Deep Learning
Single-Person Efficient 2D Human Pose Estimation

Michele Dalla Chiara - VR464051
Davide Zampieri - VR458470

A.A. 2021 - 2022

1 Introduzione

La stima della posa umana (Human Pose Estimation) è un noto task di computer vision il quale si preoccupa di produrre un insieme di coordinate che collegate determinano la posa di una persona. Una coordinata dello scheletro è nota come parte (joint/keypoint). Una connessione valida tra due parti è detta coppia (limb). Chiaramente non tutte le combinazioni delle parti danno origine a coppie valide (due keypoint come testa, anca sinistra non sono una coppia valida).

1.1 Approcci basati sul deep learning

Nell'ambito della computer vision, gli approcci basati sul deep learning, e in particolare le reti neurali convoluzionali profonde, (Convolutional Neural Networks) superano tutti gli altri paradigmi, e questo vale anche per la stima della posa umana.

Infatti, le CNN riescono a riconoscere i pattern all'interno di un'immagine di input con maggiore precisione e accuratezza rispetto a qualsiasi altro modello/algoritmo, rendendo di fatto le CNN le più adatte per attività di classificazione e rilevamento.

Per usare le CNN nella stima della posa umana, basta definire l'intero problema come un problema di regressione sulle parti del corpo (joint).

1.2 Framework EfficientPose

EfficientPose sfrutta delle CNN computazionalmente efficienti nel riconoscimento delle immagini, note come EfficientNet, per costruire un'architettura di rete scalabile in grado di eseguire la stima della posa (single-person) che risponda a differenti requisiti computazionali. Il framework comprende infatti una famiglia di cinque ConvNet che formano i modelli: EfficientPose RT, I, II, III, IV. Di seguito viene presentata l'architettura generale di EfficientPose.

1. In ingresso prende immagini sia ad alta che a bassa risoluzione per fornire due punti di vista separati.

2. Le immagini di input vengono elaborate indipendentemente attraverso due EfficientNet, dette backbone di alto e basso livello.

3. Le features risultanti sono **concatenate** per produrre le cosiddette cross-resolution features, le quali danno la possibilità di dare selettivamente enfasi sull'informazione globale e locale di un'immagine (si da egual peso alle feature delle backbone di alto e basso livello).

4. La fase di detection utilizza un blocco Mobile-DenseNet scalabile per eseguire il rilevamento in tre passaggi, dove il primo passaggio stima gli scheletri delle persone attraverso i Part Affinity Fields (per produrre configurazioni di posa fattibili) mentre il secondo e il terzo passaggio stimano le posizioni dei keypoint (con un progressivo miglioramento della precisione) producendo una keypoint heatmap per ogni parte del corpo (dalla heatmap utilizzando $\text{argmax}_p \, heatmap(p)$ con p pixel si determina la posizione del keypoint).

5. La previsione a bassa risoluzione, ottenuta dal terzo passaggio sopra citato, viene sottoposta ad up-scaling attraverso l'interpolazione bilineare migliorando il livello di dettaglio dell'output (superando una delle criticità di OpenPose).

1.2.1 Differenze con OpenPose

OpenPose è un framework open source che ha permesso al pubblico la stima della posa umana senza dover per forza usare software a pagamento o closed source. Esso comprende un'architettura multi-stage che esegue una serie di passaggi:

1. Fornita un'immagine di input, utilizza come backbone la rete neurale VGG-19 (preaddestrata su ImageNet) per estrarne le features.

2. Le features estratte vengono elaborate attraverso sei fasi di detection sequenziali, ognuna organizzata in cinque blocchi (DenseNet).

3. Le prime quattro fasi stimano gli scheletri delle persone attraverso i Part Affinity Fields (per mappare le associazioni tra i keypoint), mentre le ultime due fasi producono una keypoint heatmap per ogni parte del corpo (per ottenere stime precise delle coordinate dei keypoint).

1.3 Osservazioni riscontrate durante le prove

Durante le prove dei modelli abbiamo osservato che con EfficientPose β con $\beta >= 2$ se la persona, **ripresa in tempo reale**, non viene mostrata nella sua interezza, ad esempio dal busto in su, alcuni keypoint vengono posizionati nel pixel in alto a destra. Tale fenomeno

siamo riusciti a riscontrarlo anche quando la persona assumeva posizioni particolari. Ispezionando la funzione *extract_coordinates* (all'interno del codice di EfficientPose) si riesce ad individuare la seguente sezione:

```
# confidence inizializzato a 0.3
if real_time and conf[int(peak_y),int(peak_x)] < confidence:
    peak_x = -0.5
    peak_y = -0.5
else:
    peak_x += 0.5
    peak_y += 0.5
```

Nel codice troviamo:

- *real_time*: variabile booleana che stabilisce se l'immagine proviene da un frame di una cattura in tempo reale o meno.
- $conf$: è una delle 16 keypoint heatmap per determinare la locazione di un dato keypoint. Si tratta di un'immagine in cui il valore dei pixel indica la probabilità che quel pixel sia la posizione del keypoint.
- $confidence$: è una soglia di confidenza settata a 0.3 dentro la funzione.
- $peak\_x$ e $peak\_y$: coordinate del punto che dentro l'immagine ha il valore più alto. Da $conf$, $peak\_x$ e $peak\_y$ vengono trovati usando la $argmax$.

Questa porzione di codice stabilisce quindi che se il valore di $conf[int(peak\_y), int(peak\_x)]$ è minore di $confidence$ allora $peak\_x$ e $peak\_y$ vengono impostati a -0.5, valore corrispondente al pixel in alto a destra.

1.4 Metriche utilizzate

Nell'ambito della HPE sono state definite delle metriche per stabilire la bontà dei risultati, alcune delle quali sono: PCP, PDJ e PCK.

1.4.1 PCP

PCP (Percentage of Correct Parts) è il criterio di valutazione più popolare per la stima della posa. Nonostante questo criterio di valutazione venga considerato molto influente, esso è stato specificato in maniera ambigua, con il risultato che alcune sue implementazioni possono essere contrastanti tra loro.

Potrebbe quindi esistere una correlazione negativa tra l'accuratezza del rilevamento delle parti del corpo e la metrica PCP.

Premesso ciò, possiamo dire che la metrica PCP viene utilizzata per misurare il corretto rilevamento degli arti: se la distanza di ognuno dei due keypoint dell'arto predetti da quelli veri è al massimo la metà (o una frazione) della lunghezza dell'arto, allora l'arto viene considerato rilevato.

1.4.2 PDJ

Per risolvere i problemi di PCP, è stata proposta la metrica PDJ (Percentage of Detected Joints). Essa rappresenta la percentuale di articolazioni rilevate e si ottiene misurando la distanza di ognuno dei due keypoint di un arto predetti da quelli veri, la quale deve essere all'interno di una certa frazione del diametro del busto.

PDJ allevia lo svantaggio principale di PCP poiché i criteri di valutazione per tutte le articolazioni si basano sulla stessa soglia di distanza.

1.4.3 PCK

PCK è la metrica che misura se il keypoint predetto e quello vero si trovano entro una certa soglia di distanza. La soglia utilizzata da PCK può essere impostata rispetto alla scala del soggetto (diametro del busto) oppure anche rispetto alla diagonale del riquadro di delimitazione della testa.

Ad esempio, la metrica $PCK_h@\tau$ viene definita come la percentuale di keypoint predetti che risiedono entro una certa distanza τl da quelli veri, dove l è pari al 60% della diagonale d del riquadro di delimitazione della testa e τ è l'errore percentuale accettato rispetto a l. Ancora, $PCK@0.2$ indicherà come correttamente rilevato un keypoint la cui distanza da quello vero è al massimo pari al 20% (o una frazione) del diametro del busto.

Anche PCK allevia il problema degli arti di diverse dimensioni poiché arti più corti corrispondono a torsi (o teste) di minori dimensioni.

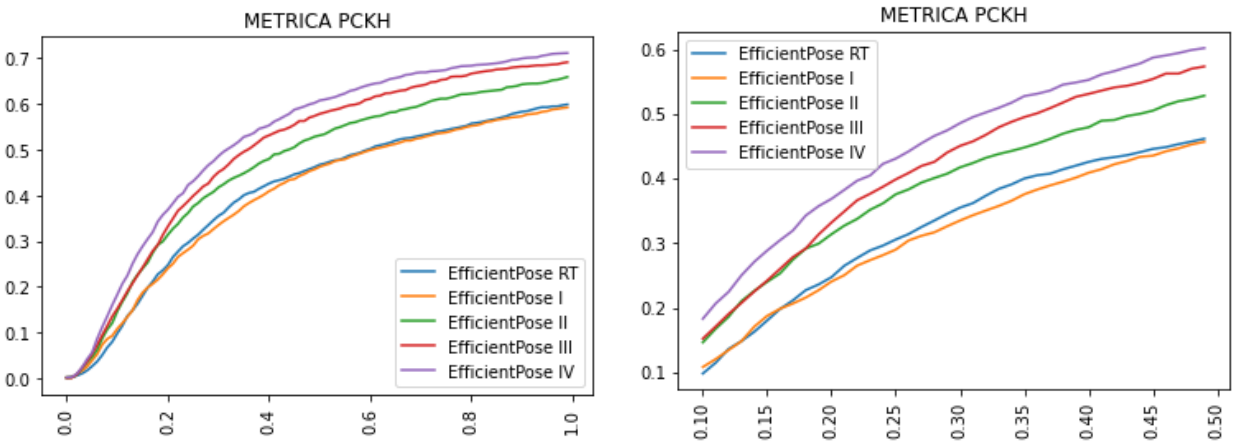
2 Replicazione risultati dell'algoritmo EfficientPose

Nella tabella 3 del paper di EfficientPose vengono mostrati i risultati dei vari modelli sul dataset MPII e viene fatta una comparazione con OpenPose in base all'efficienza e all'accuratezza. Si riportano quindi di seguito i risultati per le metriche PCK_h@0.5 e PCK_h@0.1 ottenuti sul validation set del dataset MPII per confrontarli con i risultati ottenuti dalla nostra replicazione dell'esperimento.

Nota: per la nostra analisi sono state utilizzate le immagini del train set in quanto le annotazioni per quelle del test set sono trattenute dai creatori del dataset.

	RT	**I**	**II**	**III**	**IV**
PCK_h@0.5 *(validation set)*	82.88	85.18	88.18	89.51	89.75
PCK_h@0.5 *(train set)*	46.70	46.23	53.11	57.84	60.76
PCK_h@0.1 *(validation set)*	23.56	26.49	30.17	30.90	35.63
PCK_h@0.1 *(train set)*	9.79	10.79	14.61	15.19	18.27

Figura 1: Comparazione dei valori per la metrica PCK_h (dataset MPII) e zoom



Nella tabella 4 del paper di EfficientPose vengono invece mostrati i risultati dello stato dell'arte per la metrica PCK_h@0.5 (sia per le singole parti del corpo che per il valore medio complessivo) ottenuti sul test set del dataset MPII. Di seguito, si confrontano i risultati relativi ai modelli EfficientPose RT e IV con i risultati ottenuti dalla nostra replicazione dell'esperimento.

Nota: anche in questo caso per la nostra analisi sono state utilizzate le immagini del train set.

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
RT (test set)	97.0	93.3	85.0	79.2	85.9	77.0	71.0	84.8
RT (train set)	47.7	55.0	45.6	47.0	40.0	32.9	30.6	42.7
IV (test set)	98.2	96.0	91.7	87.9	90.3	87.5	83.9	91.2
IV (train set)	55.0	62.3	61.8	65.9	54.3	56.1	46.8	57.5

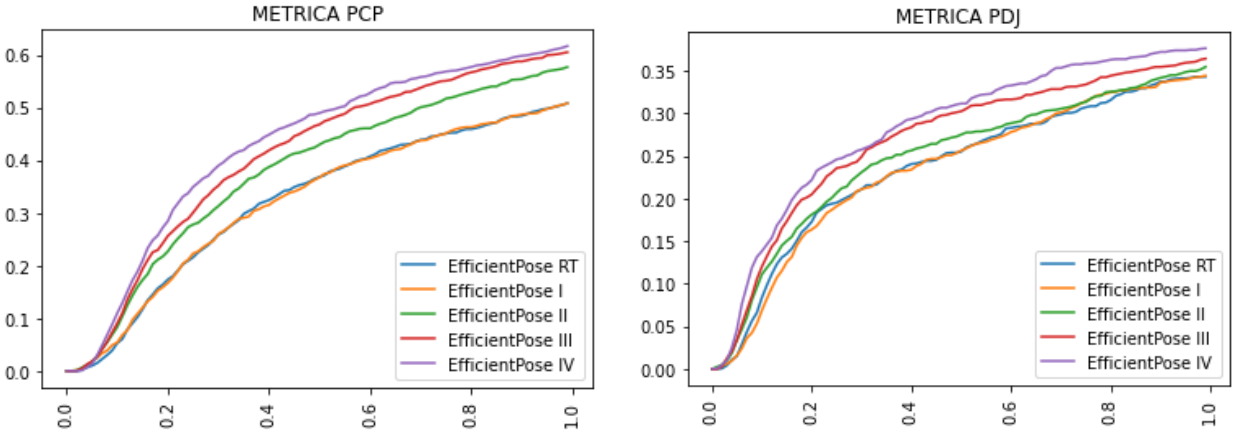
Da un confronto relativo si può notare come le parti del corpo predette meglio dal modello RT siano testa e spalle mentre quelle predette peggio siano le caviglie; quest'ultima considerazione vale anche per il modello IV.

2.1 Ulteriori analisi sul dataset MPII

Di seguito la tabella contenente i valori per la metrica AUC e i grafici con la comparazione dei valori per le metriche PCP e PDJ tra i vari modelli di EfficientPose.

	AUC per PCKH	AUC per PCP	AUC per PDJ
RT	0.405	0.320	0.235
I	0.398	0.320	0.233
II	0.463	0.376	0.247
III	0.496	0.407	0.268
IV	0.523	0.427	0.282

Figura 2: Comparazione dei valori per le metriche PCP e PDJ (dataset MPII)



3 Dataset COCO

Il dataset COCO, avente all'incirca 250k immagini (di non sole persone), nacque per la object detection e la segmetation e solo successivamente viene espanso alla multiperson pose estimation. I creatori del dataset proposero una nuova metrica per la valutazione del task HPE, nota come OKS, calcolata come:

$$OKS = \frac{\sum_i \exp \frac{-d_i^2}{2*s^2*k_i^2} * \delta(v_i > 0)}{\sum_i \delta(v_i > 0)}$$

sapendo che:

- i itera sui punti chiave rilevati;
- d_i è la distanza euclidea tra il keypoint predetto e il ground truth;
- s^2 è l'area del bounding box
- k_i è una costante che serve ad omogeneizzare la dev. standard rispetto alle diverse parti del corpo;
- v_i indica la visibilità del keypoint nell'immagine; in particolare $v_i = 0$ quando il keypoint i-esimo non è presente nelle annotations a significare che i falsi positivi non verranno conteggiati.

OKS è tuttavia da confinare all'ambito della multiperson HPE. Per questo motivo e per poter comparare i risultati con gli altri due dataset ancora le metriche impiegate sono : PCP, PDJ e PCK.

3.1 EfficientPose per COCO

Il file annotations è in formato json ed è molto articolato. I creatori del dataset hanno fornito un interfaccia molto elaborata per trattare con questa tipologia di file. Dato che non tutto il pacchetto era necessario, ciò che è stato fatto è stato prendere e riadattare il file **coco.py** per poter sfruttare alcune funzioni importanti come la possibilità di scegliere solo immagini di persone e lo scaricamento delle immagini dai server di COCO.

I passaggi chiave sono:

- creazione di un oggetto di classe COCO per sfruttare i metodi della classe
- uso dell'oggetto coco per recuperare l'id delle immagini contenenti delle persone al loro interno
- usando l'id delle immagini, vengono caricate le annotazioni (ground truth) delle immagini

- attraverso le annotazioni vengono individuate solo le immagini che verificano vincoli specifici sui keypoints poiché alcuni keypoint sono necessari in fasi successive
- una volta aver individuato dal dataset le immagini di interesse queste vengono scaricate e salvate in una cartella specifica (se richiesto)
- le immagini vengono poi analizzate da uno o da più modelli di EfficientPose e vengono poi fatte le comparative dei risultati

Il dataset COCO presenta delle differenze rispetto ai keypoint deducibili dai modelli di EfficientPose. Di seguito vengono riportate le differenze sapendo che (+) indica i keypoint aggiuntivi di COCO mentre (-) indica i keypoint mancanti.

- (+) nose
- (+) left_eye
- (+) right_eye
- (+) left_ear
- (+) right_ear
- (-) pelvis
- (-) thorax
- (-) head_top
- (-) upper_neck

I keypoint mancanti vengono calcolati partendo dai keypoint disponibili. Il calcolo di alcuni di questi si basa sugli studi anatomici svolti da Avard Tennyson Fairbanks riguardanti le proporzioni umane. Il bacino viene calcolato come punto medio tra le due anche, il torace come punto medio tra le due due spalle, l'head_top viene calcolato considerando che la distanza tra il naso e il punto più alto della testa è circa uguale alla distanza tra i due occhi, infine upper_neck lo si pensa come il punto medio tra torace e il naso.

3.2 Presentazione dei risultati

All'interno del main è presente una variabile globale che controlla se si desidera avere nell'immagine una sola persona o no. Il dataset originario, in conseguenza al valore della variabile, potrà contenere immagini differenti ma sempre in egual numero (200 immagini). I risultati che verranno proposti considereranno sia il caso persona singola sia il caso multi persona, così da poter studiare il comportamento di EfficientPose quando sono presenti più elementi all'interno dell'immagine.

Nelle tabelle sono presenti due valori per ogni modello. Il primo valore rappresenta il True Positive per le immagini single person mentre il secondo è il True Positive per le immagini multi person. Il termine True Positive nell'ambito dell'HPE si rifà ai keypoint che si trovano entro una certa distanza dal keypoint reale.

Tabella 1: Tabella per PCK@0.5

	RT	**I**	**II**	**III**	**IV**
right ankle	54.0 50.0	53.5 49.0	50.0 44.9	59.1 53.5	62.1 56.1
right knee	47.5 41.5	52.5 46.0	48.5 43.5	56.0 50.5	62.5 56.5
right hip	55.0 48.5	53.0 46.5	50.5 44.0	63.0 56.5	63.5 57.0
left hip	55.0 48.5	55.0 48.5	52.0 46.0	61.5 55.5	62.0 55.5
left knee	50.5 44.5	49.0 43.0	48.0 43.0	56.5 51.0	59.5 53.5
left ankle	54.5 51.0	53.0 46.5	50.5 44.9	57.1 51.5	62.1 56.6
pelvis	54.5 48.0	54.5 48.0	50.5 44.5	62.5 56.5	62.5 56.0
thorax	64.5 57.0	60.0 53.5	63.5 57.5	71.5 66.0	70.5 63.5
upper neck	62.5 55.0	57.5 51.0	63.5 57.5	67.5 61.5	67.5 60.0
head top	54.3 48.2	56.3 50.3	53.3 47.2	57.8 51.8	58.8 52.8
right wrist	52.5 47.0	56.5 50.0	54.5 50.0	68.5 62.5	70.5 63.5
right elbow	54.8 48.7	56.8 53.3	52.8 47.2	65.8 59.8	67.3 60.8
right shoulder	57.5 51.0	55.0 49.0	57.0 51.5	67.5 63.0	68.5 62.5
left shoulder	60.0 53.5	57.5 51.5	59.0 53.5	64.5 59.5	68.5 63.0
left elbow	58.3 52.3	55.3 49.7	55.3 49.7	64.3 58.8	66.3 59.8
left wrist	53.0 47.5	54.0 49.2	54.5 50.8	63.0 57.8	66.0 58.8

Tabella 2: Tabella per PCP@0.5

	RT	**I**	**II**	**III**	**IV**
head_t, u_neck	0.0 0.0	0.0 0.0	0.5 0.5	0.0 0.0	0.0 0.0
u_neck, thorax	13.0 11.0	12.0 10.0	12.5 10.5	11.5 9.5	12.0 10.5
thorax, r_shoulder	37.5 33.0	42.0 36.0	39.0 33.5	52.0 46.5	54.0 47.5
thorax, l_shoulder	40.0 35.5	39.0 33.0	38.0 31.5	52.0 47.0	51.0 44.5
thorax, pelvis	50.5 44.0	49.5 43.0	45.5 39.5	58.5 52.5	60 53.5
r_shoulder, r_elbow	38.7 34.7	43.7 37.2	40.7 35.2	53.3 47.7	57.3 50.8
r_elbow, r_wrist	31.7 27.6	36.2 30.2	35.2 29.6	57.7 42.7	48.2 42.2
l_shoulder, l_elbow	43.7 28.2	42.2 36.7	41.2 36.2	53.3 48.7	55.8 50.3
l_elbow, l_wrist	31.2 26.8	36.2 30.8	34.7 29.8	45.7 41.4	47.2 41.9
pelvis, r_hip	15.0 13.5	14.0 12.0	18.0 14.5	20.5 18.5	24.0 21.0
pelvis, l_hip	14.0 12.0	16.0 13.5	16.5 13.0	19.0 16.5	22.0 19.5
r_hip, r_knee	38.0 33.0	36.5 32.0	34.0 30.0	47.5 42.5	49.5 43.0
r_knee, r_ankle	34.8 30.3	34.8 30.3	32.8 29.8	46.0 40.9	49.0 43.4
l_hip, l_knee	38.5 33.0	36.5 31.0	32.5 29.0	47.0 42.0	48.5 42.5
l_knee, l_ankle	34.3 28.8	34.8 29.8	34.3 30.8	46.0 40.9	51.5 45.5

Tabella 3: Tabella per PDJ@0.5

	RT	**I**	**II**	**III**	**IV**
head_t, u_neck	48.2 41.2	48.2 41.7	46.7 40.2	51.8 45.7	50.8 44.7
u_neck, thorax	62.0 54.0	57.0 50.5	61.0 55.0	66.5 60.5	65.5 58.5
thorax, r_shoulder	56.0 49.0	53.5 47.5	54.5 49.0	65.5 61.0	67.5 61.0
thorax, l_shoulder	57.0 50.5	54.5 48.5	55.5 50.0	64.5 59.5	66.0 59.0
thorax, pelvis	52.0 45.5	51.0 44.5	48.0 42.0	59.5 53.5	60.5 54.0
r_shoulder, r_elbow	48.7.0 42.7	51.3 45.2	48.2 42.7	62.3 56.8	65.3 55.5
r_elbow, r_wrist	46.2 41.2	51.3 44.7	48.7 43.7	63.3 57.3	64.3 58.8
l_shoulder, l_elbow	51.8 45.2	48.7 43.7	48.2 42.7	59.8 54.8	62.8 57.8
l_elbow, l_wrist	45.7 40.9	47.7 42.9	47.7 42.9	58.3 53.5	60.8 56.3
pelvis, r_hip	53.5 47.0	52.5 46.0	49.5 43.0	61.5 55.0	62.0 54.0
pelvis, l_hip	54.0 47.0	53.5 47.0	50.0 44.0	61.0 55.0	61.0 55.5
r_hip, r_knee	44.0 38.0	47.5 41.5	44.5 39.0	54.0 48.0	57.0 50.5
r_knee, r_ankle	43.4 37.9	45.5 40.4	39.9 35.9	51.0 46.0	56.1 50.5
l_hip, l_knee	45.5 39.5	46.5 40.5	44.0 39.5	54.0 48.5	55.0 48.5
l_knee, l_ankle	42.4 36.9	40.9 35.4	40.9 35.9	51.5 46.5	54.0 48.0

Analizzando le tabelle è chiaro che i valori di destra di ogni riga e colonna sono più bassi rispetto ai valori di sinistra dimostrando empiricamente che se il dataset non contiene immagini di persone da sole allora EfficientPose si comporta peggio nel task di rilevamento.

Una considerazione sui risultati di PCP è doverosa. I 4 valori più bassi potrebbero essere causati da head_top, upper_neck e pelvis. In particolare:

- il modo in cui viene calcolato head_top non è ottimale, nonostante il principio di individuazione di head_top sia basato su degli studi anatomici, probabilmente perché le proporzioni proposte da Avard Tennyson Fairbanks riguardano rapporti ideali (la scultura e la pittura) di una persona.

- pelvis invece non può essere calcolato in modo diverso poiché il bacino è necessariamente a metà tra le due anche. I valori bassi del suo PCK sono spiegabili guardando i valori di PCK delle due anche. EfficientPose non inferisce bene le parti del corpo dal busto in giù; e questa osservazione è rilevabile anche in MPII.

- Gli zeri legati all'arto (head_top, upper_neck) indicano che le inferenze su questi keypoint generano almeno un Falso Positivo tra head_top e upper_neck; infatti basta che uno dei due stia al di fuori del raggio definito da una frazione della lunghezza dell'arto reale e PCP non considera correttamente predetto l'arto. Visto che entrambi head_top e upper_neck di ground truth vengono calcolati, seguendo le indicazioni riportate sopra, molto probabilmente questi zeri in realtà potrebbero essere valori più alti se disponibili i veri termini di ground truth.

Oltre ai risultati realizzati sulle diverse metriche (con $\tau = 0.5$) è stato creato un plot che mostra l'andamento delle metriche al variare del parametro τ. Dall'analisi di queste curve si riesce a stabilire quale fra i modelli di EfficientPose risulta migliore. La AUC (area under the curve) permette di comparare l'efficacia dei modelli collassando le curve ad un semplice valore (l'area sotto la curva appunto) consentendo un'individuazione più semplice tra migliori e peggiori.

Di seguito si trovano la tabella delle AUC per ogni metrica (come sempre i valori a sinistra in ogni cella indicano il valore della metrica per il dataset avente immagini di persone da sole mentre il valore di destra si riferisce a immagini aventi più persone) e le immagini delle curve sull'andamento delle metriche al variare del parametro τ.

Tabella 4: Tabella dell'AUC

	RT	**I**	**II**	**III**	**IV**
PCK	50.35 45.12	49.67 44.48	49.15 44.27	57.17 52.19	59.12 53.37
PCP	26.40 23.07	27.15 23.47	26.46 22.93	34.09 30.64	36.08 31.80
PDJ	44.69 39.42	44.45 39.29	43.66 38.84	52.47 47.64	54.17 48.46

Figura 3: Valori di PCK al variare di tau per il dataset single-person

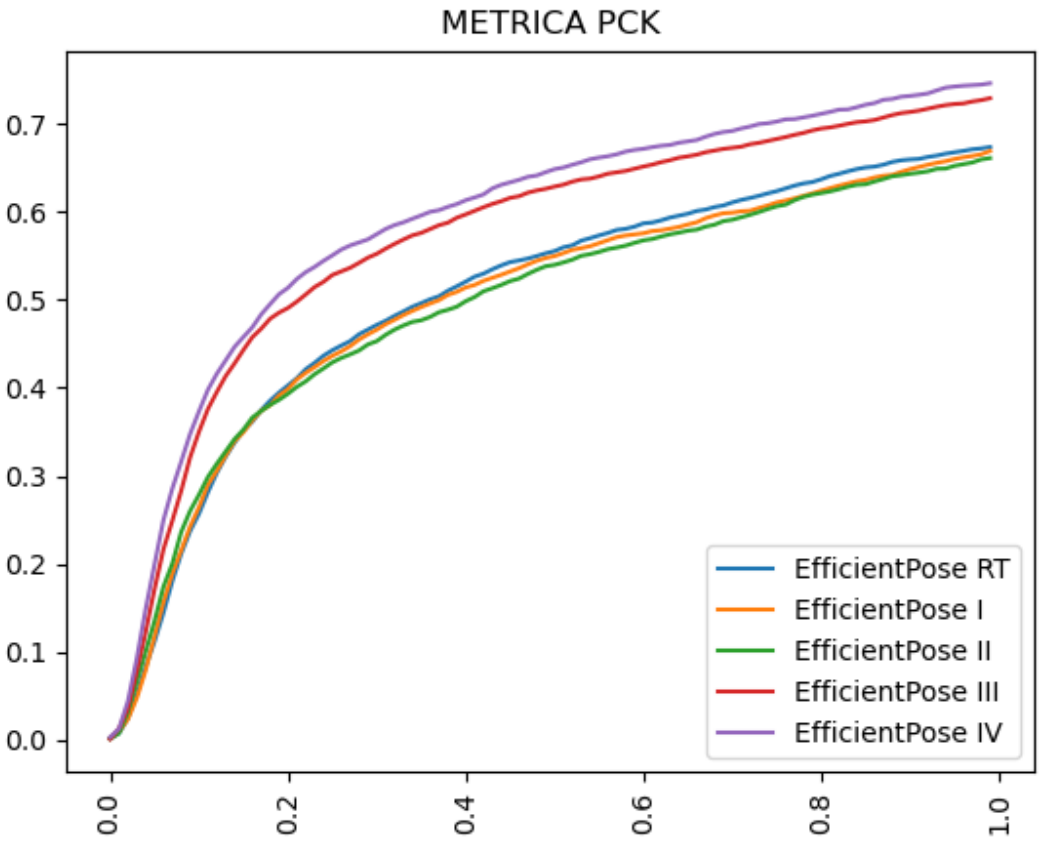


Figura 4: Valori di PCP al variare di tau per il dataset single-person

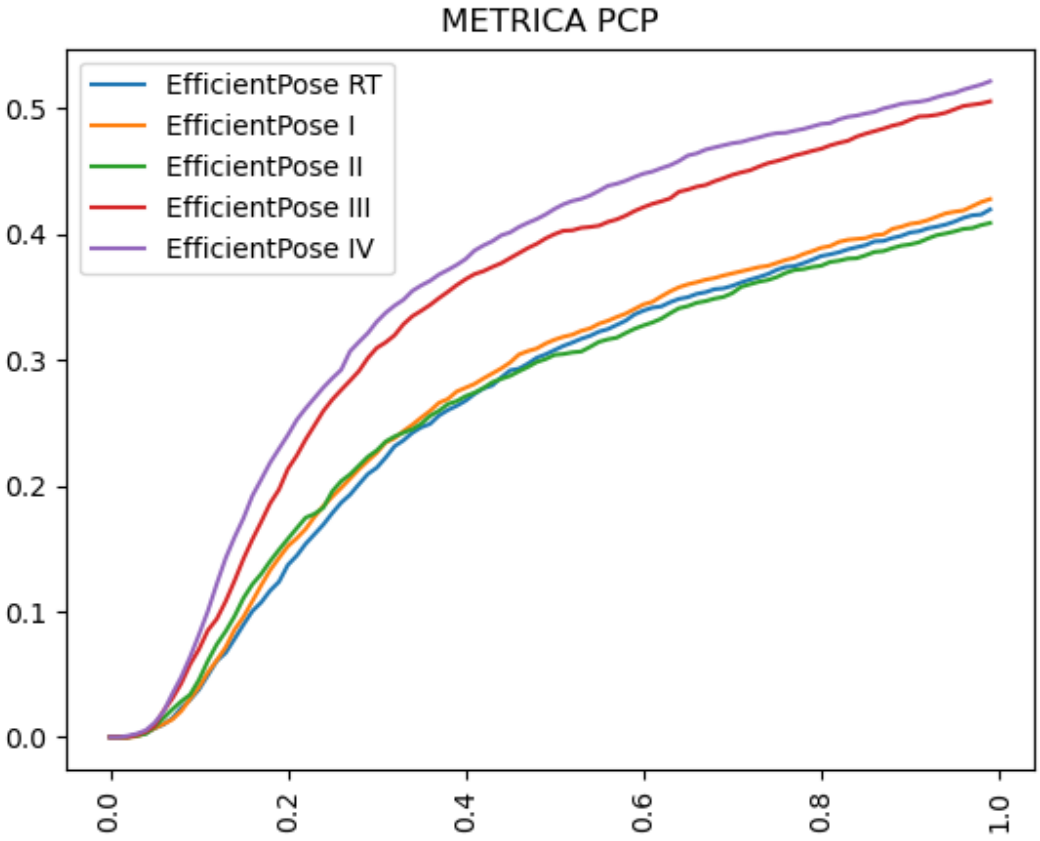


Figura 5: Valori di PDJ al variare di tau per il dataset single-person

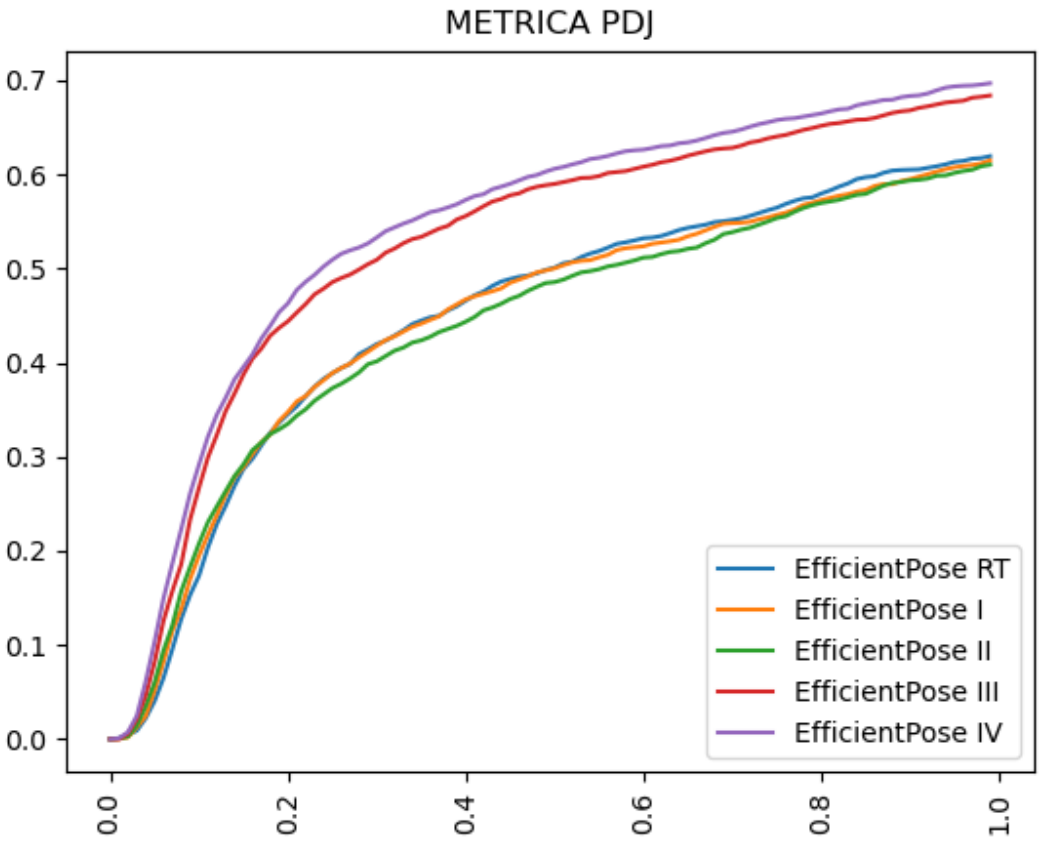


Figura 6: Valori di PCK al variare di tau per il dataset multi-person

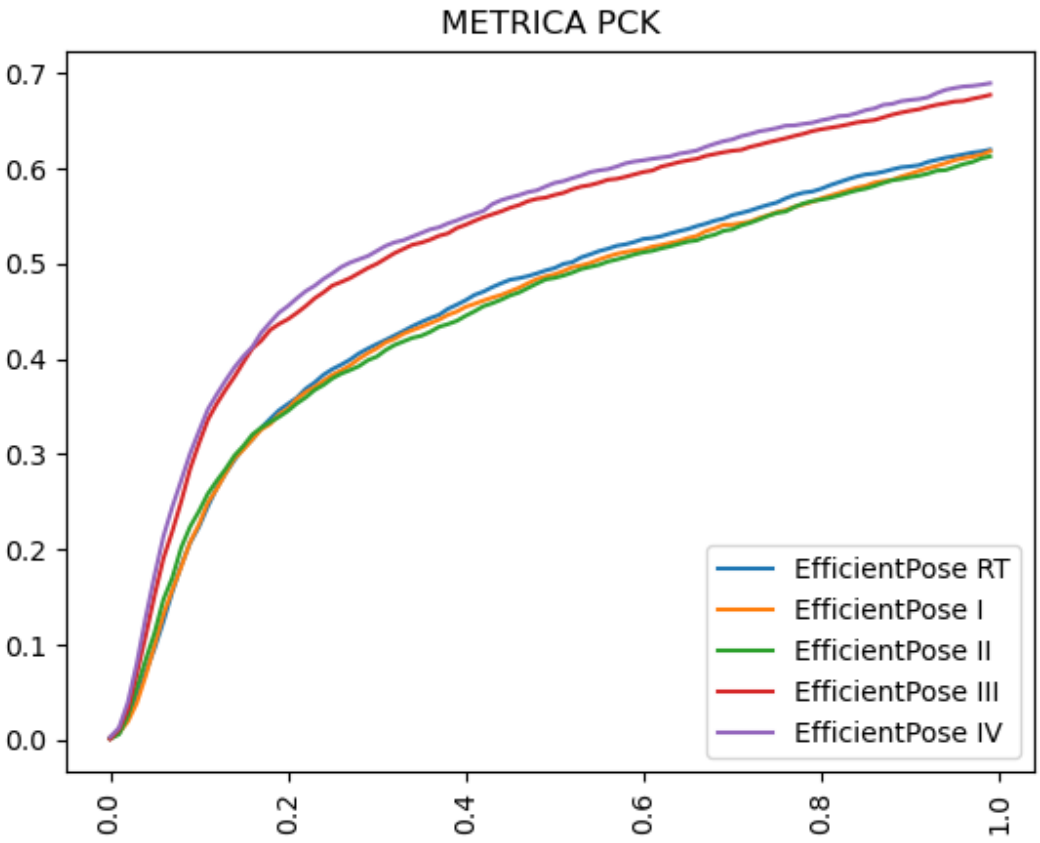


Figura 7: Valori di PCK al variare di tau per il dataset multi-person

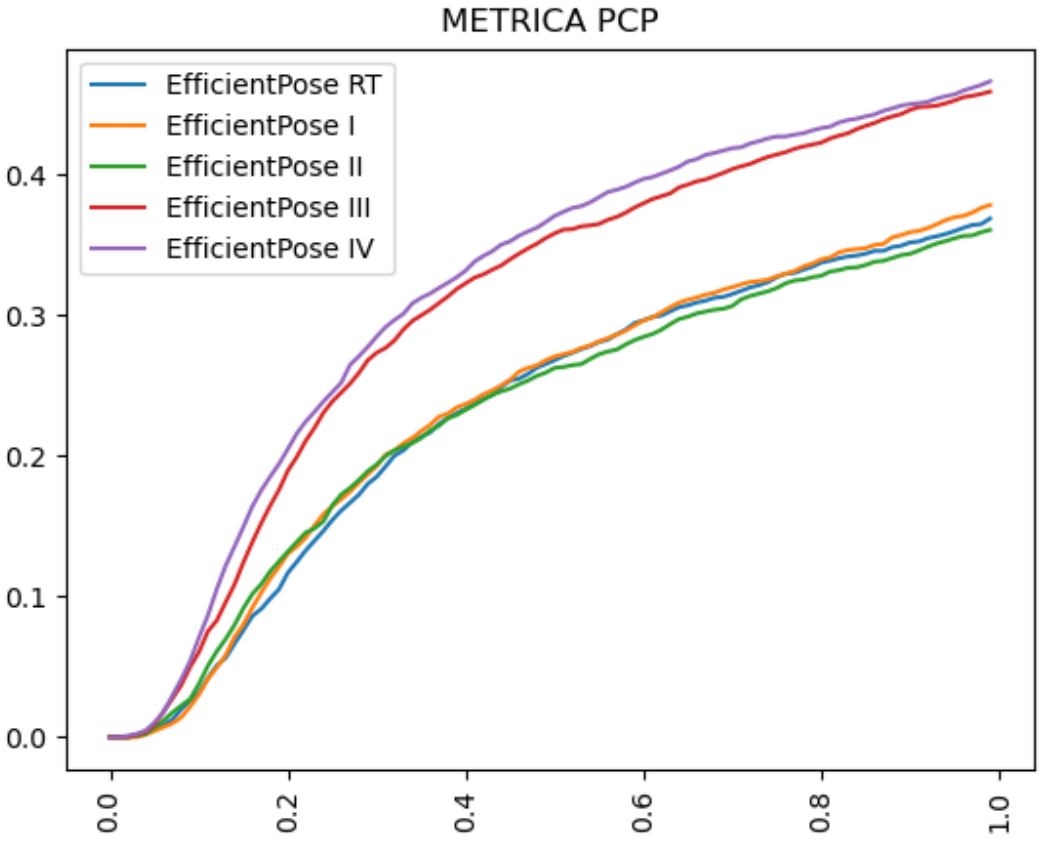
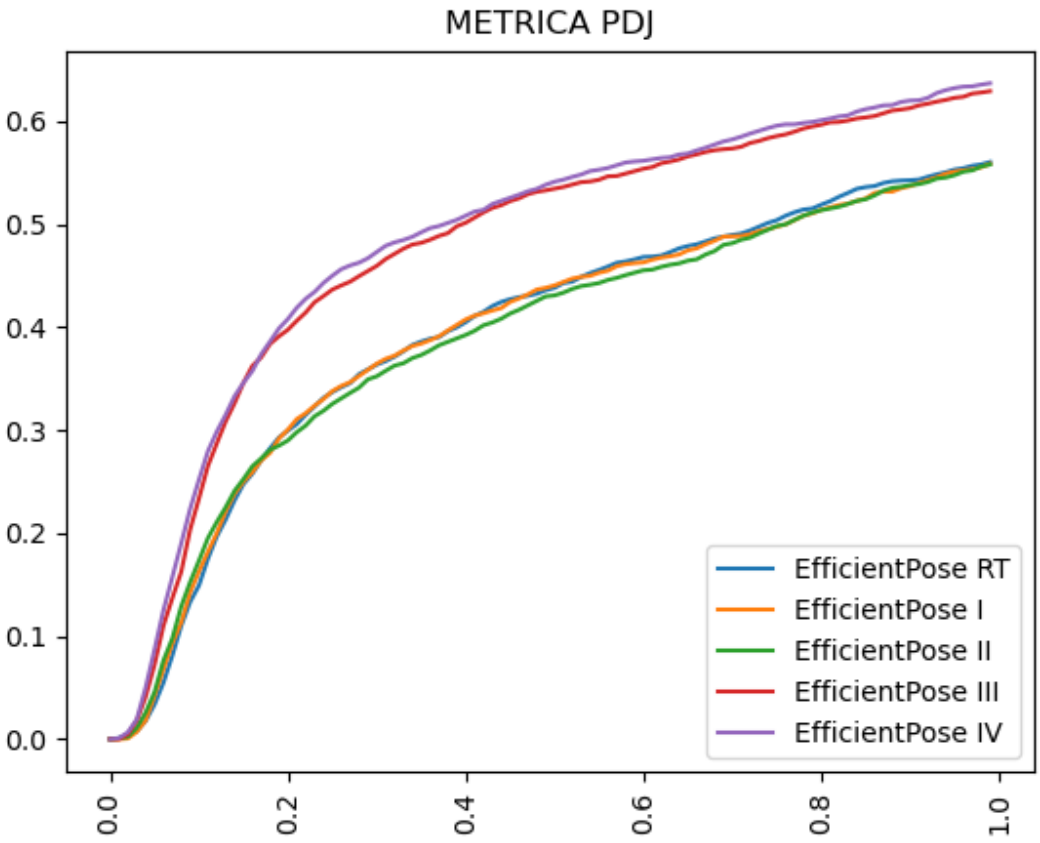


Figura 8: Valori di PCK al variare di tau per il dataset multi-person



3.2.1 Breve riflessione sui risultati

Considerando i valori delle metriche al variare di τ (tralasciando se il dataset è single o multi person) è interessante osservare che i modelli invece di migliorare in modo progressivo all'aumentare del modello (RT < I < ... < IV) presentano una sorta di stazionarietà (almeno per EfficientPose RT, I e II) tale da poterli dividere in due gruppi: il gruppo di "ristagno" formato da RT,I e II e il secondo gruppo in cui è presente un miglioramento di circa due punti percentuali (in ogni metrica) da parte di EfficientPose IV rispetto a III. In generale il miglioramento offerto da EfficientPose IV rispetto a RT è del 20% su PCK, del 40% su PCP e del 22% per PDJ (miglioramenti rispetto al valore base di PCK).

4 Leeds Sports Pose Dataset

Questo dataset contiene 2000 immagini di sportivi raccolte utilizzando tag come: atletica, badminton, baseball, ginnastica, parkour, calcio, tennis, pallavolo. Le immagini sono state ritagliate in modo da racchiudere il soggetto principale e da avere una lunghezza di circa 150 pixel. Ogni immagine è stata poi annotata con 14 keypoint. Il file *joints.mat* è il file di MATLAB che contiene le annotazioni; si tratta di una matrice $3 \times 14 \times 2000$ denominata *joints* in cui:

- La prima dimensione si riferisce alle posizioni x e y e ad un valore binario che indica la visibilità di ciascun keypoint.
- La seconda dimensione si riferisce all'ordine dei keypoint (Right ankle, Right knee, Right hip, Left hip, Left knee, Left ankle, Right wrist, Right elbow, Right shoulder, Left shoulder, Left elbow, Left wrist, Neck, Head top).
- La terza dimensione si riferisce al numero dell'immagine.

4.1 Comparazione dei risultati

Il metodo di valutazione utilizzato dagli autori del dataset si basa sulla metrica PCP@0.5 (i keypoint di un arto predetti devono trovarsi entro il 50% della lunghezza dell'arto calcolata come distanza tra i keypoint veri). Nella seguente tabella si presenta la comparazione tra i risultati ottenuti con il metodo proposto dagli autori del dataset *LSP* e i risultati ottenuti utilizzando i modelli del framework EfficientPose *RT, I, II, III, IV*.

Nota: le colonne con due numeri mostrano i risultati rispettivamente per gli arti sinistro e destro.

	Total	**Torso**	**Upper Leg**	**Lower Leg**	**Upper Arm**	**Forearm**	**Head**
LSP	55.1	78.1	64.8 66.7	60.3 57.3	48.3 46.5	34.5 31.2	62.9
RT	67.6	95.5	84.5 86.5	79.0 78.5	75.0 79.5	61.5 70.5	74.0
I	67.7	95.5	84.5 86.0	78.0 79.5	73.5 78.0	59.0 67.0	75.5
II	70.1	97.0	87.5 89.0	86.0 86.0	78.0 83.5	67.5 73.5	71.5
III	72.2	97.0	90.5 89.5	89.5 91.0	82.5 83.5	71.0 76.0	72.0
IV	72.5	96.5	89.5 89.5	91.0 91.0	78.0 80.0	71.5 74.5	78.0

Si può notare come EfficientPose migliori in media di un 12-17% il metodo proposto dagli autori del dataset. Ciò potrebbe essere dovuto al fatto che il dataset LSP sembra soddisfare perfettamente i requisiti richiesti da EfficientPose per ottenere la precisione ottimale, ovvero:

- Assicurarsi che nell'immagine sia presente una sola persona.

- Assicurarsi che l'intero corpo della persona sia chiaramente visibile e si trovi vicino al centro dell'immagine.
- Evitare che il soggetto sia occluso, anche parzialmente, da altri oggetti.

4.2 Differenze tra LSP e output di EfficientPose

Il framework EfficientPose restituisce predizioni per 16 keypoint (top of head, upper neck, shoulders, elbows, wrists, thorax, pelvis, hips, knees, ankles). Quindi, per far corrispondere i keypoint predetti con quelli annotati sono state effettuate le seguenti operazioni:

- Il keypoint *neck* di LSP viene ricavato dal punto medio tra *upper neck* e *thorax* di EfficientPose.
- Il keypoint *pelvis* di LSP non è annotato e quindi viene ricavato dal punto medio tra *right hip* e *left hip* (sempre di LSP).

4.3 Comparazione dei modelli di EfficientPose

Nei grafici sottostanti viene presentata la comparazione tra i vari modelli di EfficientPose mostrando i risultati per le metriche PCP, PDJ e PCK facendo variare la soglia (da 0 a 1 con passo 0.1) per poter calcolare l'area sotto la curva (AUC) e valutare così la bontà dei modelli. Infatti, la metrica AUC misura quanto un modello è in grado di distinguere le varie articolazioni del corpo: maggiore è il valore di AUC, migliore è il modello.

	AUC per PCP	**AUC per PDJ**	**AUC per PCK**
RT	0.565	0.793	0.823
I	0.569	0.794	0.825
II	0.595	0.817	0.851
III	0.604	0.825	0.854
IV	0.613	0.826	0.856

Figura 9: Comparazione dei valori per la metrica PCP (dataset LSP)

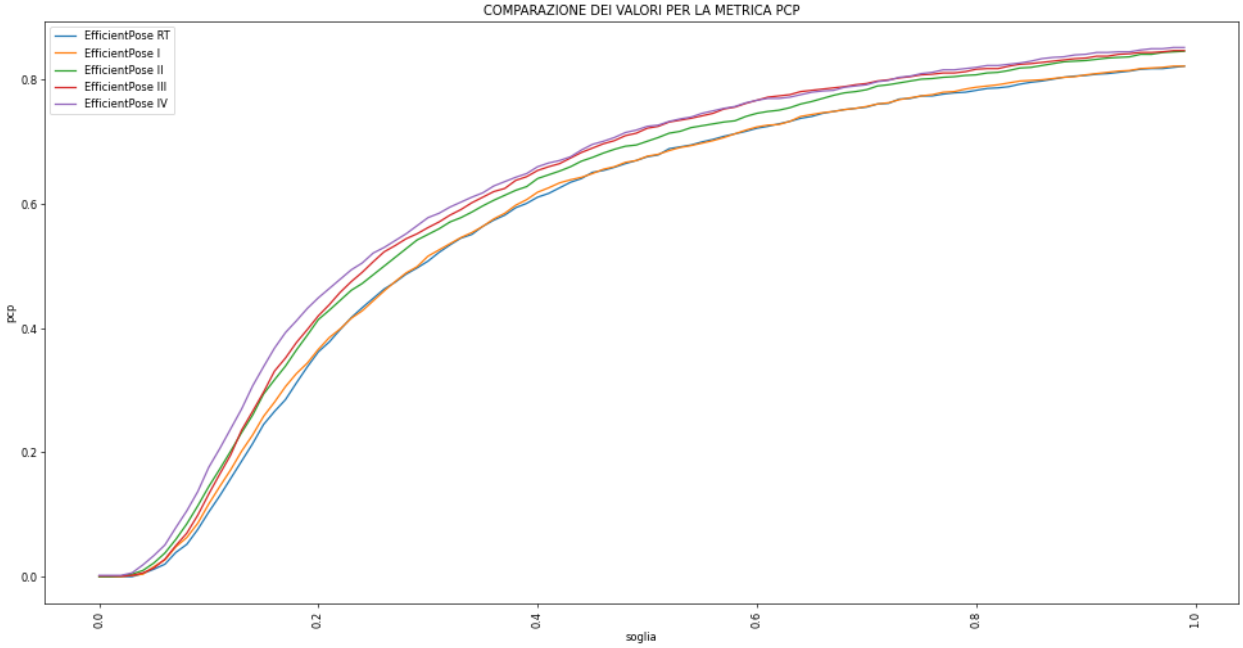


Figura 10: Comparazione dei valori per la metrica PDJ (dataset LSP)

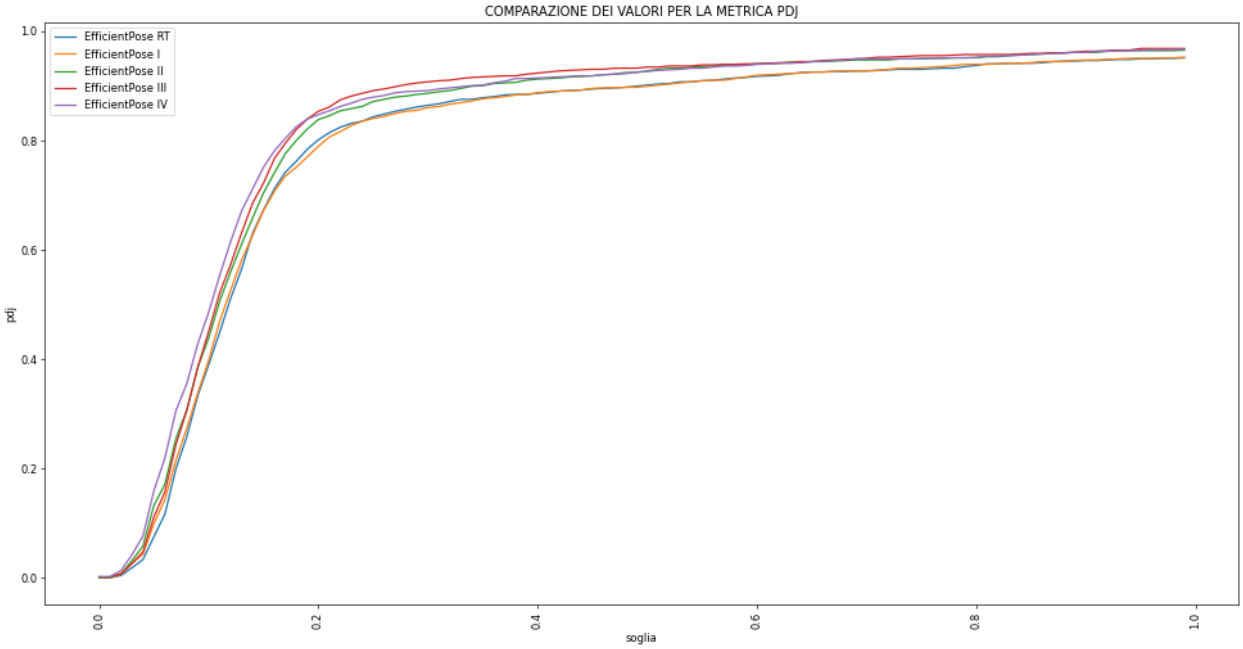
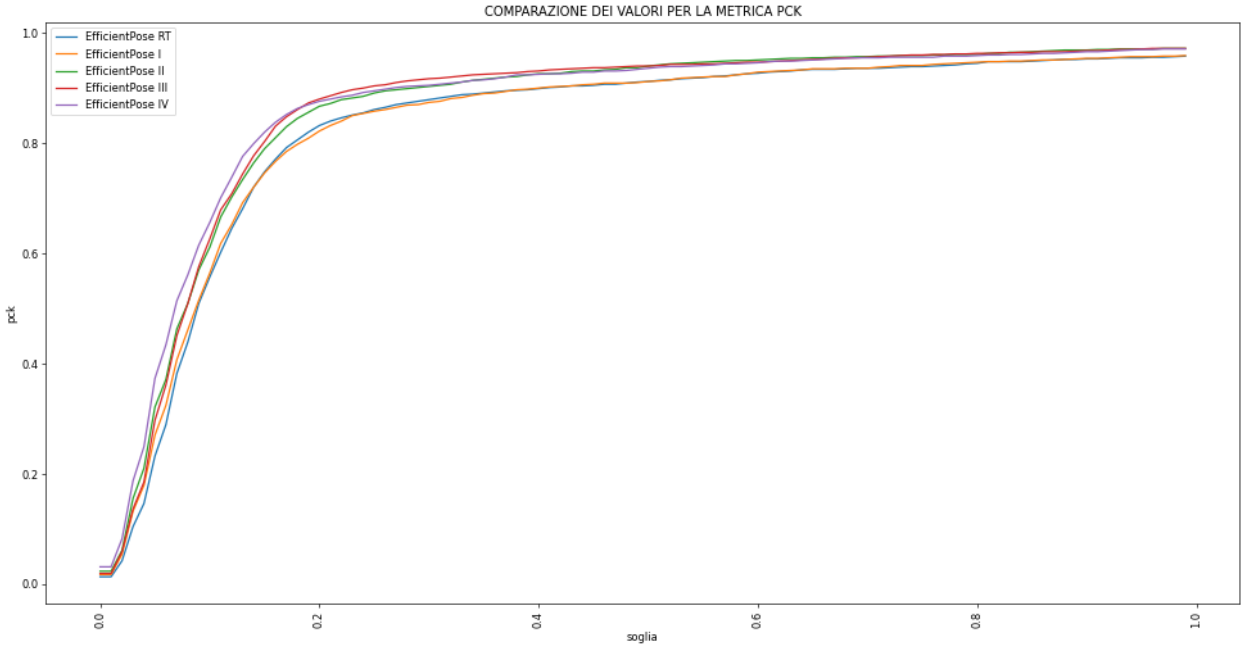


Figura 11: Comparazione dei valori per la metrica PCK (dataset LSP)

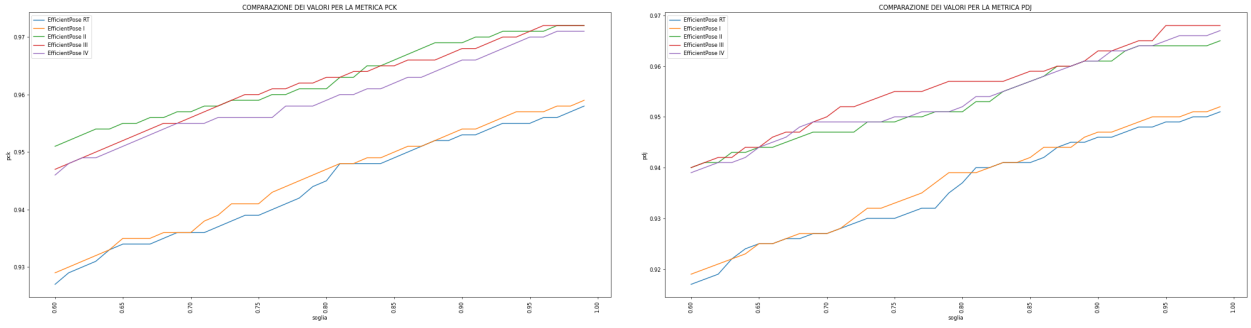


4.4 Note tecniche

La tabella sottostante mostra le tempistiche medie di inferenza impiegate da ciascun modello di EfficientPose (in termini relativi rispetto al modello più veloce, ovvero RT). In particolare, i dati si riferiscono ad un'analisi eseguita su 200 immagini utilizzando il framework Pytorch.

Modello	*RT*	*I*	*II*	*III*	*IV*
Tempo di inferenza	1x	1.6x	6.3x	42.6x	129.5x

Figura 12: Trend delle metriche PCK e PDJ (dataset LSP)



Osservando le tempistiche impiegate per l'inferenza e i risultati delle metriche PCK e PDJ viene confermato che EfficientPose II realizza il miglior compromesso tra efficienza e pre-

cisione poiché riesce sostanzialmente a raggiungere la precisione di EfficientPose IV benché abbia un costo computazionale più vicino ad EfficientPose RT.

5 Riferimenti

1. Groos, Daniel and Ramampiaro, Heri and Ihlen, Espen AF (2021) EfficientPose: Scalable single-person pose estimation. In: Applied Intelligence 51:2518–2533 (Springer)

2. Andriluka M, Pishchulin L, Gehler P, Schiele B (2014) 2d human pose estimation: new benchmark and state of the art analysis. In: IEEE Conference on computer vision and pattern recognition (CVPR)

3. Y. Yang and D. Ramanan (2013) Articulated human detection with flexible mixtures of parts. In: PAMI

4. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick (2014) Microsoft COCO: Common Objects in Context. In: Proc. ECCV

5. Sam Johnson and Mark Everingham (2010) Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In: Proceedings of the 21st British Machine Vision Conference (BMVC)