

# Sviluppo di strategie di deep learning per applicazioni di videosorveglianza

Michele Dalla Chiara - VR464051

Davide Zampieri - VR458470

A.A. 2021 - 2022

## PRIMA PARTE

- Framework HRNet

La maggior parte dei framework per la stima della posa umana esistenti ottengono una rappresentazione ad alta risoluzione a partire da quelle a bassa risoluzione. Il framework proposto dai creatori di HRNet, invece, mantiene rappresentazioni ad alta risoluzione durante l'intero processo con il risultato che la heatmap dei keypoint predetti è potenzialmente più accurata e più precisa nella localizzazione dei keypoint stessi.

Come altri framework, anche HRNet è basato sulle reti neurali convoluzionali profonde, le quali hanno dimostrato di poter raggiungere prestazioni all'avanguardia. Al contrario di come avviene nella maggior parte delle soluzioni esistenti, invece, HRNet collega le sottoreti ad alta e bassa risoluzione in parallelo (anziché in serie) garantendo così una superiorità in termini di precisione di rilevamento dei keypoint ed una maggiore efficienza in termini di complessità e parametri di calcolo.

In particolare, il framework applica delle operazioni di fusione ripetute in modo tale che ciascuna delle rappresentazioni ad alta e bassa risoluzione riceva informazioni da altre rappresentazioni parallele.

- Replicazione dell'esperimento

Nell'esperimento, vengono studiate due tipologie di rete: una più piccola (HRNet-W32) e una più grande (HRNet-W48). Inoltre, i test sono stati svolti su vari dataset tra cui MPII Human Pose Estimation. Il task preso in esame è la *single-person pose estimation* e la metrica utilizzata per le comparazioni è *PCKh@0.5*.

Per replicare l'esperimento è stato necessario modificare alcuni parametri del file di configurazione fornito dagli autori dello studio:

- ❖ GPUS: (0,1,2,3) → (0,)
- ❖ WORKERS: 24 → 2
- ❖ BATCH\_SIZE\_PER\_GPU: 32 → 1

Ciò è stato dovuto principalmente al fatto che il codice è stato sviluppato e testato in un ambiente con a disposizione 4 GPU NVIDIA P100. Per la replicazione dell'esperimento, invece, è stata utilizzata una macchina del laboratorio (tramite [virtualab](#)) che, avendo a disposizione una singola GPU, è in grado di riservare meno memoria di quella richiesta dalla configurazione originale.

## Risultati

- Risultati di HRNet sul validation set di MPII

Una volta scaricati i file dei modelli Pytorch e le annotazioni per le immagini di MPII, il framework HRNet è stato testato sul validation set del dataset MPII (il quale è formato da 2729 immagini), ottenendo i risultati presentati nella seconda riga della tabella sottostante.

<b>Arch</b>	<b>Hea</b>	<b>Sho</b>	<b>Elb</b>	<b>Wri</b>	<b>Hip</b>	<b>Kne</b>	<b>Ank</b>	<b>Mean @0.5</b>	<b>Mean @0.1</b>
<i>W32 (test)</i>	98.6	96.9	92.8	89.0	91.5	89.0	85.7	92.3	-
<i>W32 (valid.)</i>	97.1	95.9	90.3	86.4	89.1	87.1	83.3	90.3	37.7

I risultati del test sulla rete più grande (HRNet-W48) non vengono mostrati in quanto sovrapponibili a quelli del test sulla rete più piccola (HRNet-W32). Gli autori dello studio motivano questo fatto osservando che le prestazioni sul dataset MPII tendono a saturarsi. In entrambe le reti la dimensione dell'input è 256x256.

- Ulteriori analisi

Riprendendo il [codice](#) scritto per analizzare il comportamento del framework EfficientPose sul dataset MPII, è stata prodotta la seguente analisi sui valori delle metriche PCP e PDJ:

<b>Parte del corpo</b>	<b>PCP</b>	<b>PDJ</b>
('head_top', 'upper_neck')	98.3	99.0
('upper_neck', 'thorax')	49.8	99.7
('thorax', 'right_shoulder')	76.8	98.7
('thorax', 'left_shoulder')	76.5	98.9
('thorax', 'pelvis')	98.6	99.6
('right_shoulder', 'right_elbow')	89.2	97.2
('right_elbow', 'right_wrist')	81.0	96.3
('left_shoulder', 'left_elbow')	90.8	97.5
('left_elbow', 'left_wrist')	83.7	95.6
('pelvis', 'right_hip')	41.0	99.3
('pelvis', 'left_hip')	41.0	98.8
('right_hip', 'right_knee')	90.3	97.2
('right_knee', 'right_ankle')	85.5	93.6

('left_hip', 'left_knee')	90.0	95.6
('left_knee', 'left_ankle')	84.2	92.3
<b>Totale (@0.5)</b>	78.4	97.3
<b>AUC</b>	0.674	0.879

Dall'osservazione dei risultati si conferma che PDJ allevia lo svantaggio principale di PCP, ovvero la possibile correlazione negativa tra l'accuratezza del rilevamento delle parti del corpo e il valore della metrica. Nel calcolo di PDJ, infatti, i criteri di valutazione per tutte le parti del corpo si basano sulla stessa soglia di distanza.

Si può quindi constatare la precisione nella localizzazione dei keypoint che gli autori del modello affermano di poter garantire grazie all'uso di rappresentazioni ad alta risoluzione durante l'intero processo.

- Risultati del confronto con EfficientPose

Infine, viene presentata una comparazione con il modello EfficientPose, il quale prende in ingresso immagini sia ad alta che a bassa risoluzione (elaborate indipendentemente attraverso due backbone) per fornire due punti di vista separati; ed inoltre, sottopone ad up-scaling la heatmap contenente le predizioni per migliorare il livello di dettaglio dell'output e superare così una delle criticità di OpenPose (su cui è basato il modello).

Metrica	HRNet W32	EP-RT	EP-I	EP-II	EP-III	EP-IV
$PCKh$ @0.5 (valid.)	90.3	82.9	85.2	88.2	89.5	89.8
$PCKh$ @0.1 (valid.)	37.7	23.6	26.5	30.2	30.9	35.6
$PCKh$ @0.5 (test)	92.3	84.8	-	-	-	91.2

Il modello HRNet si comporta quindi meglio dei modelli presenti nel framework EfficientPose. Ciò è riscontrabile anche visitando il [link](#), in cui il modello HRNet viene dato in posizione #7 mentre il modello EfficientPose viene dato in posizione #17.

- Risultati di HRNet sul dataset COCO val2017

**PCK ~ tau=0.5**

<b>Parte del corpo</b>	<b>HRNet</b>	<b>EfficientPose IV (addestrato su MPII)</b>
right_ankle	94.4 %	56.1
right_knee	96.0 %	56.5
right_hip	97.0 %	57.0
left_hip	97.0 %	55.5
left_knee	96.5 %	53.5
left_ankle	95.5 %	56.6
pelvis	97.0 %	56.0
thorax	97.0 %	63.5
upper_neck	97.0 %	60.0
head_top	97.0 %	52.8
right_wrist	95.5 %	63.5
right_elbow	97.0 %	60.8
right_shoulder	97.0 %	62.5
left_shoulder	97.0 %	63.0
left_elbow	96.5 %	59.8
left_wrist	97.0 %	58.8

**PCP ~ tau=0.5**

<b>Coppia parti del corpo</b>	<b>HRNet</b>	<b>EfficientPose IV (addestrato su MPII)</b>
head_top, upper_neck	93.0 %	0.0 %
upper_neck, thorax	83.5 %	10.5 %
thorax, right_shoulder	88.0 %	47.5 %
thorax, left_shoulder	90.0 %	44.5 %
thorax, pelvis	96.0 %	53.5 %
right_shoulder, right_elbow	94.5 %	50.8 %

right_elbow, right_wrist	90.5 %	42.2 %
left_shoulder, left_elbow	92.0 %	50.3 %
left_elbow, left_wrist	85.9 %	41.9 %
pelvis, right_hip	55.0 %	21.0 %
pelvis, left_hip	61.5 %	19.5 %
right_hip, right_knee	93.0 %	43.0 %
right_knee, right_ankle	90.9 %	43.4 %
left_hip, left_knee	90.5 %	42.5 %
left_knee, left_ankle	92.9 %	45.5 %

**PDJ ~ tau=0.5**

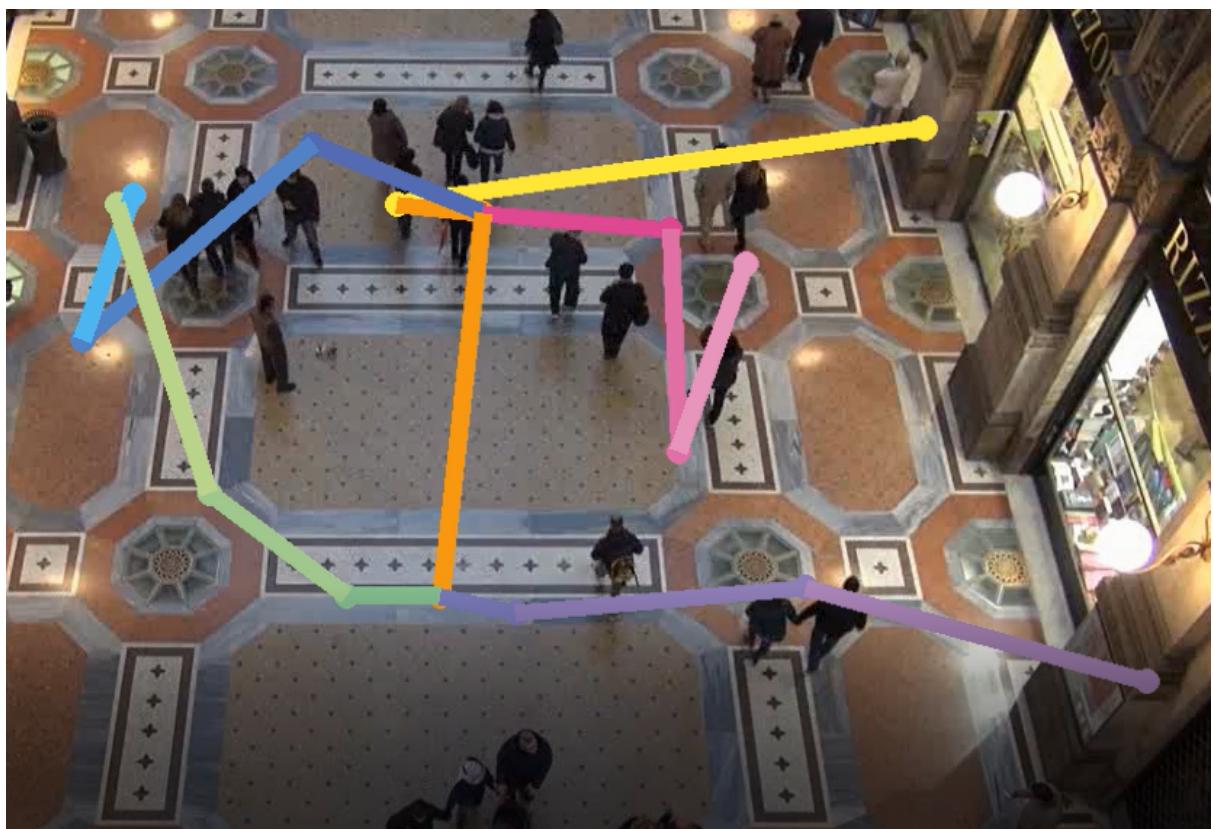
Coppia parti del corpo	HRNet	EfficientPose IV (addestrato su MPII)
head_top, upper_neck	97.0 %	44.7 %
upper_neck, thorax	97.0 %	58.5 %
thorax, right_shoulder	97.0 %	61.0 %
thorax, left_shoulder	97.0 %	59.0 %
thorax, pelvis	97.0 %	54.0 %
right_shoulder, right_elbow	97.0 %	55.5 %
right_elbow, right_wrist	95.5%	58.8 %
left_shoulder, left_elbow	96.5 %	57.8 %
left_elbow, left_wrist	96.5 %	56.3 %
pelvis, right_hip	97.0 %	54.0 %
pelvis, left_hip	97.0 %	55.5 %
right_hip, right_knee	96.0 %	50.5 %
right_knee, right_ankle	93.4 %	50.5 %
left_hip, left_knee	96.0 %	48.5 %
left_knee, left_ankle	94.9 %	48.0 %

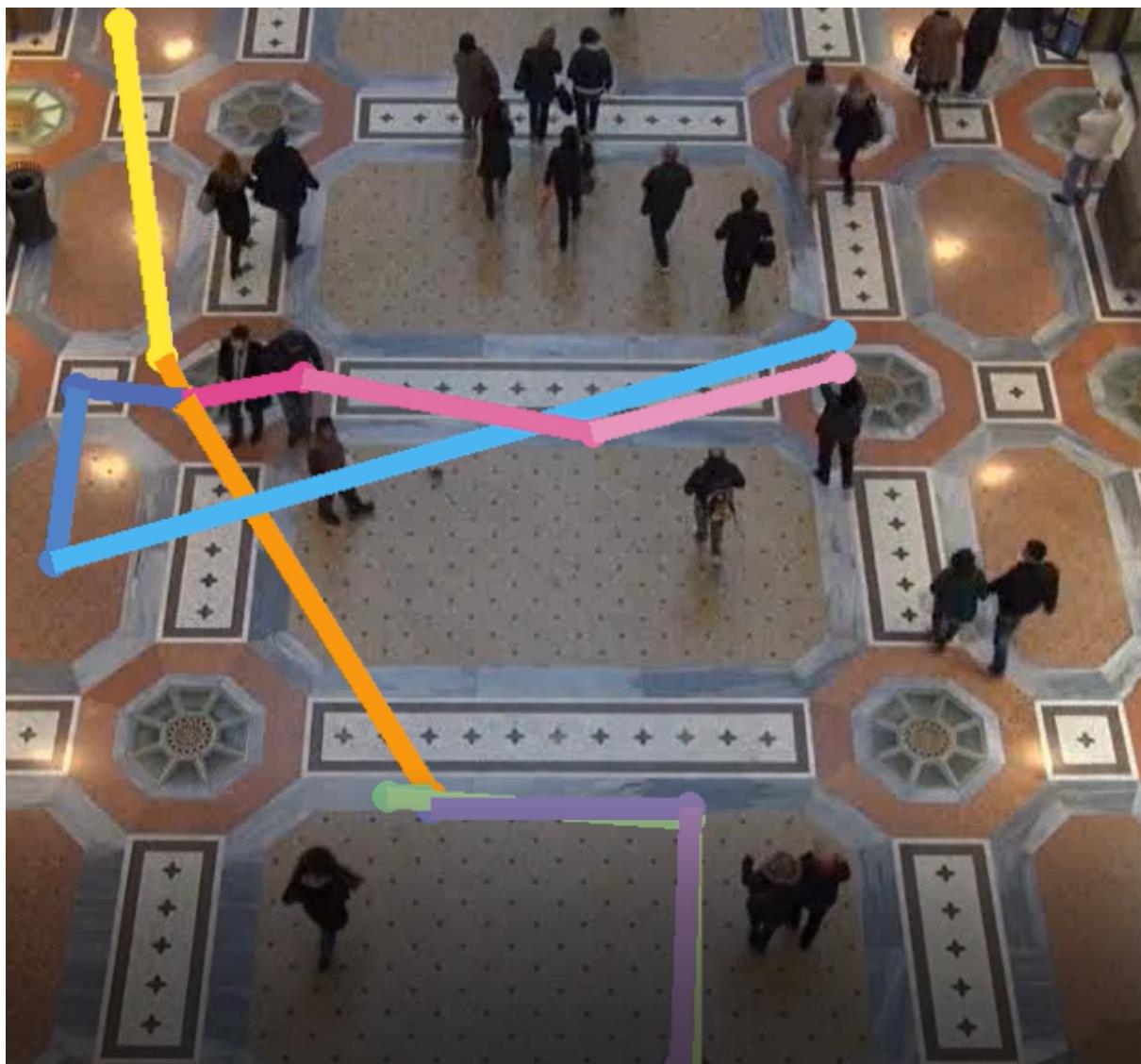
### AUC (più si avvicina a 1 e meglio è)

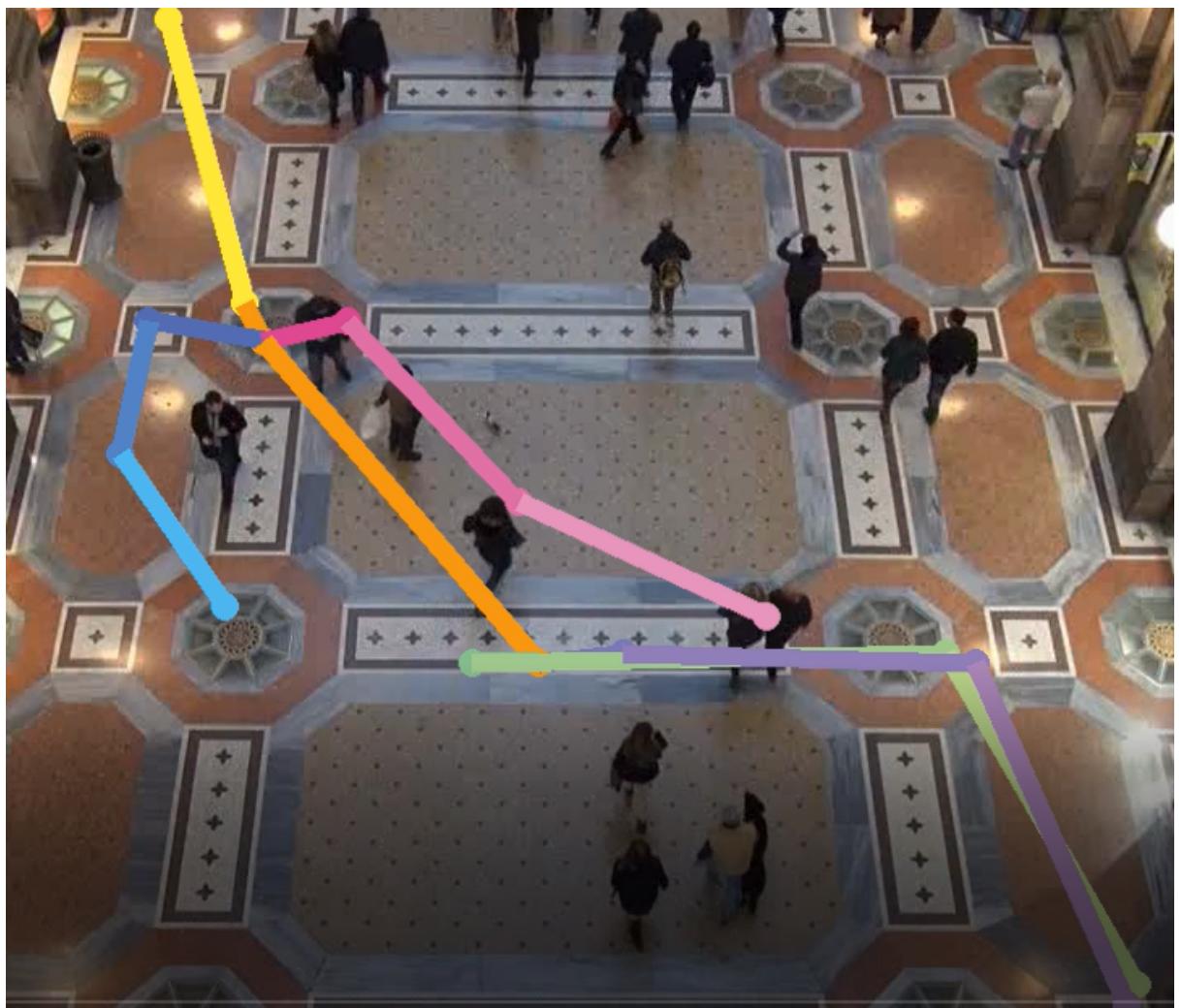
AUC per metrica	HRNet	EfficientPose IV (addestrato su MPII)
AUC per PCK	0.9117	0.5337
AUC per PCP	0.7424	0.3180
AUC per PDJ	0.8942	0.4846

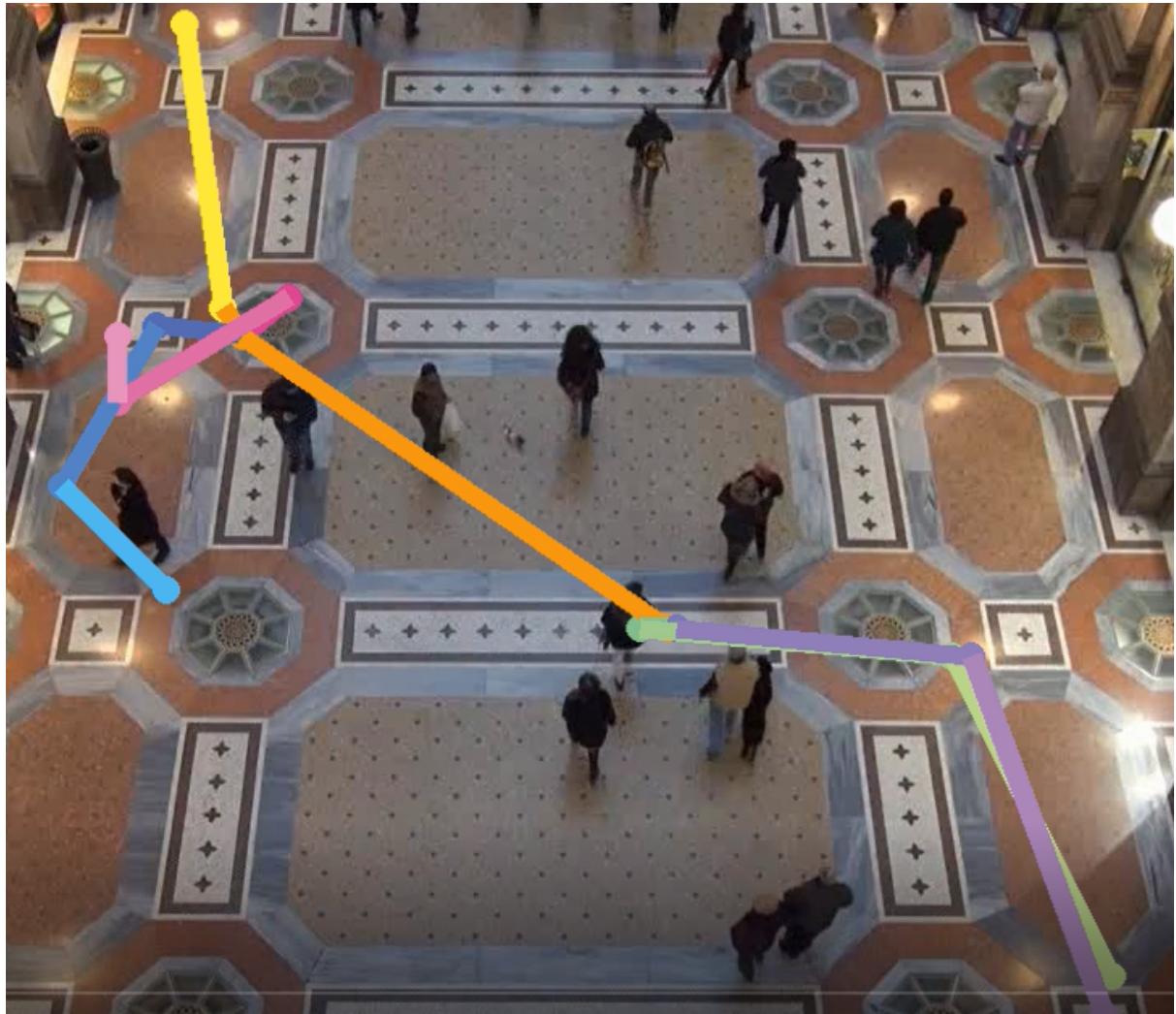
- Risultati test qualitativo su EfficientPose

C'è poco da dire, i risultati sono stati poco significativi. Trattandosi di un video contenente più persone in movimento EfficientPose RT non è riuscito per nessun frame del video a stimare la posa di nessuna persona. Di seguito viene proposto qualche frame









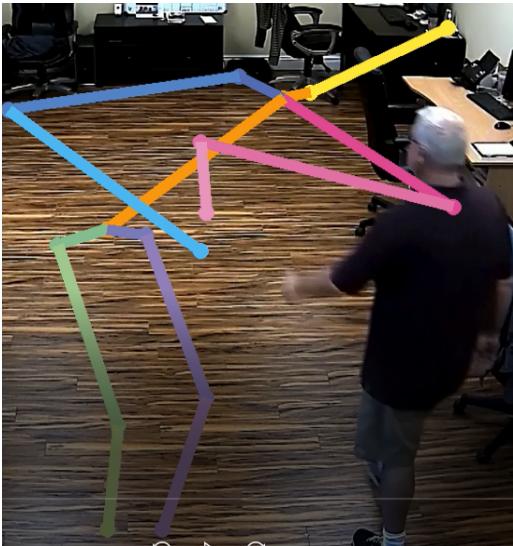
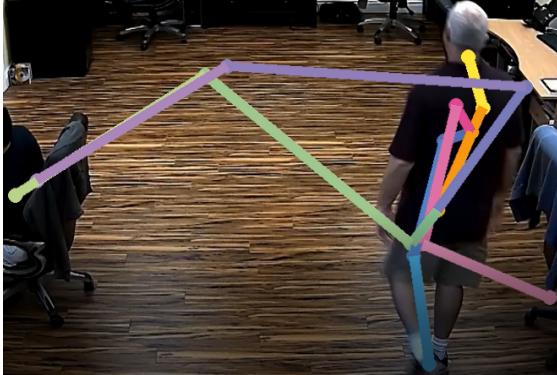
Viene considerato EfficientPose RT in quanto è il più efficiente fra tutti quanti e in un task di stima per video o per un rilevamento in tempo reale è fondamentale l'efficienza.

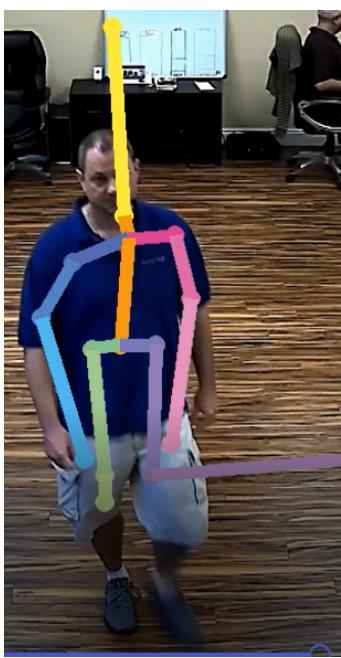
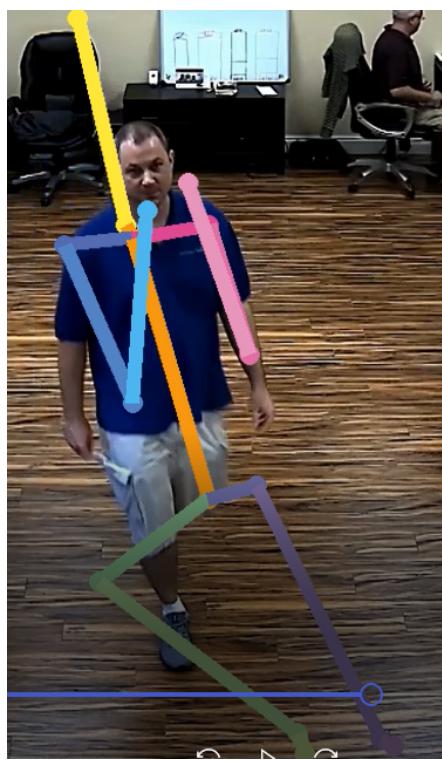
Risultati simili ce li si aspettava dato che il video non rispetta i criteri di ottimalità di EfficientPose:

- ❖ Persona in foreground
- ❖ Parti del corpo non devono essere coperte da altri oggetti nella scena
- ❖ Deve esserci una sola persona all'interno dell'immagine

Di seguito proponiamo una tabella che mette a confronto le pose stimate per i video test1.mp4 e test2.mp4 (recuperabili nella sezione ‘Dataset utilizzati’).

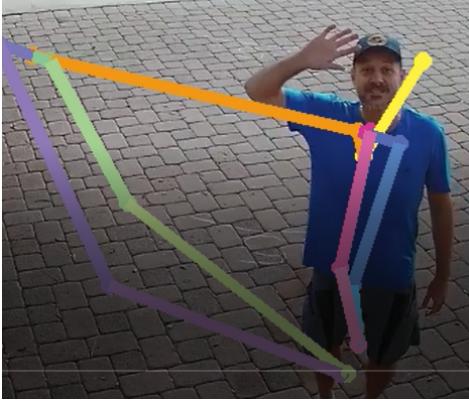
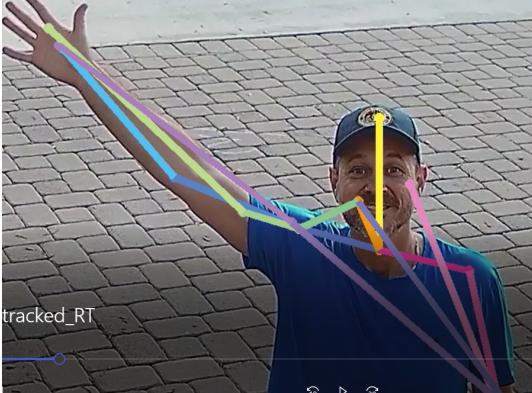
Video test1.mp4

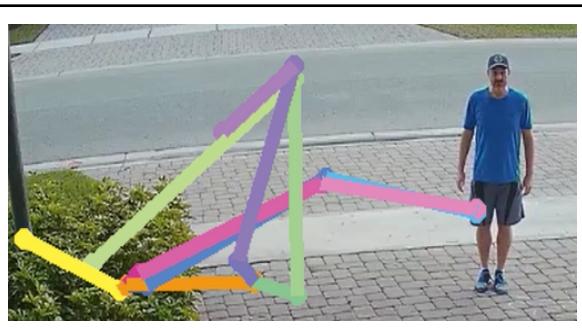
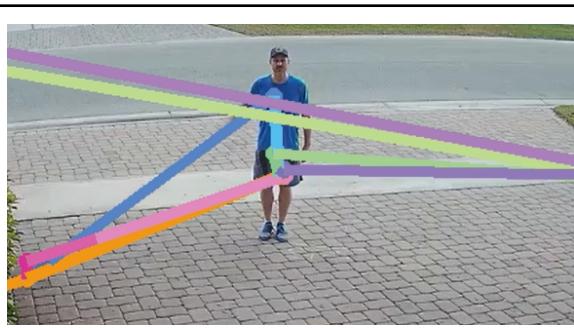
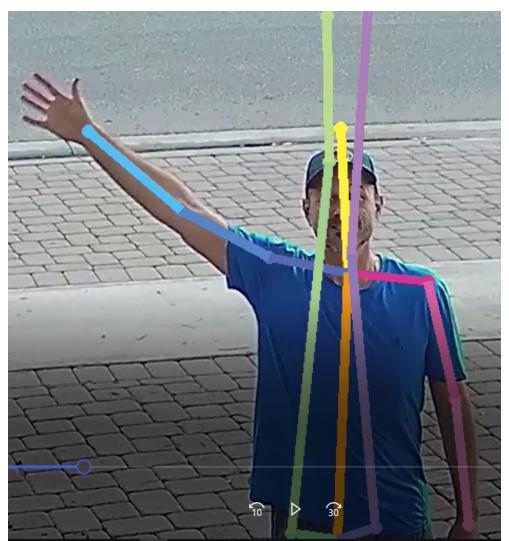
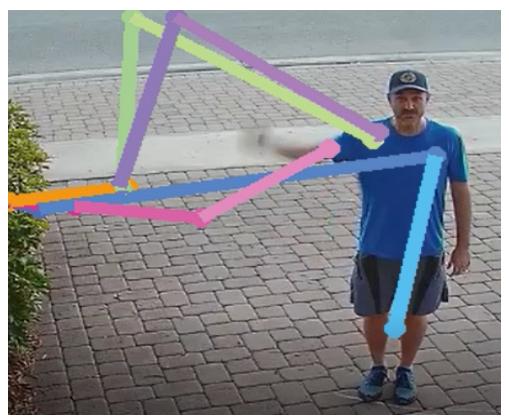
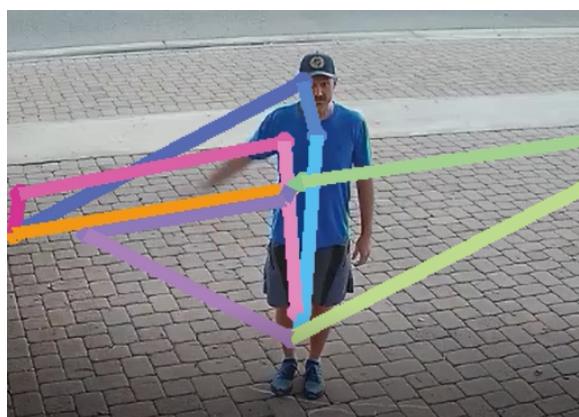
EFFICIENTPOSE RT	EFFICIENTPOSE II
	
	

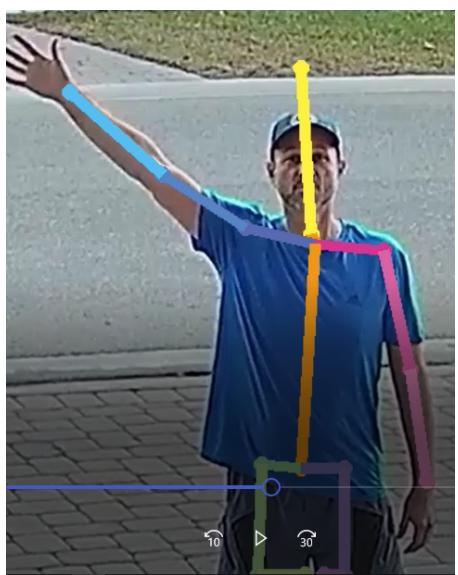
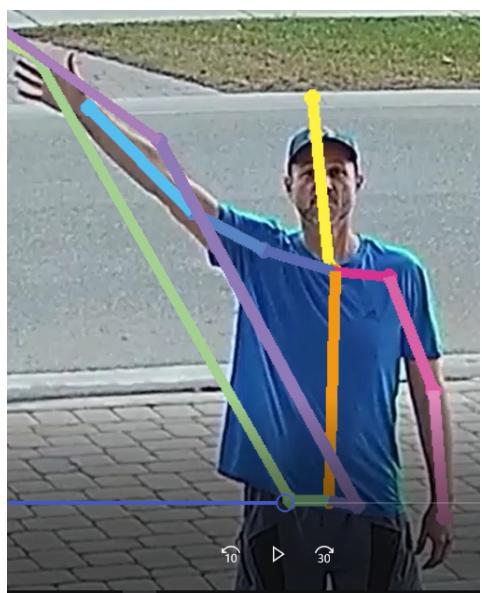




Video test2.mp4

EFFICIENTPOSE RT	EFFICIENTPOSE II
	
	
 <p>tracked_RT</p> <p>10 ▶ 30</p>	 <p>tracked_II</p> <p>10 ▶ 30</p>







## SECONDA PARTE

- Confronto tra EfficientPose applicato senza e dopo la detection (sul dataset MPII)

**N.B:** i valori in grassetto all'interno delle tabelle indicano il valore più alto di ogni riga.

PCKh ~ tau=0.5

Parte del corpo	EfficientPose RT (con la detection)	EfficientPose RT (senza detection)	EfficientPose IV (senza detection)
<i>right_ankle</i>	<b>59,0%</b>	30,6%	50,0%
<i>right_knee</i>	<b>64,2%</b>	34,1%	54,9%
<i>right_hip</i>	<b>54,8%</b>	42,9%	53,3%
<i>left_hip</i>	54,8%	37,1%	<b>55,2%</b>
<i>left_knee</i>	<b>65,4%</b>	31,7%	57,3%
<i>left_ankle</i>	<b>59,0%</b>	30,6%	43,5%
<i>pelvis</i>	<b>61,5%</b>	43,8%	57,1%
<i>thorax</i>	<b>74,3%</b>	58,2%	67,3%
<i>upper_neck</i>	<b>73,4%</b>	54,5%	60,0%

<i>head_top</i>	<b>76,9%</b>	47,7%	55,0%
<i>right_wrist</i>	53,7%	43,1%	<b>68,8%</b>
<i>right_elbow</i>	55,1%	44,4%	<b>62,0%</b>
<i>right_shoulder</i>	<b>67,0%</b>	52,7%	62,7%
<i>left_shoulder</i>	<b>66,1%</b>	57,3%	61,8%
<i>left_elbow</i>	59,3%	46,8%	<b>61,5%</b>
<i>left_wrist</i>	55,1%	50,9%	<b>63,0%</b>

PCP ~ tau=0.5

Coppia parti del corpo	EfficientPose RT (con la detection)	EfficientPose RT (senza detection)	EfficientPose IV (senza detection)
<i>head_top, upper_neck</i>	<b>73,1%</b>	48,6%	55,0%
<i>upper_neck, thorax</i>	<b>16,5%</b>	9,1%	8,2%
<i>thorax, right_shoulder</i>	<b>52,3%</b>	42,7%	51,8%
<i>thorax, left_shoulder</i>	<b>56,9%</b>	42,7%	51,8%
<i>thorax, pelvis</i>	<b>76,0%</b>	53,3%	67,6%
<i>right_shoulder, right_elbow</i>	57,9%	42,6%	<b>60,2%</b>
<i>right_elbow, right_wrist</i>	42,1%	31,5%	<b>57,4%</b>
<i>left_shoulder, left_elbow</i>	57,4%	46,8%	<b>58,7%</b>
<i>left_elbow, left_wrist</i>	44,9%	45,4%	<b>56,5%</b>
<i>pelvis, right_hip</i>	<b>30,8%</b>	20,0%	22,9%
<i>pelvis, left_hip</i>	24,0%	20,0%	<b>28,6%</b>
<i>right_hip, right_knee</i>	<b>65,4%</b>	35,4%	53,7%
<i>right_knee, right_ankle</i>	<b>63,9%</b>	29,0%	50,0%
<i>left_hip, left_knee</i>	<b>66,7%</b>	35,4%	54,9%

<i>left_knee, left_ankle</i>	<b>57,4%</b>	25,8%	46,8%
----------------------------------	--------------	-------	-------

PDJ ~ tau=0.5

Coppia parti del corpo	EfficientPose RT (con la detection)	EfficientPose RT (senza detection)	EfficientPose IV (senza detection)
<i>head_top, upper_neck</i>	<b>86,4%</b>	65,4%	65,4%
<i>upper_neck, thorax</i>	<b>92,2%</b>	70,2%	72,1%
<i>thorax, right_shoulder</i>	<b>81,6%</b>	64,4%	70,2%
<i>thorax, left_shoulder</i>	<b>85,4%</b>	69,2%	71,2%
<i>thorax, pelvis</i>	<b>80,6%</b>	58,7%	72,1%
<i>right_shoulder, right_elbow</i>	<b>71,6%</b>	57,3%	68,0%
<i>right_elbow, right_wrist</i>	67,6%	50,5%	<b>70,9%</b>
<i>left_shoulder, left_elbow</i>	<b>74,5%</b>	57,3%	66,0%
<i>left_elbow, left_wrist</i>	<b>68,3%</b>	53,9%	66,7%
<i>pelvis, right_hip</i>	<b>80,2%</b>	60,8%	76,5%
<i>pelvis, left_hip</i>	<b>81,2%</b>	56,9%	76,5%
<i>right_hip, right_knee</i>	<b>73,4%</b>	42,5%	62,5%
<i>right_knee, right_ankle</i>	<b>76,7%</b>	32,8%	57,4%
<i>left_hip, left_knee</i>	<b>80,0%</b>	37,7%	59,0%
<i>left_knee, left_ankle</i>	<b>72,4%</b>	40,7%	55,9%

Considerando il modello RT, in quanto è il più efficiente fra tutti, applicato dopo aver effettuato una detection con yolov5 (viene ritagliata l'immagine attorno alla persona individuata con il maggior grado di confidenza) si può notare come esso raggiunga performance sovrapponibili a quelle ottenute dal modello IV, il più preciso tra tutti, applicato

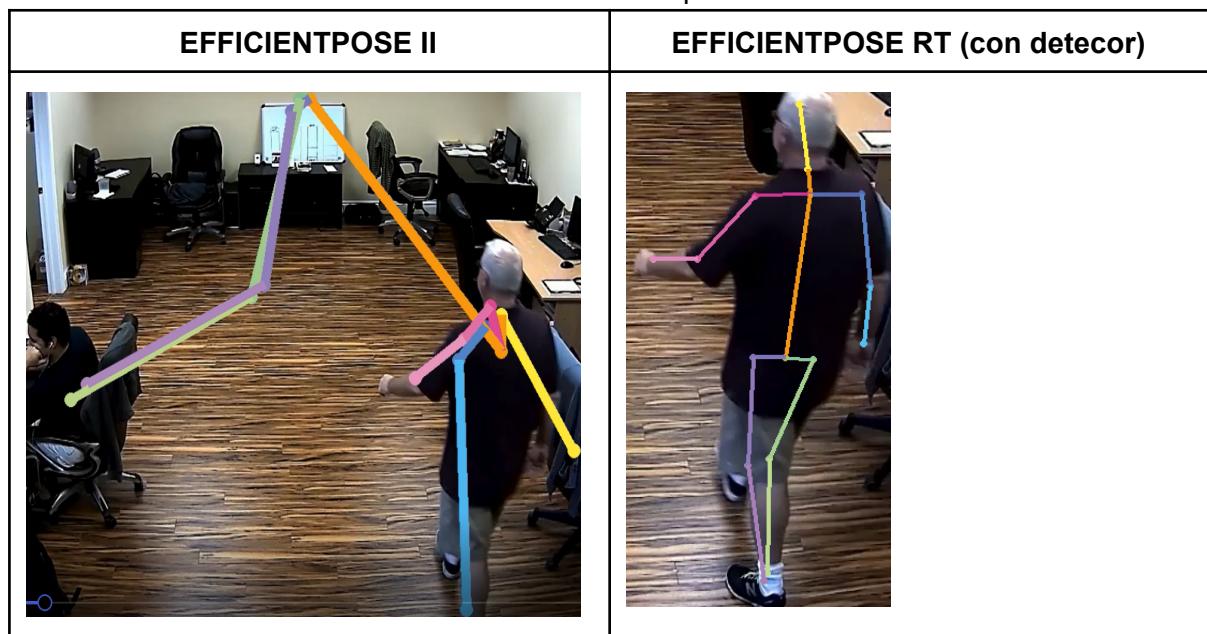
sulle immagini originali; inoltre, riesce in ogni caso a migliorare in maniera considerevole i risultati che si avrebbero senza effettuare la detection.

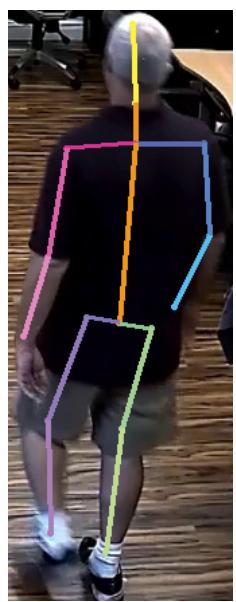
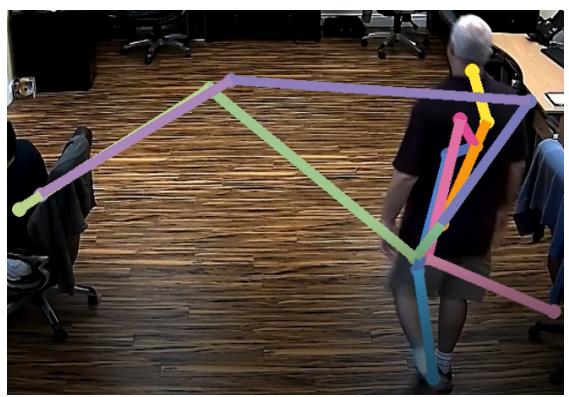
Tempi di inferenza su 110 immagini del train set del dataset MPII (considerare circa 250ms aggiuntivi per ogni immagine nel caso in cui viene effettuata la detection):

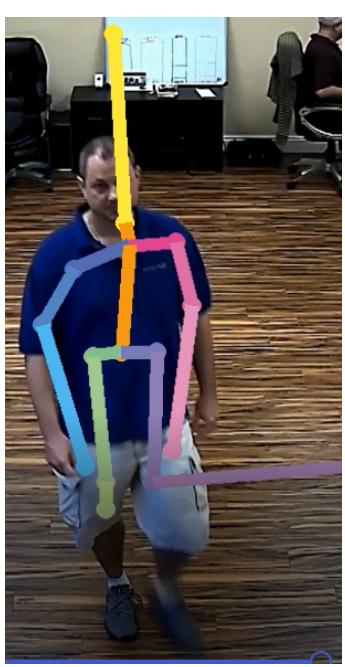
Modello	<i>RT</i>	<i>I</i>	<i>II</i>	<i>III</i>	<i>IV</i>
<b>Tempo di inferenza (assoluto)</b>	11 minuti	12 minuti	13 minuti	17 minuti	29 minuti
<b>Tempo di inferenza (relativo)</b>	1x	1,1x	1,2x	1,5x	2,6x

- Test qualitativo - Confronto fra RT (con detector) ed EP II

Video test1.mp4

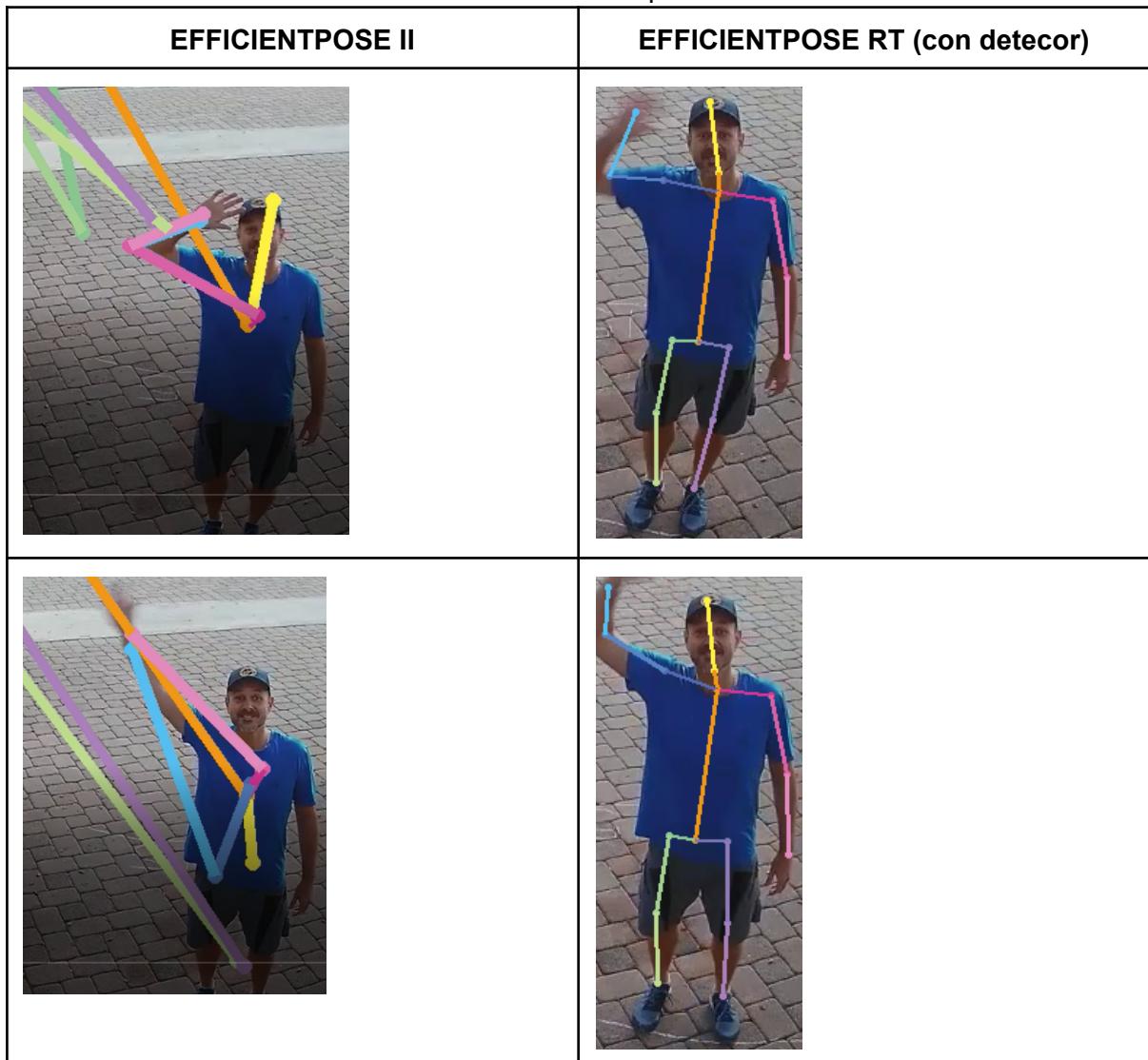


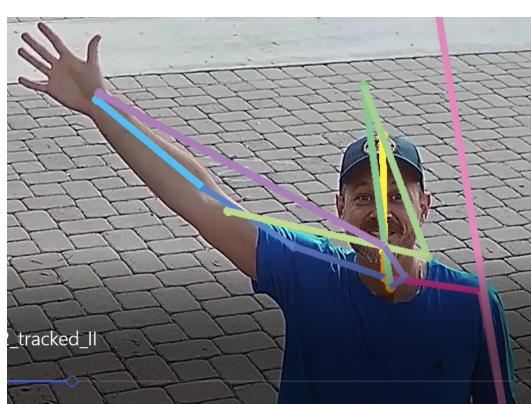


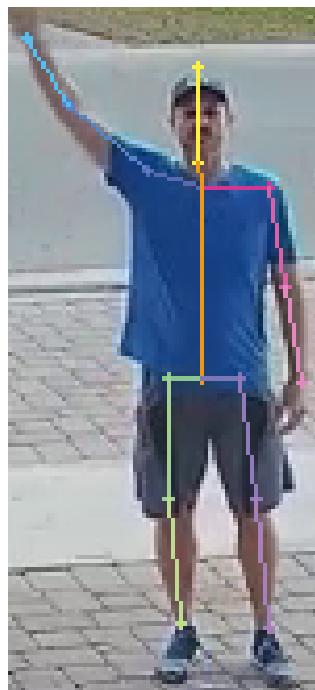
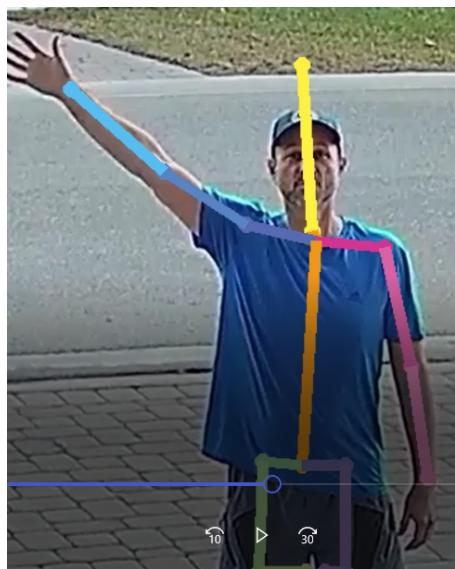
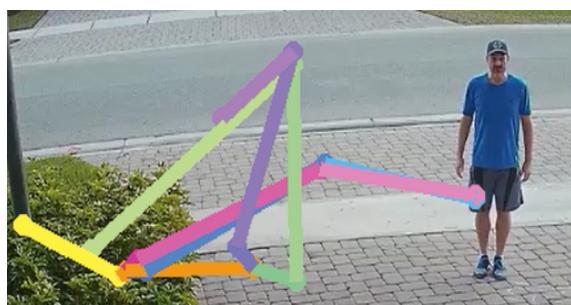


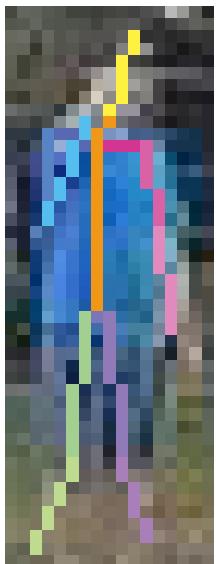


Video test2.mp4









Dal test qualitativo è evidente come l'impiego del detector abbia portato a risultati migliori (anche rispetto ad un modello EP superiore e quindi tecnicamente più preciso) proprio perché viene eseguito un crop (ritaglio) intorno alla persona; in questo modo vengono rispettati due requisiti importanti affinché EfficientPose generi risultati qualitativamente buoni.

- Integrazione di Yolo in EfficientPose per immagini multi-person

L'integrazione del detector Yolo in EfficientPose per fare stima della posa su immagini in cui sono presenti più persone è un'operazione che va inevitabilmente a pesare sulla velocità effettiva dell'algoritmo.

Per vedere di quanto è stato realizzato il seguente confronto tra le due versioni dell'algoritmo sul dataset COCO. L'impiego di COCO è giustificato dalla comoda interfaccia che consente di scegliere le immagini che soddisfano determinati criteri (ad esempio la presenza delle annotazioni per un certo numero di persone).

**EP: 90 immagini**

Modello	RT	I	II	III	IV
Tempo medio di inferenza (assoluto)	0.3547 s	0.5187 s	1.4735 s	3.2978 s	44.065 s
Tempo medio di inferenza (relativo)	1x	1.5x	4.2x	9.3x	124.2x

**YOLO+EP: 30 immagini con 1 persona annotata**

Modello	RT	I	II	III	IV
Tempo medio di inferenza (assoluto)	0.9377 s	1.1963 s	2.8565 s	6.4667 s	94.831 s
Tempo medio di inferenza (relativo)	1x	1.3x	3x	6.9x	101.1x

### YOLO+EP: 30 immagini con 2 persone annotate

Modello	<i>RT</i>	<i>I</i>	<i>II</i>	<i>III</i>	<i>IV</i>
Tempo medio di inferenza (assoluto)	1.4673 s	2.1595 s	6.457 s	15.53 s	210.95 s
Tempo medio di inferenza (relativo)	1x	1.5x	4.4x	10.6x	143.8x

### YOLO+EP: 30 immagini con 3 persone annotate

Modello	<i>RT</i>	<i>I</i>	<i>II</i>	<i>III</i>	<i>IV</i>
Tempo medio di inferenza (assoluto)	1.9269 s	2.8225 s	8.1688 s	19.394 s	214.65 s
Tempo medio di inferenza (relativo)	1x	1.5x	4.2x	10.1x	111.4x

### Confronto tra EP e YOLO+EP

Modello	EP base (single-person)	YOLO+EP (1 persona)	YOLO+EP (2 persone)	YOLO+EP (3 persone)
<i>RT</i>	1x	2.6x	4.1x	5.4x
<i>I</i>	1x	2.3x	4.2x	5.4x
<i>II</i>	1x	1.9x	4.4x	5.5x
<i>III</i>	1x	2x	4.7x	5.9x
<i>IV</i>	1x	2.2x	4.8x	4.9x

# RIFERIMENTI

## Idee

- Abbiamo deciso di considerare altri due video (test1.mp4 e test2.mp4 recuperabili nella sezione ‘Dataset utilizzati’) per studiare fino a che punto è possibile spingersi nel rilevamento della posa con EfficientPose. Il video test1 presenta un gruppo di persone di cui due (poi diventeranno tre) quasi immobili e un’altra che si muove nella scena. Il video test2 invece ha solo una persona nella scena che lentamente si allontana all’interno della scena.

## Letteratura

- [\[1902.09212\] Deep High-Resolution Representation Learning for Human Pose Estimation \(arxiv.org\)](#)
- [EfficientPose: Scalable single-person pose estimation | SpringerLink](#)

## Dataset utilizzati

- [MPII Human Pose Database \(mpg.de\)](#)
- [COCO - Common Objects in Context \(cocodataset.org\)](#)
- Il video gveii.mp4 (per il test qualitativo): [datasets - OneDrive \(sharepoint.com\)](#)
- Il video test1.mp4 (per il test qualitativo):  
<https://www.youtube.com/watch?v=9wxEmqyVIB8>
- Il video test2.mp4 (per il test qualitativo):  
<https://www.youtube.com/watch?v=cqlIt4OJVMq>