Penerapan 3D Human Pose Estimation Indoor Area untuk Motion Capture dengan Menggunakan YOLOv4-Tiny, EfficientNet Simple Baseline, dan VideoPose3D

Gerry Steven, Liliana, Anita Nathania Purbowo
Program Studi Informatika Fakultas Teknologi Industri Universitas Kristen Petra
Jl. Siwalankerto 121 – 131 Surabaya 60236
Telp. (031) – 2983455, Fax. (031) – 8417658
E-mail: sgerrys93@gmail.com, lilian@petra.ac.id, anita.nathania@petra.ac.id

ABSTRAK

Human pose estimation merupakan topik penelitian yang bertujuan untuk mengestimasi letak masing-masing sendi yang akan dihubungkan sehingga membentuk kerangka manusia yang diambil dari gambar dan video. Pengembangann topik ini dapat diaplikasikan menjadi deteksi aktivitas manusia, pelacakan manusia, dan dunia hiburan seperti *motion capture* animasi dan perfilman. Namun topik ini juga memiliki tantangan: salah satunya bentuk tubuh yang beragam termasuk dengan pakaian, dan lingkungan yang kompleks yang menyebabkan susah membedakan *foreground* dan *background.*

Metode yang digunakan dalam penelitian ini adalah *YOLOv4-Tiny*, *EfficientNet Simple Baseline*, dan *VideoPose3D.* Dimana *YOLOv4-Tiny* akan menerima masukan gambar untuk mendapatkan koordinat *bounding box* manusia yang selanjutnya akan dimasukan ke dalam modifikasi *EfficientNet Simple Baseline* sehingga akan menghasilkan 16 koordinat 2D. Setelah didapatkan koordinat 2D, akan dimasukan ke dalam model *VideoPose3D* untuk mengestimasikan 15 koordinat 3D.

Hasil yang didapatkan penelitian ini adalah modifikasi menghasilkan kecepatan yang lebih cepat dengan waktu 4.54ms sedangkan Simple Baseline yang menghasilkan kecepatan 5,15ms. Akan tetapi akurasi modifikasi akan kurang akurat dengan rata-rata *Percentage of Correct Keypoints head* (PCKh@0.2) tertinggi sebesar 86.89% dibandingkan Simple Baseline dengan rata-rata PCKh@0.2 sebesar 89.62%. Hal ini juga mempengaruhi hasil estimasi koordinat 3D menggunakan VideoPose3D dimana menggunakan modifikasi EfficientNet menghasilkan rata-rata *Mean Per Joints Position Error* (MPJPE) sebesar 25,3mm sedangkan Simple Baseline menghasilkan MPJPE sebesar 28.1 mm.

Kata Kunci: Human Pose Estimation, YOLO, EfficientNet, Simple Baseline, VideoPose3D

ABSTRACT

Human pose estimation is a research topic that has goal to estimate every human's keypoint coordinate that can be connected and make a human skeleton. The development of this topic can be applicated to human activity recognition, human tracking, and motion capture for film and animation. There are several challenges for this topic: diverse human pose, diverse body appearance from clothing and similar parts, and complex environment that may cause foreground occlusion.

There are several methods to be used in this research: YOLOv4-Tiny, EfficientNet Simple Baseline, and VideoPose3D. YOLOv4-Tiny will process image input to get bounding box coordinate. This coordinate will be inputted to EfficientNet Simple Baseline modification to get 16 keypoint 2D coordinates. After that, VideoPose3D will processed 2D coordinates into 15 keypoints 3D coordinates.

The result from this research is EfficientNet Simple Baseline modification is faster with 4.54ms time compared to its original with time of 5.15ms. Although faster, its modification has its own downside. In term of accuracy, modification still less accurate than its original with highest average Percentage of Correct Keypoints head (PCKh@0.2) 86.89%, and original with PCKh@0.2 89.62%. This affect 3D human pose estimation using VideoPose3D, where using EfficientNet modification resulting Mean Per Joints Position Error (MPJPE) 25.3 mm compared to original Simple Baseline resulting MPJPE 28.1mm.

Keyword: Human Pose Estimation, YOLO, EfficientNet, Simple Baseline, VideoPose3D

1. PENDAHULUAN

Human pose estimation merupakan estimasi bagian-bagian tubuh manusia terutama letak sendi-sendi yang akan dihubungkan sehingga membentuk kerangka manusia yang diambil oleh sensor-sensor seperti video dan gambar [30]. Topik penelitian ini merupakan topik penelitian yang cukup maju karena solusi *deep learning* yang melebihi metode *computer vision* pada biasanya, seperti: klasifikasi gambar [12, 13], segmentasi sematik [16] dan deteksi objek [20]. Dengan pengembangan *human pose estimation* lebih lanjut dapat diaplikasikan menjadi deteksi aktivitas manusia, [14, 17], pelacakan manusia [11, 32], dan dunia hiburan seperti *motion capture* animasi dan perfilman.

Terdapat beberapa tantangan dalam melakukan *human pose estimation.* Yang paling sering adalah fleksibilitas tubuh yang berbeda sehingga bentuk tubuh dapat bergerak menjadi suatu pose yang kompleks atau jarang, bentuk tubuh yang beragam termasuk dengan pakaian, dan lingkungan yang kompleks yang menyebabkan susah membedakan *foreground* dan *background* [1]. Solusi yang digunakan untuk *human pose estimation* sebelumnya menggunakan teknik *computer vision* [6, 7, 22, 24, 28]. Akan tetapi, seiring berkembangnya *deep learning, deep learning* lebih akurat dibandingkan metode *computer vision* biasanya khususnya dalam pengambilan fitur pada lingkungan yang cukup kompleks.

Terdapat beberapa penelitian *human pose estimation* sebelumnya. Penelitian *human pose estimation* dibagi menjadi *single person estimation* dan *multi person estimation.* Kemudian, setiap estimasi dibagi menjadi 2D dan 3D. Untuk estimasi 2D, Metode *convolutional neural network* yang digunakan ada yang menggunakan *heatmap* sebelum melakukan prediksi [2, 3, 5, 27, 29] dan *direct regression* [4, 8, 17]. Sedangkan untuk estimasi 3D, Sebagian besar menggunakan *input* dari koordinat 2D dan diestimasikan hasil 3D melalui *network* yang digunakan seperti

[26, 31, 18, 19]. Dimana didapatkan hasil akurasi yang cukup tinggi.

Karena itulah, dalam penelitian ini menggunakan metode *deep learning* untuk melakukan *human pose estimation.* Penelitian ini pertama-tama menggunakan YOLOv4-Tiny sebagai deteksi manusia dimana akan diambil *bounding box.* Setelah didapatkan *bounding box*, kemudian menggunakan metode *heatmaps* untuk estimasi 16 koordinat *keypoint* 2D. Estimasi menggunakan *convolutional neural network heatmaps simple baseline* dan modifikasi *backbone* dengan mengubah *backbone* dari *ResNet* menjadi EfficientNet. Modifikasi ini didasari karena Penggabungan kedua metode ini terinspirasi dari [9] yang memodifikasi *framework OpenPose* dengan *EfficientNet* dimana menghasilkan *parameter* yang lebih sedikit dan hasil yang lebih bagus. Alasan lainnya, EfficientNet memiliki besar model yang rendah dengan akurasi tinggi pada ImageNet [10], sedangkan ResNet memiliki model yang lebih besar dan akurasi yang kurang dari EfficientNet [23]. Sehingga diharapkan dengan modifikasi ini akan menghasilkan *model* yang lebih ringan dan akurat. *Model* kemudian dilanjutkan dengan estimasi 3D 15 koordinat *keypoint* 3D menggunakan VideoPose3D dengan *input* hasil estimasi 2D sebelumnya. Sehingga pada akhirnya, *model* akan menghasilkan hasil estimasi 15 koordinat *keypoint* 3D dari video atau gambar yang dimasukan.

2. TINJAUAN STUDI

Bazarevsky et al. [3] melakukan penelitian *human pose estimation* pada *embedded device* dengan menggunakan *BlazePose*. Dengan menggunakan arsitektur CNN akan dicari manusia dalam setiap *frame* dan diprediksi setiap sendi menggunakan *heatmap*. Arsitektur *heatmap* menggunakan *stacked hourglass*. Apabila tidak terdeteksi manusia dalam suatu *frame*, akan diulang dari deteksi manusia. Hasil penelitian ini didapatkan hasil prediksi yang cukup cepat dan akurat dibandingkan dengan *framework OpenPose* dengan *Percentage Correct Keypoints* (PCK)@0.2 sebesar 84,1 pada *augmented reality* (AR) *dataset* dan PCK@0.2 sebesar 84,5 pada Yoga *dataset* dengan kecepatan 10 FPS. Kekurangan dari penelitian ini adalah kurang konsistennya kerangka manusia jika diestimasikan dari samping.

Xiao et al. [27] melakukan penelitian *multi human pose estimation* 2D dengan *heatmap* menggunakan *backbone* ResNet yang diikuti dengan 3 *deconvolutional layer* dengan *batch normalization* dan ReLu *activation layer.* Hasil dari penelitian ini lebih bagus dibandingkan dengan dibandingkan 8-*stage-houglass* dan *cascaded pyramid network* (CPN) yang biasa digunakan untuk *heatmap* dengan hasil 72,2 *Average Precision* (AP). Kekurangan dari penelitian ini adalah *backbone network* yang digunakan harus cukup dalam seperti ResNet-152 untuk mendapatkan akurasi yang cukup tinggi.

Groos et al. [9] melakukan modifikasi terhadap *framework OpenPose* dengan improvisasi menggunakan *EfficientNet*. Hasil dari penelitian ini didapatkan hasil yang lebih bagus daripada *framework OpenPose* sebelumnya, salah satunya *model EfficientPose RT* dengan PCK@50 sebesar 82.88 dengan Parameter 0.46M. Kekurangan dari penelitian ini karena menggunakan *EfficientNet*, maka semakin kecil *model*, harus mengorbankan akurasi untuk kecepatan.

Mehta et al. [18] mengimplementasikan *human pose estimation* menjadid *motion capture real-time* menggunakan kamera RGB-D. Peneliti menggunakan *ResNet-50* untuk estimasi 2D dan *location maps* dari koordinat 2D sebelumnya diprediksi. Hasil penelitian ini didapatkan hasil *Mean Per Joints Position Error* (MPJPE) sebesar 124,7mm. Kekurangan dari penelitian ini adalah masih besarnya MPJPE dibandingkan penelitian lain dan kurang bisa mendeteksi sendi bagian kaki.

Pavllo, Feichtenhofer, Grangier, & Auli [19] melakukan 3D *human pose estimation* dengan *dilated temporal convolutions* dari koordinat 2D. *Model* menggunakan *n receptive field* untuk mengambil koordinat 2D dari *n frame* sebelumnya sebagai *input* untuk estimasi 3D. Hasil dari penelitian ini adalah *model* yang lebih akurat disbanding metode sebelumnya yaitu MPJPE sebesar 46.8mm pada dataset *human3.6m*. Kekurangan dari penelitian ini adalah akurasi pose orang sedang duduk masih kurang bagus dibandingkan metode yang dibandingkan.

3. METODOLOGI

3.1. YOLOv4-Tiny

YOLOv4-Tiny (*You Only Look Once*) [25] merupakan model *object detection* YOLOv4 yang dilakukan *scaling* menjadi lebih kecil dengan tujuan untuk mengorbankan akurasi untuk kecepatan yang paling optimal. *Scaling* model dilakukan dengan menggantikan *backbone* YOLOv4 dengan menggunakan CSPOSANet yang memiliki arsitektur PCB. Sehingga, nantinya jumlah *convolutional layer* akan semakin mengecil dengan jumlah 29 *convolutional layer* dibandingkan YOLOv4 dengan jumlah 137 *convolutional layer* dan perubahan YOLOv3 *head*, dimana sebelumnya memiliki 3 *head* pada YOLOv4, pada YOLOv4-Tiny hanya memiliki 2 *head*.

3.2. Simple Baseline

Simple baseline [27] merupakan metode untuk *human pose estimation* 2D. Metode ini nantinya akan menghasilkan *heatmap* letak masing-masing sendi yang akan diprediksi dan diambil hasil maksimal dari *heatmap*.

Simple baseline menggunakan 3 *deconvolutional layer* pada *network* akhir *backbone*. *Deconvolutional layer* akan menghasilkan *heatmap* dan kemudian bagian akhirnya menggunakan *convolutional layer* 1x1 untuk mendapatkan hasil prediksi maksimal dari *heatmap* menjadi koordinat setiap sendi.

3.3. EfficientNet

EfficientNet [23] merupakan salah satu model *convolutional neural network*. EfficientNet terinspirasi karena model sebelumnya seperti ResNet semakin dalam, menghasilkan akurasi yang bagus. Maka dari itu, peneliti mengembangkan EfficientNet dimana keluasan dan kedalaman model akan dibuat sama atau *scale uniform*.

$$\begin{aligned} &depth: d = \alpha^{\phi} \\ &width: w = \beta^{\phi} \\ &resolution: r = \gamma^{\phi} \\ &\text{s.t. } \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \\ &\alpha \geq 1, \beta \geq 1, \gamma \geq 1 \end{aligned} \tag{1}$$

Peneliti juga membentuk *baseline* yaitu EfficientNet-B0 yang terinspirasi oleh MBConv. Kemudian dari *baseline* ini, akan dilakukan pencarian *small grid search* untuk mencari nilai *depth, width, resolution*. Dari hasil pencarian ditemukan bahwa nilai masing-masing untuk *EfficientNet-B0* berurutan adalah 1,2; 1,1; 1,15 *Baseline* ini dapat dibesarkan menjadi B7 menggunakan *uniform scaling* pada Rumus 1.

3.4. VideoPose3D

VideoPose3D [19] merupakan model *temporal convolutional* yang digunakan untuk mengestimasikan koordinat 3D dari *n frame* koordinat 2D. metode ini menggunakan Convolutional Neural

Network (CNN) dikarenakan lebih bagus dalam hal waktu dan *batch parallelization*, sehingga gradien antara *input* dan *output* memiliki panjang yang sama sebesar apapun *sequence*-nya, yang dapat diatur *temporal receptive field* yang berguna untuk estimasi 3D, dan terakhir dapat diimplementasi *dilated convolutions* untuk *dependency* jangka panjang yang efisien.

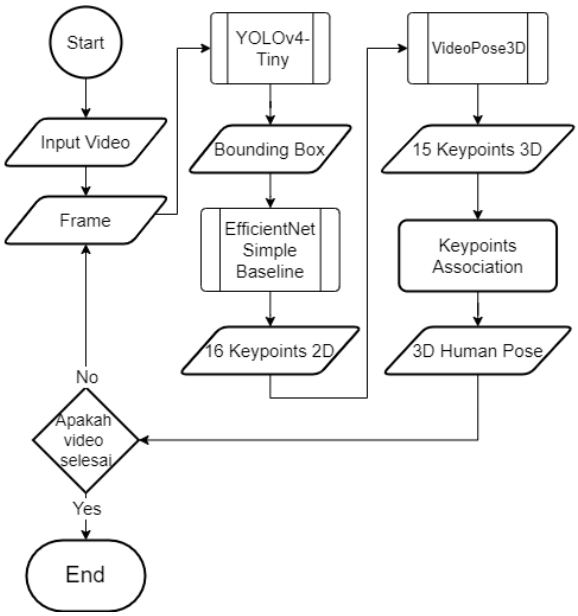
Metode ini dapat menggunakan 1 atau *n frame* untuk memprediksi koordinat 3D dalam *frame* yang diatur dalam *temporal receptive field* untuk mendapatkan hasil estimasi yang diperhitungkan dengan *frame-frame* sebelumnya.

3.5. Desain Sistem

Pada bagian ini akan membahas mengenai desain sistem mulai dari desain sistem, dataset yang digunakan, serta modifikasi *EfficientNet Simple Baseline*.

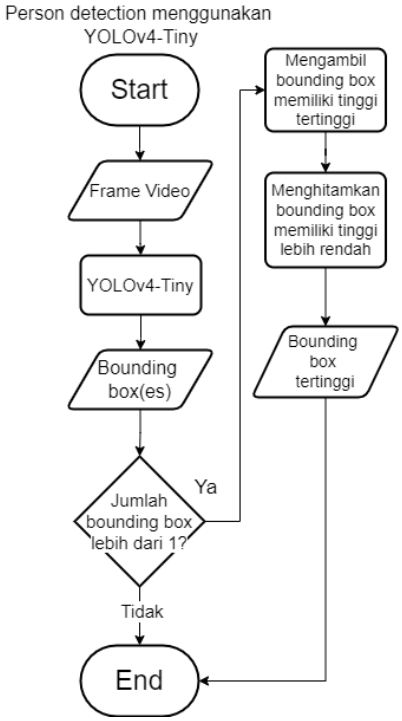
3.5.1. Desain Sistem

Sistem pertama-tama dimulai dengan dimasukan video Gerakan manusia. Video selanjutnya akan diproses setiap *frame* untuk dimasukkan pada *person detector* YOLOv4-Tiny. Setelah mendapatkan *bounding* box akan dimasukkan 2D *human pose estimation EfficientNet Simple Baseline.* Setelah menghasilkan 16 koordinat 2D *keypoints*, koordinat selanjutnya dimasukan pada 3D *human pose estimation VideoPose3D*. Akhirnya, akan didapatkan 15 3D koordinat setiap *keypoints* pada *frame* tersebut. Proses ini dapat dilihat pada Gambar 1.



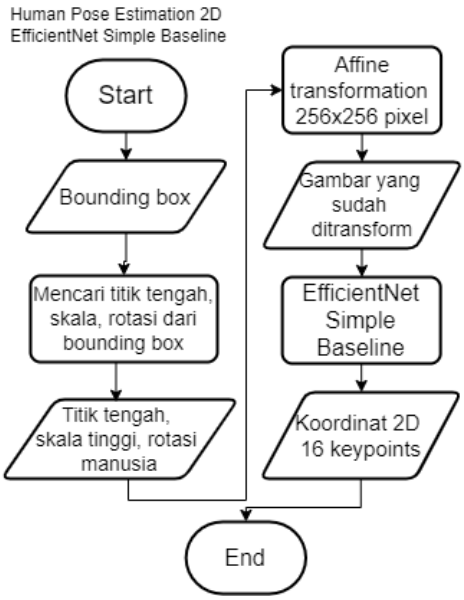
Gambar 1. Diagram sistem *human pose estimation*

Dalam proses *person detection* menggunakan *YOLOv4-Tiny* seperti pada Gambar 2, pertama-tama mendapatkan *frame video* dan dimasukkan ke dalam model *YOLOv4-Tiny* yang akan menghasilkan *bounding box. Bounding box* kemudian akan diseleksi. Apabila terdapat 2 atau lebih *bounding box*, akan diambil *bounding box* dengan tinggi tertinggi dan kemudian sisanya akan dihitamkan. Hasil dari model ini akan mengembalikan nilai koordinat *bounding box* manusia.



Gambar 2. Diagram *pre-defined process person detector* menggunakan *YOLOv4-Tiny*

Apabila didapatkan *bounding box* manusia, selanjutnya akan mencari titik tengah, tinggi skala, dan rotasi *bounding box*. Setelah didapatkan, akan dilakukan *affine transformation* dimana *frame* akan ditransformasikan dengan titik tengah gambar dari titik tengah *bounding box* sebesear 256x256 *pixel*. Nantinya, *frame* yang ditransformasikan akan dimasukkan ke dalam *model EfficientNet Simple Baseline* sehingga akan didapatkan 16 2D koordinat 2D. Proses ini dapat dilihat pada Gambar 3.

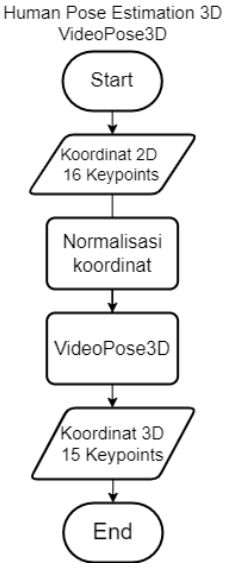


Gambar 3. Diagram *pre-defined process human pose estimation* 2D *EfficientNet Simple Baseline*

Setelah mendapatkan 16 2D koordinat *keypoints*, langkah selanjutnya adalah normalisasi koordinat tersebut sehingga proses

perhitungan dan prediksi koordinat akan lebih cepat. Setelah normalisasi, koordinat tersebut akan dimasukkan ke dalam model *VideoPose3D* untuk diprediksi 15 3D koordinat setiap *keypoint*. Proses ini dapat dilihat pada Gambar 4.

Keseluruhan proses akan dilakukan terus menerus hingga seluruh *frame* video diproses.



Gambar 4. Diagram *pre-defined human pose estimation* 3D *VideoPose3D*

3.5.2. Dataset

Dataset yang digunakan untuk penelitian adalah *dataset* COCO, MPII, dan HumanEva-I. Setiap *dataset* akan dijelaskan pada masing-masing sub-bab berikut.

3.5.2.1. Dataset COCO

Dataset COCO [15] merupakan *dataset* gambar-gambar objek pada dunia nyata yang ditangkap dan dianotasi. Anotasi dari gambar ini merupakan segmentasi atau koordinat *bounding box* masing-masing objek pada setiap gambar.

Penelitian ini akan menggunakan *dataset* COCO untuk *person detector*, sehingga dari *dataset* ini gambar-gambar yang memiliki orang akan disaring dan dimasukan ke dalam model *training*.

3.5.2.2. Dataset MPII

Dataset MPII [1] merupakan *dataset* yang berisikan gambar-gambar gerakan manusia dalam dunia nyata. Setiap manusia dalam suatu gambar memiliki anotasi koordinat kepala, titik tengah, skala, dan koordinat-koordinat 2D 16 sendi setiap orang pada masing-masing gambar.

Penelitian ini akan menggunakan *dataset* MPII untuk estimasi koordinat 2D dikarenakan MPII memiliki gambar orang dalam lingkungan yang beragam sehingga model *EfficientNet Simple Baseline* akan semakin *robust.*

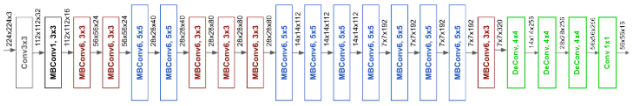
3.5.2.3. Dataset HumanEva-I

Dataset HumanEva-I [21] merupakan *dataset* video-video gerakan-gerakan manuusia yang ada pada dalam ruang. Video ini diambil dengan merekam orang menggunakan *motion capture* untuk diambil koordinat 3D. Video ini dianotasikan setiap *frame* pada 3 kamera yang memiliki sudut pandang berbeda 15 koordinat sendi 3D.

Penelitian ini menggunakan *dataset* HumanEva-I untuk melakukan estimasi koordinat 3D menggunakan model *VideoPose3D.*

3.5.3. Modifikasi EfficientNet Simple Baseline

Modifikasi *EfficientNet Simple Baseline* dilakukan dengan mengambil *EfficientNet* sebagai *backbone* kemudian pada akhir jaringan akan diimplementasikan *Simple Baseline* dari [27] sehingga akan menghasilkan *heatmap*. Dan dari *heatmap* tersebut diambil prediksi maksimal setiap masing-masing sendi menghasilkan koordinat 2D. Untuk arsitektur selengkapnya dapat melihat pada Gambar 5.



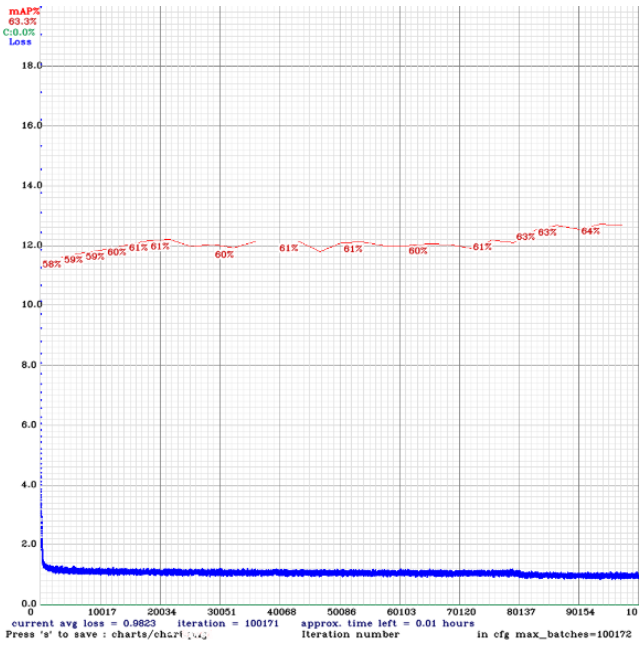
Gambar 5. Arsitektur modifikasi *EfficientNet Simple Baseline*

4. PENGUJIAN

Dalam bagian ini akan membahas mengenai pengujian tingkat akurasi terhadap model yang sudah dibuat. Pengujian dilakukan untuk masing-masing model *YOLOv4-Tiny*, modifikasi *EfficientNet Simple Baseline*, dan *VideoPose3D* masing-masing *dataset* validasi.

4.1. Pengujian YOLOv4-Tiny

Pengujian *person detector* untuk YOLOv4-Tiny dilakukan dengan metrik mAP (*mean average precision*). Dengan menggunakan mAP *model* akan semakin bagus apabila mAP semakin tinggi. Untuk hasil mAP didapatkan setiap n iterasi *training* dimana model akan melakukan evaluasi dengan *dataset* validasi dan didapatkan hasil mAP. Hasil mAP tertinggi dari model ini adalah 63.3% seperti ditunjukan Gambar 6.



Gambar 6. Grafik mAP YOLOv4-Tiny

4.2. Pengujian Simple Baseline

Dalam sub-bab ini akan menjelaskan dan membandingkan pengujian *model* Simple Baseline dengan *model* modifikasi EfficientNet Simple Baseline B0 dan B6. Pengujian *model* ini dilakukan dengan menggunakan metrik pengujian PCKh@0.5 yang dijelaskan sebelumnya, kecepatan *training*,dan kecepatan model memproses.

Untuk pengujian model dengan metrik PCKh@0.5 dilakukan dengan cara melakukan evaluasi dengan dataset validasi setiap kali epoch selesai melakukan training. Pengujian kecepatan dilakukan dengan menghitung kecepatan *weights* memproses suatu *frame*. Adapun semua hasil yang diperoleh sebagai berikut.

Tabel 1. Hasil validasi PCKh@0.5 setiap keypoint diprediksi, waktu training, dan waktu proses.

Keterangan	Nilai		
Backbone	ResNet-152	EfficientNet-B0	EfficientNet-B5
Head	97.03	95.71	96.28
Shoulder	95.94	94.24	94.72
Elbow	90.05	85.82	86.84
Wrist	84.98	79.13	80.53
Hip	89.16	85.53	86.64
Knee	85.31	80.27	81.56
Ankle	81.27	76.07	77.23
Mean	89.62	85.93	86.89
Mean@0.1	35.03	29.14	30.35
Waktu train	29.17 Jam	25.6 Jam	18.6 Jam
Waktu proses	5.15ms	4.54ms	15.06ms

Berdasarkan Tabel 1, *Simple Baseline* original memiliki akurasi yang lebih tinggi dibandingkan modifikasi *EfficientNet Simple Baseline*. Akan tetapi jika diukur dalam segi kecepatan training dan proses, modifikasi EfficientNet akan lebih unggul dibandingkan originalnya.

Selain itu, dalam pengujian ini juga dilakukan perbandingan antara *EfficientNet* B0 dan B5 untuk mengetahui apakah semakin baru versi model *EfficientNet* apakah akan semakin bagus juga hasilnya. Dari pengujian tersebut B5 lebih unggul dalam segi kecepatan *training* dan lebih sedikit dalam akurasi, akan tetapi dari segi pemrosesan B0 lebih unggul.

4.3. Pengujian dengan VideoPose3D

Dalam sub-bab ini akan membandingkan MPJPE *VideoPose3D* dengan *ground truth* koordinat 2D *HumanEva-I*, original *Simple Baseline*, dan modifikasi *EfficientNet Simple Baseline*. Pengukuran MPJPE dilakukan dengan menghitung selisih hasil prediksi koordinat 3D *VideoPose3D* dengan *ground truth* koordinat 3D *HumanEva-I*. Untuk hasil dari masing-masing prediksi dapat dilihat pada Tabel 2, Tabel 3, dan Tabel 4.

Tabel 2. Pengujian *model* VideoPose3D MPJPE dengan *ground truth* HumanEva-I 2D

Nama Subjek	Box	Jog	Walking	Average
S1	22.74 mm	21.29 mm	13.77 mm	19.3 mm
S2	29.39 mm	14.58 mm	10.32 mm	18.1 mm
S3	32.31 mm	16.68 mm	-	24.5 mm
Total Average				20.63 mm

Tabel 3. Pengujian model VideoPose3D MPJPE dengan prediksi koordinat 2D HumanEva-I oleh Simple Baseline

Nama Subjek	Box	Jog	Walking	Average
S1	27.60 mm	26.81 mm	18.97 mm	24.5 mm
S2	32.69 mm	20.06 mm	14.31 mm	22.4 mm
S3	37.61 mm	20.35 mm	-	29.0 mm
Total Average				25.3 mm

Tabel 4. Pengujian model VideoPose3D MPJPE dengan prediksi koordinat 2D HumanEva-I oleh EfficientNet Simple Baseline

Nama Subjek	Box	Jog	Walking	Average
S1	31.72 mm	28.81 mm	21.29 mm	27.3 mm
S2	36.84 mm	22.49 mm	17.71 mm	25.7 mm
S3	40.09 mm	22.55 mm	-	31.3 mm
Total Average				28.1 mm

Berdasarkan tabel-tabel diatas, dapat disimpulkan bahwa nilai MPJPE dipengaruhi oleh nilai PCKh@0.2 sebelumnya. Semakin besar nilai PCKh@0.2 maka akan semakin bagus nilai MPJPE. Dari tabel-tabel diatas, MPJPE original *Simple Baseline* paling kecil sehingga paling bagus.

5. KESIMPULAN

Pada penelitian ini, dapat disimpulkan beberapa hal:

- Penggunaan YOLOv4-Tiny sudah cukup bagus sebagai *person detector*. Akan tetapi hal ini bisa ditingkatkan lebih lagi dengan menggunakan *person detector* yang khusus untuk benar-benar untuk manusia. YOLOv4-Tiny merupakan *object detector*, sehingga dapat ditingkatkan kembali dengan menggunakan *person detector* yang khusus.
- Modifikasi EfficientNet Simple Baseline dalam segi keakuratan masih kurang bagus dibandingkan ResNet-152

Simple Baseline. Dimana Simple Baseline memiliki rata-rata PCKh@0.5 sebesar 89.62% dan EfficientNet Simple Baseline memiliki rata-rata PCKh@0.5 tertinggi sebesar 86.89%. Akan tetapi jika diukur dalam segi kecepatan, maka Modifikasi EfficientNet Simple Baseline akan lebih bagus dimana memiliki kecepatan dengan waktu 4.54ms sedangkan Simple Baseline memiliki kecepatan dengan waktu 5,15ms. Selain itu, EfficientNet versi B0 paling bagus dipakai daripada versi terbarunya, dikarenakan diukur dari waktu proses, *parameter* B0 lebih kecil, serta selisih akurasi yang tidak signifikan.

- Akurasi estimasi 3D dipengaruhi oleh hasil estimasi 2D dengan hasil VideoPose3D dengan Simple Baseline menghasilkan rata-rata MPJPE sebesar 25.3 mm sedangkan EfficientNet Simple Baseline menghasilkan rata-rata MPJPE sebesar 28.1 mm.

6. SARAN

Untuk pengembangan topik penelitian lebih lanjut, terdapat beberapa saran yang dapat dipertimbangkan:

- Menggunakan *dataset* yang memiliki lebih banyak aneka ragam gerakan untuk *human pose estimation* 2D sehingga dapat mengestimasikan gerakan lebih beragam.
- Menggunakan *dataset* yang memiliki lebih banyak data untuk *human pose estimation* 3D agar MPJPE dapat lebih kecil. Salah satu contoh *dataset* ini yang cukup terkenal adalah *Human 3.6m.*

7. REFERENSI

[1] Andriluka, M., Pishchulin, L., Gehler, P., & Schiele, B. (2014). 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition* (pp. 3686-3693). DOI: 10.1109/CVPR.2014.471

[2] Artacho, B., & Savakis, A. (2020). Unipose: Unified human pose estimation in single images and videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7035-7044). DOI: 10.1109/CVPR42600.2020.00706

[3] Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., & Grundmann, M. (2020). Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*.

[4] Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., & Sun, J. (2018). Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7103-7112). DOI: 10.1016/j.cviu.2019.102897

[5] Debnath, B., O'Brien, M., Yamaguchi, M., & Behera, A. (2018, November). Adapting MobileNets for mobile based upper body pose estimation. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 1-6). IEEE. DOI: 10.1109/AVSS.2018.8639378

[6] Eichner, M., Marin-Jimenez, M., Zisserman, A., & Ferrari, V. (2012). 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. *International journal of computer vision*, *99*(2), 190-214. DOI: 10.1007/s11263-012-0524-9

[7] Eichner, M., & Ferrari, V. (2010, September). We are family: Joint pose estimation of multiple persons. In *European conference on computer vision* (pp. 228-242). Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-642-15549-9_17

[8] Fan, X., Zheng, K., Lin, Y., & Wang, S. (2015). Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1347-1355). DOI: 10.1109/CVPR.2015.7298740.

[9] Groos, D., Ramampiaro, H., & Ihlen, E. A. (2021). EfficientPose: Scalable single-person pose estimation. *Applied intelligence*, *51*(4), 2518-2533. DOI: 10.1007/s10489-020-01918-7

[10] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778). DOI: 10.1109/CVPR.2016.90

[11] Insafutdinov, E., Andriluka, M., Pishchulin, L., Tang, S., Levinkov, E., Andres, B., & Schiele, B. (2017). Arttrack: Articulated multi-person tracking in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6457-6465). DOI: 10.1109/CVPR.2017.142

[12] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*. DOI: 10.1145/3065386

[13] Li, B., Dai, Y., Cheng, X., Chen, H., Lin, Y., & He, M. (2017, July). Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)* (pp. 601-604). IEEE. DOI: 10.1109/ICMEW.2017.8026282

[14] Li, B., He, M., Dai, Y., Cheng, X., & Chen, Y. (2018). 3D skeleton based action recognition by video-domain translation-scale invariant mapping and multi-scale dilated CNN. *Multimedia Tools and Applications*, *77*(17), 22901-22921. DOI: 10.1007/s11042-018-5642-0

[15] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham. DOI: 10.1007/978-3-319-10602-1_48

[16] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440). DOI: 10.1109/CVPR.2015.7298965

[17] Luvizon, D. C., Picard, D., & Tabia, H. (2018). 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5137-5146). DOI: 10.1109/CVPR.2018.00539

[18] Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H. P., ... & Theobalt, C. (2017). Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, *36*(4), 1-14. DOI: 10.1145/3072959.3073596

[19] Pavllo, D., Feichtenhofer, C., Grangier, D., & Auli, M. (2019). 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7753-7762). DOI: 10.1109/CVPR.2019.00794

[20] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, *28*. DOI: 10.1109/TPAMI.2016.2577031

[21] Sigal, L., Balan, A. O., & Black, M. J. (2010). Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, *87*(1), 4-27. DOI: 10.1007/s11263-009-0273-6

[22] Sun, M., & Savarese, S. (2011, November). Articulated part-based model for joint object detection and pose estimation. In *2011 International Conference on Computer Vision* (pp. 723-730). IEEE. DOI: 10.1109/ICCV.2011.6126309

[23] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105-6114). PMLR 97:6105-6114.

[24] Wang, F., & Li, Y. (2013). Beyond physical connections: Tree models in human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 596-603). DOI: 10.1109/CVPR.2013.83

[25] Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2021). Scaled-yolov4: Scaling cross stage partial network. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition* (pp. 13029-13038). DOI: 10.1109/CVPR46437.2021.01283

[26] Wang, M., Chen, X., Liu, W., Qian, C., Lin, L., & Ma, L. (2018). Drpose3d: Depth ranking in 3d human pose estimation. *arXiv preprint arXiv:1805.08973*.

[27] Xiao, B., Wu, H., & Wei, Y. (2018). Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 466-481). DOI: 10.1007/978-3-030-01231-1_29

[28] Yang, Y., & Ramanan, D. (2011, June). Articulated pose estimation with flexible mixtures-of-parts. In *CVPR 2011* (pp. 1385-1392). IEEE. DOI: 10.1109/CVPR.2011.5995741

[29] Zhang, H., Ouyang, H., Liu, S., Qi, X., Shen, X., Yang, R., & Jia, J. (2019). Human pose estimation with spatial contextual information. *arXiv preprint arXiv:1901.01760*.

[30] Zheng, C., Wu, W., Yang, T., Zhu, S., Chen, C., Liu, R., ... & Shah, M. (2020). Deep learning-based human pose estimation: A survey. *arXiv preprint arXiv:2012.13392*.

[31] Zhou, K., Han, X., Jiang, N., Jia, K., & Lu, J. (2019). Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 2344-2353). DOI: 10.1109/ICCV.2019.00243

[32] Zimmermann, C., Welschehold, T., Dornhege, C., Burgard, W., & Brox, T. (2018, May). 3d human pose estimation in rgbd images for robotic task learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 1986-1992). IEEE. DOI: 10.1109/ICRA.2018.84628338