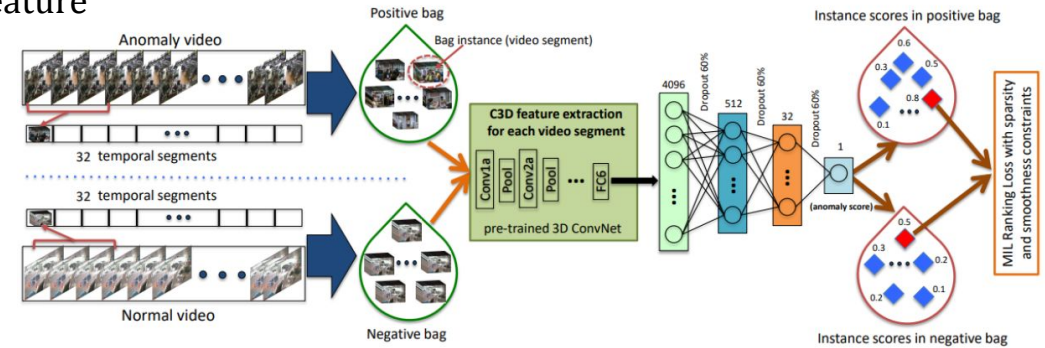
per mitigare gli errori dei metodi **MIL**:

- **Self-Training**
- **Multitask**
- **Ibridi**

Gestione Domain Gap: metodi **Encoder-Based** o feature pre-addestrate più sofisticate (es. **CLIP**)

Gestione Contesto:

- Convoluzioni dilatate
- GCN(Graph Convolutional Network)
- Moduli di Self-Attention e Transformer



schema metodo (Sultani et al. 2018)

Modello Base - RTFM(Robust Temporal Feature Magnitude)

Modello basato su **MIL** e **Encoder-Agnostic**

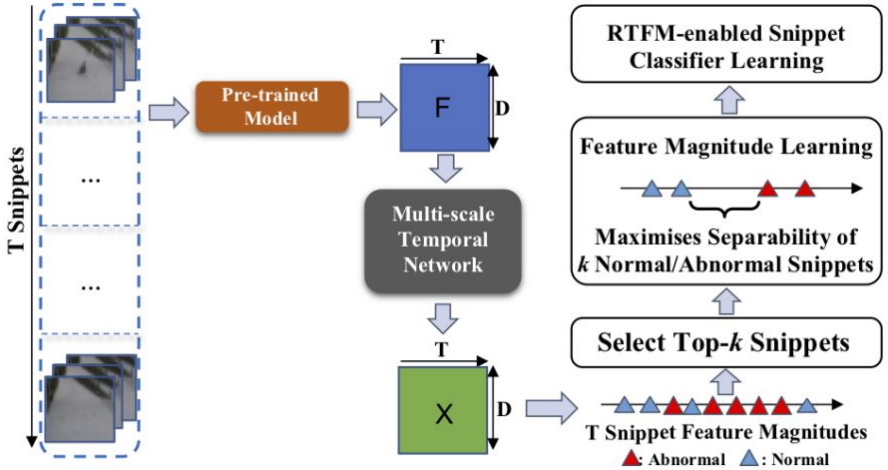
MTN(Multi-Scale Temporal Network):

- **dipendenze temporali locali**
  (piramide di convoluzioni dilatate)
- **dipendenze temporali globali**
  (Non Local-Block basato su self-attention)

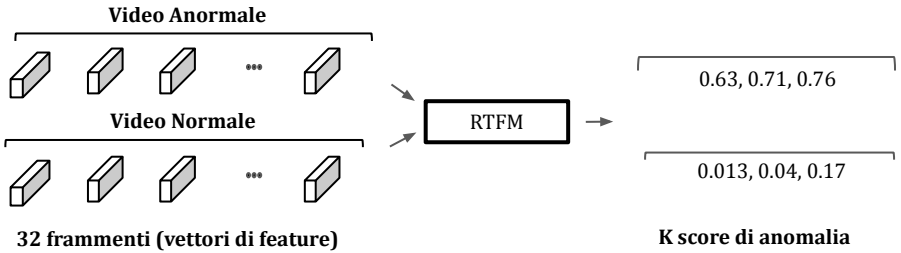
Top-k: frammenti con magnitudine media più alta delle feature

Feature Magnitude Learning: massimizza con un certo margine la distanza tra le magnitudini medie delle feature dei top-k frammenti anormali e dei top-k frammenti normali

Snippet-Classifier: classifica i top-k frammenti

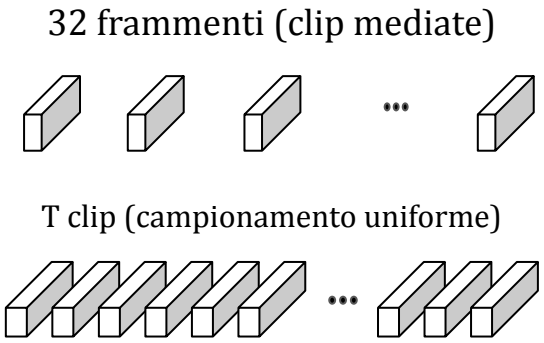


schema metodo (Tian et al. 2021)



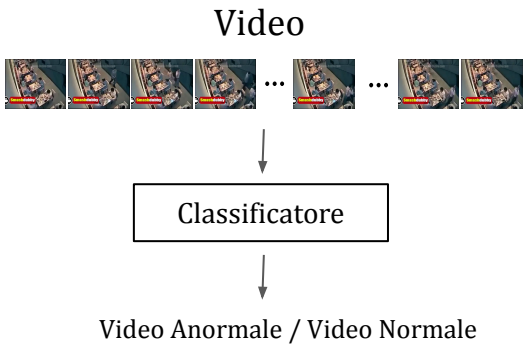
Metodo Proposto

Fine-Grained Input



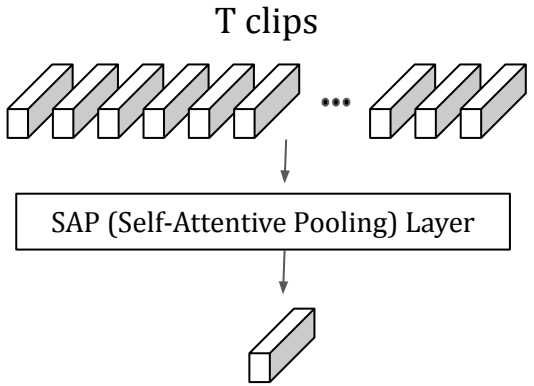
- dovrebbe riconoscere meglio le anomalie brevi
- le clip anormali non sono sopraffatte da quelle normali
- dimensione variabile ed efficienza

Video Classification



- task ausiliario per aiutare il task principale
- task più semplice (supervisionato e i dati non sono sbilanciati)

Aggregazione Temporale



- aggrega l'informazione temporale, imparando l'importanza delle singole clip tramite self-attention

Esperimenti

- **Dataset:** Shanghai-Tech, UCF-Crime e XD-Violence
- **Split:** Train e Test
- **Etichette:** video-level (train), frame-level (test)
- **Metriche:** AUC e AP
- **Feature:** C3D e I3D

Iper-parametri (fissati):

- 5000 iterazioni simile a Fan et al. 2023
- Come Tian et al. 2021:
  - learning rate: 1e-3 Shanghai-Tech, UCF-Crime, 1e-4 XD-Violence
  - weight-decay: 5e-4
  - batch-size: 32
- Video Classification: learning rate: 1e-4, weight-decay: 5e-4

Tuning Iper-parametri:

- Tuning sul test (Park et al. 2023, Fan et al. 2023)
- Per I3D Grid-Search con 1000 iterazioni (per avere tempi gestibili di addestramento)
- Stesse configurazioni ottenute per I3D per C3D (C3D richiede più tempo di train rispetto a I3D)
- Un'ulteriore prova su XD-Violence con C3D

Esperimenti

Configurazioni considerate per prove con I3D

iperparametri	spazio di ricerca
k	{3, 5, 7, 9, 11}
m	{50, 100, 150}
vc_alpha	{1e-2, 1e-3, 1e-4}
n_clips	{16, 32, 64} per Shanghai-Tech, altrimenti {64, 128, 256}
sap_size	{128, 256, 512}

Configurazioni limitate per C3D su XD-Violence

iperparametri	spazio di ricerca
k	{3, 5, 7}
m	{100}
vc_alpha	{1e-2, 1e-3, 1e-4}
n_clips	{64, 128}
sap_size	{128, 256, 512}

Statistiche sul numero di clip (Video Anormali di Train)

Dataset	Minimo	Mediana	Media	Massimo
Shanghai-Tech	13	22	23,89	59
UCF-Crime	7	134	245,72	8869
XD-Violence	2	112	173,69	8581

Risultati - Confronto con Modello Base

Video Anomaly Detection

Shanghai-Tech

Modello	Feature	Iperparametri	AUC%
Modello Base	C3D	k=3, m=100 e 32 frammenti	91,51
Modello Proposto	C3D	k=5, m=50, vc_alpha=1e-4, n_clips=64, sap_size=256	**97,03 (+5,52)**
Modello Base	I3D	k=3, m=100 e 32 frammenti	97,21
Modello Proposto	I3D	k=5, m=50, vc_alpha=1e-4, n_clips=64, sap_size=256	**97,98 (+0,77)**

UCF-Crime

Modello	Feature	Iperparametri	AUC%
Modello Base	C3D	k=3, m=100 e 32 frammenti	83,28
Modello Proposto	C3D	k=11, m=100, vc_alpha=1e-3, n_clips=64, sap_size=256	**85,05 (+1,77)**
Modello Base	I3D	k=3, m=100 e 32 frammenti	84,3
Modello Proposto	I3D	k=11, m=100, vc_alpha=1e-3, n_clips=64, sap_size=256	**85,56 (+1,26)**

XD-Violence

Modello	Feature	Iperparametri	AP%
Modello Base	C3D	k=3, m=100 e 32 frammenti	**75,89**
Modello Proposto	C3D	k=3, m=100, vc_alpha=1e-4, n_clips=64, sap_size=256	75,48 (-0,41)
Modello Base	I3D	k=3, m=100 e 32 frammenti	77,81
Modello Proposto	I3D	k=5, m=50, vc_alpha=1e-2, n_clips=128, sap_size=256	**80,76 (+2,95)**

usando la stessa configurazione di I3D con C3D si otteneva un risultato di 70,42 AP%

Video Classification

Shanghai-Tech

Modello	Feature	AUC %
Modello Proposto	C3D	98,57
Modello Proposto	I3D	99,75

UCF-Crime

Modello	Feature	AUC %
Modello Proposto	C3D	95,70
Modello Proposto	I3D	92,90

XD-Violence

Modello	Feature	AP %
Modello Proposto	C3D	96,21
Modello Proposto	I3D	98,78

Risultati - Confronto con Stato Dell'Arte

Shanghai-Tech

Nome Metodo	Anno	Feature	AUC %
Zhongh et. al.	2019	C3D	76,44
Zhang et. al.	2019	C3D	82,50
Sultani et. al	2018	C3D	83,47
ARNET	2020	C3D	85,01
CLAWS	2020	C3D	89,67
RTFM	2021	C3D	91,51
MIST	2021	C3D	93,13
MSL	2022	C3D	94,81
NG-MIL	2023	C3D	96,02
CNN-ViT	2023	C3D	96,08
LSTC	2023	C3D	96,56
Metodo Proposto	**2023**	**C3D**	**97,03**
CLIP-TSA	2022	C3D	97,19
ARNET	2020	I3D Flow	82,33
Zhongh et. al.	2019	TSN Flow	84,13
Zhongh et. al.	2019	TSN	84,44
Sultani et. al	2018	I3D	85,33
ARNET	2020	I3D	85,38
ARNET	2020	I3D + Flow	91,24
MIST	2021	I3D	94,83
MSL	2022	I3D	96,08
RTFM	2021	I3D	97,21
CNN-ViT	2023	I3D	97,43
NG-MIL	2023	I3D	97,43
S3R	2022	I3D	97,48
LSTC	2023	I3D	97,92
CLIP-TSA	2022	I3D	97,98
Metodo Proposto	**2023**	**I3D**	**97,98**
PEL	2023	I3D	98,14
MSL	2022	VideoSwin	97,32
CNN-ViT	2023	CLIP	98,06
CNN-ViT	2023	CLIP + C3D	98,13
CLIP-TSA	2022	CLIP	98,32
CNN-ViT	2023	CLIP + I3D	98,66

UCF-Crime

Nome Metodo	Anno	Feature	AUC %
MotionAware	2019	C3D	74,40
Sultani et. al	2018	C3D	75,41
Zhang et. al.	2019	C3D	78,70
Zhongh et. al.	2019	C3D	81,08
MIST	2021	C3D	81,40
MSL	2022	C3D	82,85
CLAWS	2020	C3D	83,03
RTFM	2021	C3D	83,28
NG-MIL	2023	C3D	83,43
LSTC	2023	C3D	83,47
CLIP-TSA	2022	C3D	83,94
Metodo Proposto	**2023**	**C3D**	**85,05**
CNN-ViT	2023	C3D	85,78
MotionAware	2019	I3D	75,40
Sultani et. al	2018	I3D	77,92
Zhongh et. al.	2019	TSN Flow	78,08
MotionAware	2019	PWC-Flow	79,00
MotionAware	2019	I3D + PWC-Flow	79,80
Zhongh et. al.	2019	TSN	82,12
MIST	2021	I3D	82,30
Wu et. al.	2020	I3D	82,44
RTFM	2021	I3D	84,30
CLIP-TSA	2022	I3D	84,66
MSL	2022	I3D	85,30
Metodo Proposto	**2023**	**I3D**	**85,56**
NG-MIL	2023	I3D	85,63
LSTC	2023	I3D	85,88
S3R	2022	I3D	85,99
CNN-ViT	2023	I3D	86,50
PEL	2023	I3D	86,76
MGFN	2023	I3D	86,98
MSL	2022	VideoSwin	85,62
MGFN	2023	VideoSwin	86,67
CLIP-TSA	2022	CLIP	87,58
CNN-ViT	2023	CLIP	87,63
CNN-ViT	2023	C3D + CLIP	88,02
CNN-ViT	2023	I3D + CLIP	88,97

XD-Violence

Nome Metodo	Anno	Feature	AP %
Sultani et. al	2018	C3D	73,20
Metodo Proposto	**2023**	**C3D**	**75,48**
MSL	2022	C3D	75,53
RTFM	2021	C3D	75,89
NG-MIL	2023	C3D	75,91
CLIP-TSA	2022	C3D	77,66
Wu et. al.	2020	I3D	75,41
Sultani et. al	2018	I3D	75,68
RTFM	2021	I3D	77,81
CLIP-TSA	2022	I3D	78,19
MSL	2022	I3D	78,28
NG-MIL	2023	I3D	78,51
Wu et. al.	2020	I3D + VGG-ish	78,64
MGFN	2023	I3D	79,19
S3R	2022	I3D	80,26
Metodo Proposto	**2023**	**I3D**	**80,76**
PEL	2023	I3D	85,59
MSL	2022	VideoSwin	78,58
MGFN	2023	VideoSwin	80,11
CLIP-TSA	2022	CLIP	82,19

- Miglior Risultato
- Secondo Miglior Risultato
- Terzo Miglior Risultato

Risultati - Analisi Modifiche Effettuate

Modello	Shanghai-Tech (AUC%)	UCF-Crime (AUC%)	XD-Violence (AP%)
RTFM + VC + SAP (clip)	**97,98 (+0,77)**	**85,56 (+1,26)**	**80,76 (+2,95)**
RTFM + VC + GAP (clip)	97, 88 (+0,67)	85,24 (+0,94)	79,13 (+1,32)
RTFM + VC + SAP (frammenti)	97,78 (+0,57)	85,47 (+1,17)	78,35 (+0,54)
RTFM (frammenti)	97,21	84,3	77,81

GAP(Global Average Pooling): aggrega il tempo mediando le clip, le quali hanno uguale importanza.

SAP(Self-Attentive Pooling): aggrega il tempo imparando l'importanza delle singole clip tramite **self-attention.**

risultati del tuning del metodo proposto con i frammenti

Dataset	Configurazioni Iperparametri
Shanghai-Tech	k=3, m=150, vc_alpha=1e-2, sap_size=256
UCF-Crime	k=7, m=150, vc_alpha=1e-2, sap_size=256
XD-Violence	k=5, m=150, vc_alpha=1e-4, sap_size=256

Risultati - Analisi di Efficienza

- Vengono stimati i tempi di inferenza: (mediati su 30 prove)
- Considerando una clip da 16 frame, 224 x 224, con 10 crop augmentation.
- In **CPU** non si riesce a processare nemmeno una clip al secondo, con meno di 16 FPS.
- In **GPU** si ottengono tempi real-time: processando 10,42 clip/s e 166,70 FPS

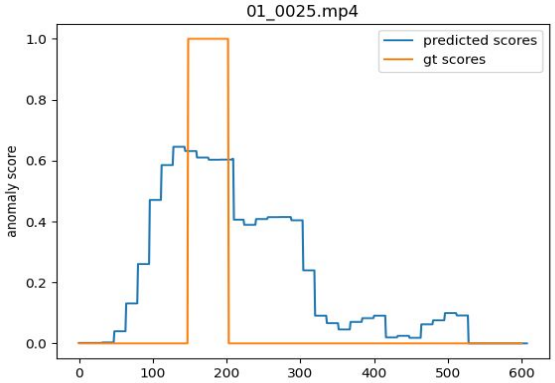
Tempi di inferenza su una clip

Modello	CPU (ms)	GPU (ms)
I3D	1439	95,32
Modello Proposto	2,09	0,66
I3D + Modello Proposto	1441,09	95,98

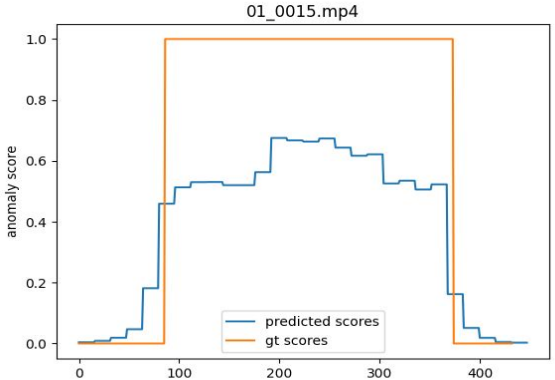
CPU: AMD Ryzen 7 5800 X, GPU: RTX 3060 Ti

Risultati - Esempi di Predizione

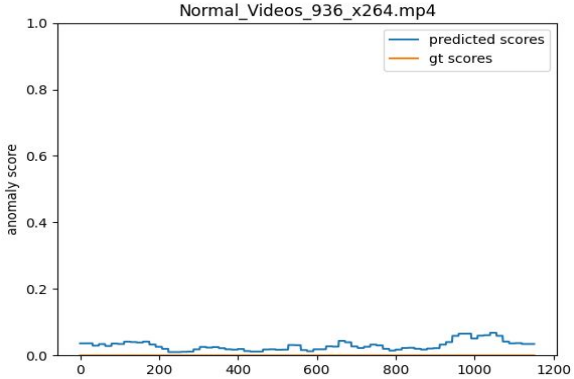
bike



skateboard

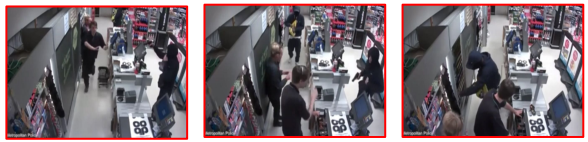


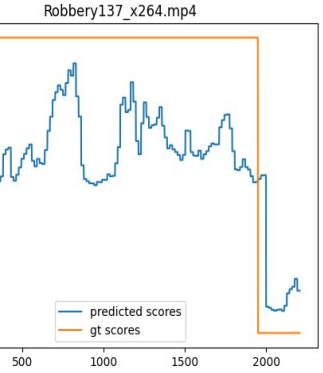
normal



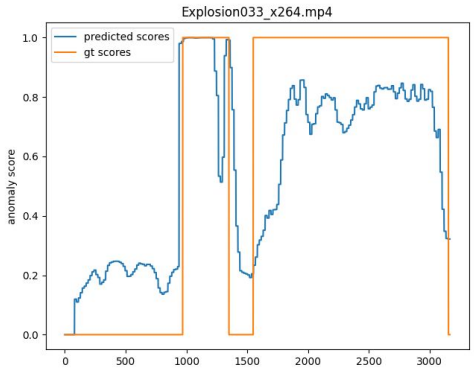
Risultati - Esempi di Predizione

robbery robbery robbery

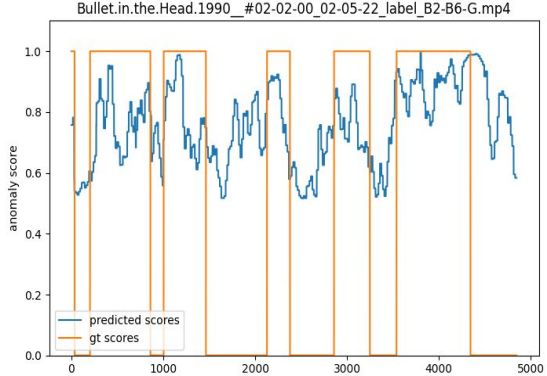




explosion explosion



car accident shooting explosion



Conclusioni e Sviluppi Futuri

Limitazioni:

- Il campionamento delle clip è rischioso, comporta perdita e duplicazione di informazione
- Il campionamento uniforme delle clip comporta salti temporali
- Il modello non è pensato per streaming video
- Le feature pre-addestrate non sono task-specific

Sviluppi Futuri:

- Campionare insiemi di clip consecutive
- Task di video classification multiclasse
- Considerare altre feature
- Considerare tublet (patch di clip)
- Fine-tuning delle feature

Sfide Aperte:

- Video a bassa risoluzione
- Mancanza etichette frame-level
- Metodo di valutazione
- Rilevamento spaziale delle anomalie
- Rilevamento online delle anomalie

Grazie dell’attenzione

Riferimenti Bibliografici

- Park, Seongheon, et al. "Normality Guided Multiple Instance Learning for Weakly Supervised Video Anomaly Detection." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023.

- Deshpande, Kapil, et al. "Anomaly detection in surveillance videos using transformer based attention model." International Conference on Neural Information Processing. Singapore: Springer Nature Singapore, 2022.

- Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

- Wu, Peng, et al. "Not only look, but also listen: Learning multimodal violence detection under weak supervision." *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer International Publishing, 2020.

- Tian, Yu, et al. "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

- Fan, Yidan, et al. "Weakly-Supervised Video Anomaly Detection with Snippet Anomalous Attention." *arXiv preprint arXiv:2309.16309* 2023.

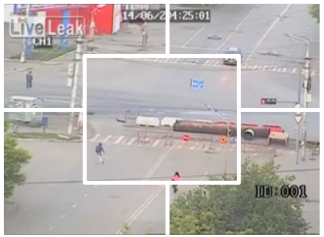
Slides - Extra

Feature per il Riconoscimento di Attività

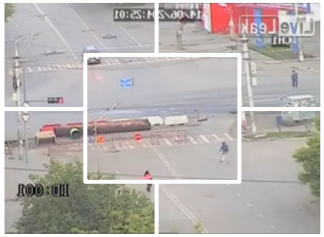
C3D (Convolutional 3D) è una **CNN(Convolutional Neural Network) 3D** addestrata **from-scratch** sul dataset **Sports-1M**.

I3D (Inflated 3D Conv-Net) è basato su **Inception-V1** precedentemente addestrata su **ImageNet**. Viene addestrato tramite **fine-tuning** su **Kinetics-400** espandendo i pesi 2D in 3D.

10 Crop Augmentation

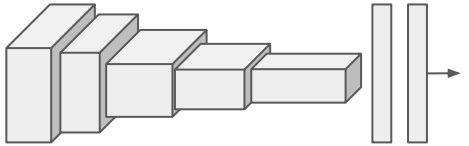


5 crop frame originale

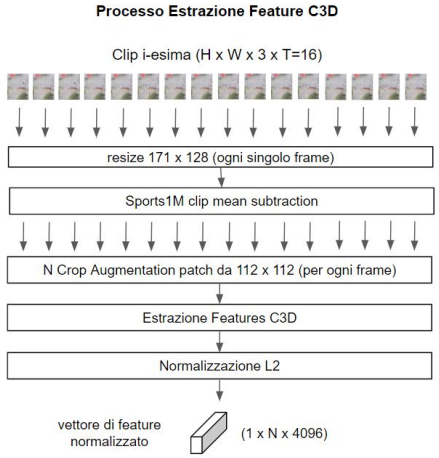


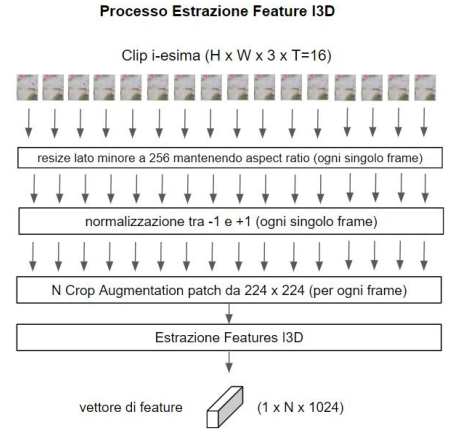
5 crop frame specchiato

Classificatore di Attività

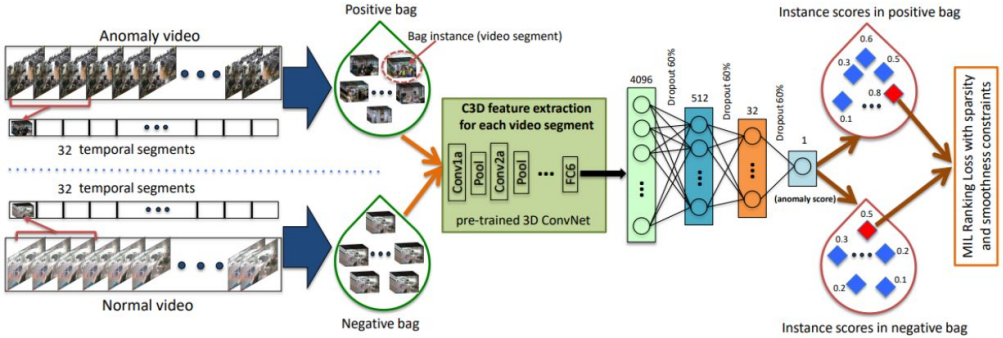


running





Deep MIL Ranking Framework



schema metodo (Sultani et al. 2018)

- score distanti tra frammenti normali e anormali
- score alti per i frammenti anomali
- score bassi per i frammenti normali

$$L(B_a, B_n) = \max(0, 1 - \max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) + \max_{i \in \mathcal{B}_n} f(\mathcal{V}_n^i))$$

$$+ \lambda_1 \sum_{i=1}^{n-1} (f(\mathcal{V}_a^i) - f(\mathcal{V}_a^{i+1}))^2$$ vincolo di smoothness (tra score vicini)

$$+ \lambda_2 \sum_{i=1}^{n} f(\mathcal{V}_a^i)$$ vincolo sulla sparsità delle anomalie

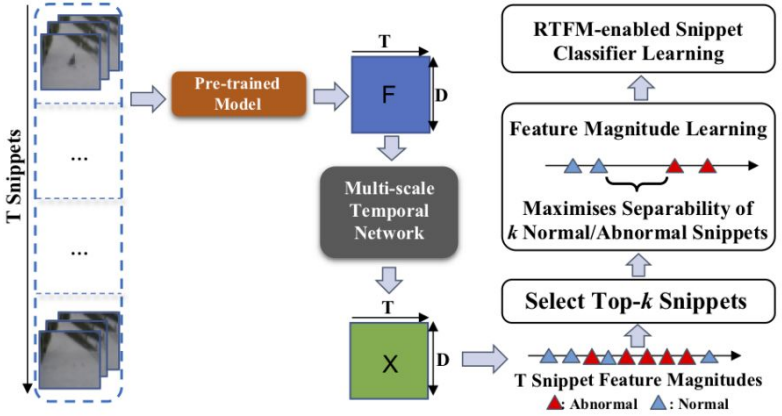
frammento i-esimo nella positive bag

$$\max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) > \max_{i \in \mathcal{B}_n} f(\mathcal{V}_n^i)$$

frammento i-esimo nella negative bag

RTFM (Robust Temporal Feature Magnitude)

- Il Classificatore è un **MLP(Multi Layer Perceptron)** addestrato tramite **Binary Cross Entropy**.
- **FML(Feature Magnitude Learning):**
  - basato sulla magnitudine media delle feature dei top-k frammenti
  - massimizza la separabilità dei frammenti normali e anormali con un certo margine



schema metodo (Tian et al. 2021)

feature pre-addestrate: $\mathbf{F} \in \mathcal{F} \subset \mathbb{R}^{T \times D}$ T frammenti, D numero canali delle feature

estrattore di feature temporali: $s_\theta : \mathcal{F} \rightarrow \mathcal{X}$ con $\mathcal{X} \subset \mathbb{R}^{T \times D}$

feature temporali estratte da un video: $X = s_\theta(\mathbf{F})$

$$g_{\theta,k}(X) = \max_{\Omega_k(X) \subseteq \{x_t\}_{t=1}^{T}} \frac{1}{k} \sum_{x_t \in \Omega_k(X)} ||X_t||_2$$

$$d_{\theta,k}(X^+, X^-) = g_{\theta,k}(X^+) - g_{\theta,k}(X^-)$$

$$FML(X^+, X^-) = max(0, m - d_{\theta,k}(X^+, X^-))$$

bag positiva bag negativa

Aggregazione Temporale

GAP (Global Average Pooling) Layer:

$$\mu = \frac{1}{T}\sum_{t=1}^{T}\mathbf{h}_t$$ vettore di output

$$h_t \in \mathbb{R}^D$$

vettore di feature di dimensione D per la t-esima clip, con un numero totale di T clip

SAP (Self-Attentive Pooling) Layer: (okabe et al. 2018)

$$\mathbf{e}_t = \mathbf{v}^T f(Wh_t + b) + k$$ score di attention

$$\mathbf{a}_t = \frac{\exp(\mathbf{e}_t)}{\sum_{t=1}^{T}\exp(\mathbf{e}_t)}$$ score di attention normalizzato

$$\hat{\mu} = \sum_{t=1}^{T}\mathbf{a}_t \times \mathbf{h}_t$$ vettore di output

- **f(.)** è una funzione di attivazione (**tanh**)
- i vari parametri sono i pesi del layer lineare **W** ed il bias **b**, il context vector **v** ed uno scalare **k**

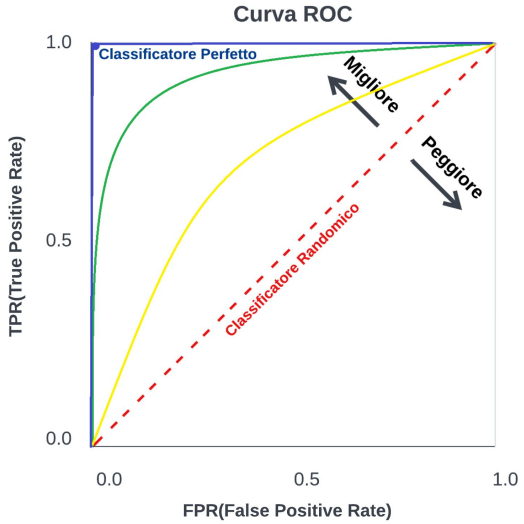
Okabe, Koji, Takafumi Koshinaka, and Koichi Shinoda. "Attentive statistics pooling for deep speaker embedding." *arXiv preprint arXiv:1803.10963* (2018).

Metriche di Valutazione

Curva **ROC(Receiver Operating Characteristic)**: grafico che mostra il **TPR** e **FPR** a diversi threshold

AUC(Area Under the ROC Curve): area sotto la curva **ROC**

AP(Average Precision): area sotto la curva di **Precision-Recall**



TPR(True Positive Rate) o Recall:
è il rapporto tra i veri positivi **(TP)** e i positivi totali (effettivi)

FPR(False Positive Rate) o Specificity:
è il rapporto tra i veri negativi **(TN)** e i negativi totali (effettivi)

Precision: è il rapporto tra i veri positivi **(TP)** e il totale degli esempi (predetti) positivi

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP})$$

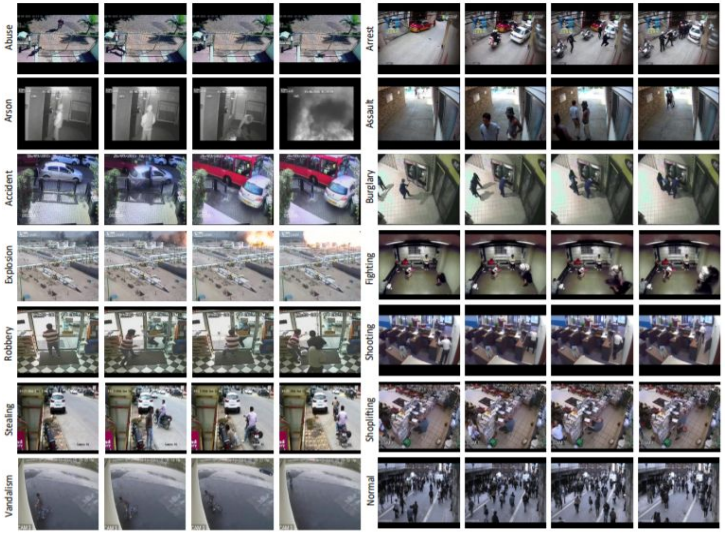
$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$$

Principali Dataset di Benchmark

Shanghai-Tech

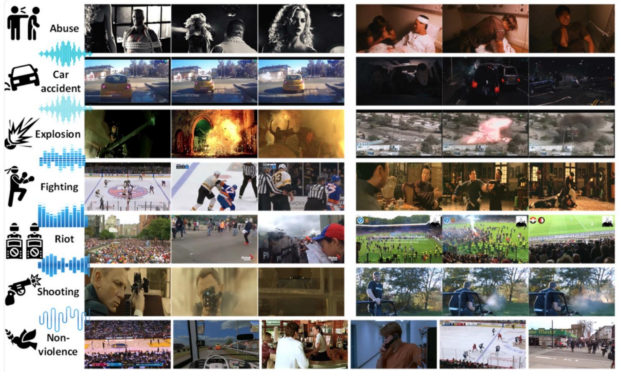
(Deshpande et al. 2022)

UCF-Crime



(Sultani et al. 2018)

XD-Violence



(Wu et al. 2020)

dataset	anno	# video	durata	fps	risoluzione media	# anomalie	sorgente dati
Shanghai-Tech	2017	437	317.398 frame	24	856 × 480	11	telecamere CCVT
UCF-Crime	2018	1900	128 ore	30	320 x 240	13	telecamere CCVT
XD-Violence	2020	4754	217 ore	24	623,49 x 326,83	6	telecamere CCVT, di auto o a mano, sport, giochi e film

Statistiche dei Dataset

dataset	# video	train	test
Shanghai-Tech	437	238 (63 A / 175 N)	199 (44 A / 155 N)
UCF-Crime	1900	1610 (810 A / 800 N)	290 (140 A / 150 N)
XD-Violence	4754	3954 (1905 A / 2049 N)	800 (500 A / 300 N)

Split Train/Test (A:anormali, N:normali)