Fast inference methods in Whole-Body Pose Estimation

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

Presented By
김태우

Outline

- Background

- **One-Stage 3D Whole-Body Mesh Recovery with Component Aware Transformer (CVPR 2023)**

- Background

- **Effective Whole-body Pose Estimation with Two-stages Distillation (ICCVW 2023)**

Background

- SMPL-X
  - 사람을 3차원으로 표현하기 위한 표준 모델 중 하나로, 몸만 표현하는 SMPL 모델에 더해 손과 얼굴의 표정까지 표현

그림. 사람의 손과 표정까지 표현하는 SMPL-X

1) Rong, Yu, Takaaki Shiratori, and Hanbyul Joo. "Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration." *ICCV* (2021).

Background

- Whole-body-estimation
  - 신체(body) 뿐만 아니라, 손, 얼굴을 포함한 전신에 대한 3d pose를 예측

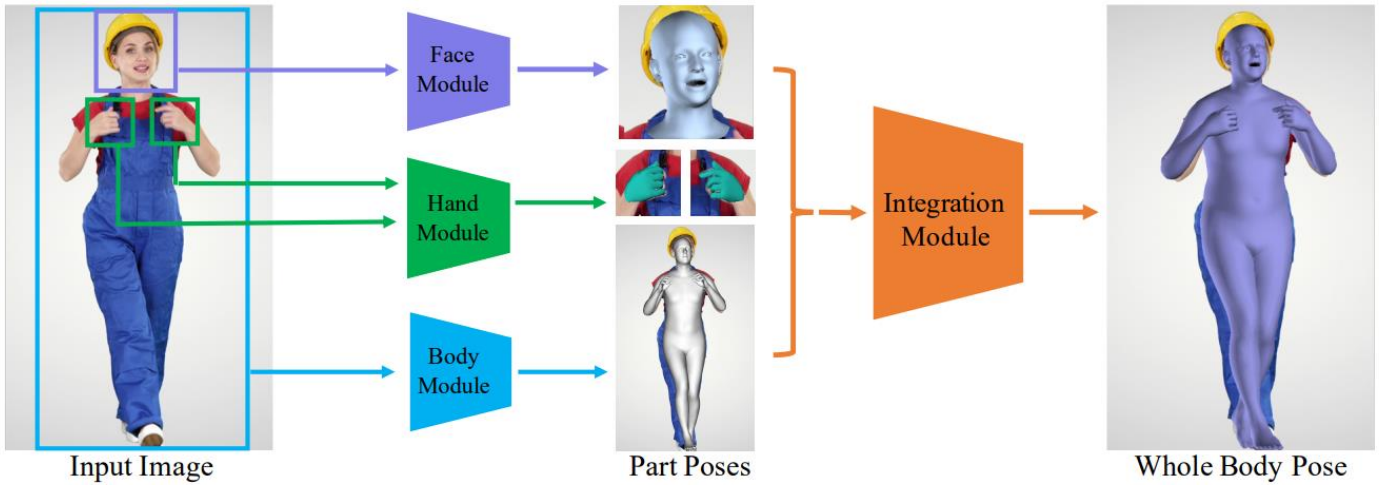


그림. Frankmocap[1] 모듈: 기존의 Whole-Body-Estimation 방법

1) Rong, Yu, Takaaki Shiratori, and Hanbyul Joo. "Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration." *ICCV* (2021).

Background

- Whole-Body-Estimation
  - 한계점: 신체(body) 와 달리, 손과 얼굴은 전체 이미지에서 차지하는 영역이 작음
    - 해상도가 작아져 예측하기 어려움

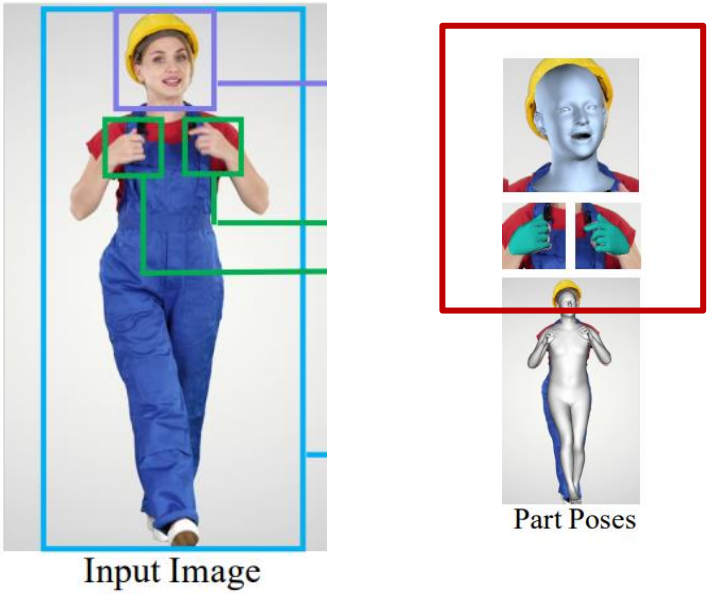


그림. Frankmocap[1] 모듈: 기존의 Whole-Body-Estimation 방법

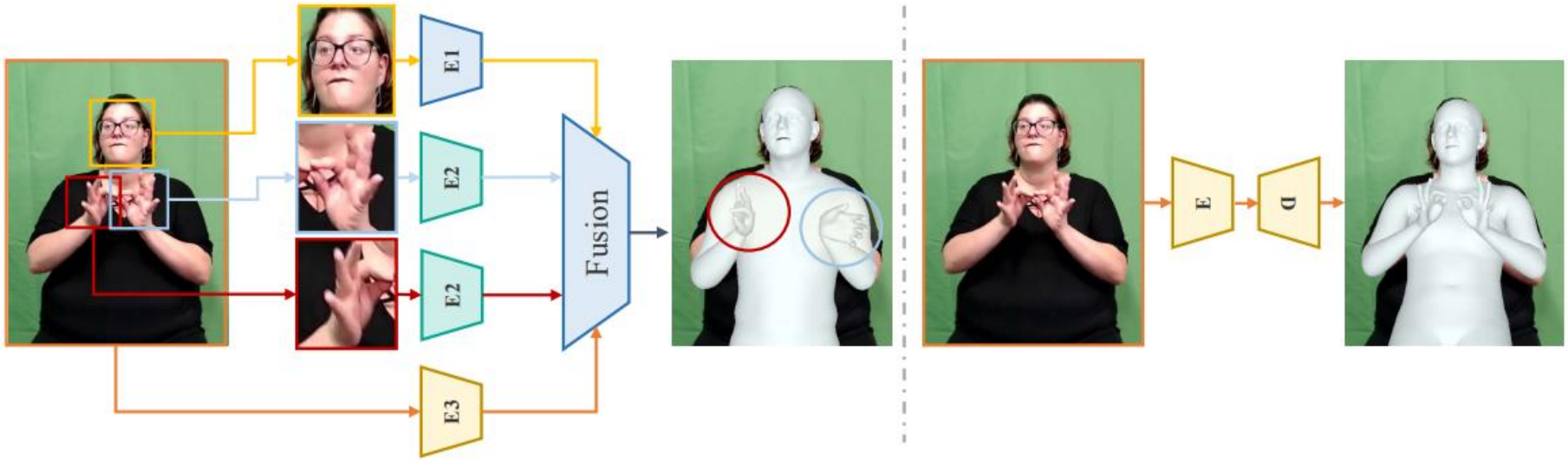
OSX: One-Stage 3D Whole-Body Mesh Recovery with Component Aware Transformer (CVPR 2023)

Abstract

- 인코더 디코더 형태 기반의 OSX
  - 기존 : Detection 과정 후 각각의 전용 모델에 넣은 multi-stage 구조

    - One-stage 인코더 디코더 형태를 새롭게 제안

    - 손과 신체 토큰 사이의 attention 을 고려하여 손목이 과도하게 접히는 현상 (불가능한 각도) 이 적음

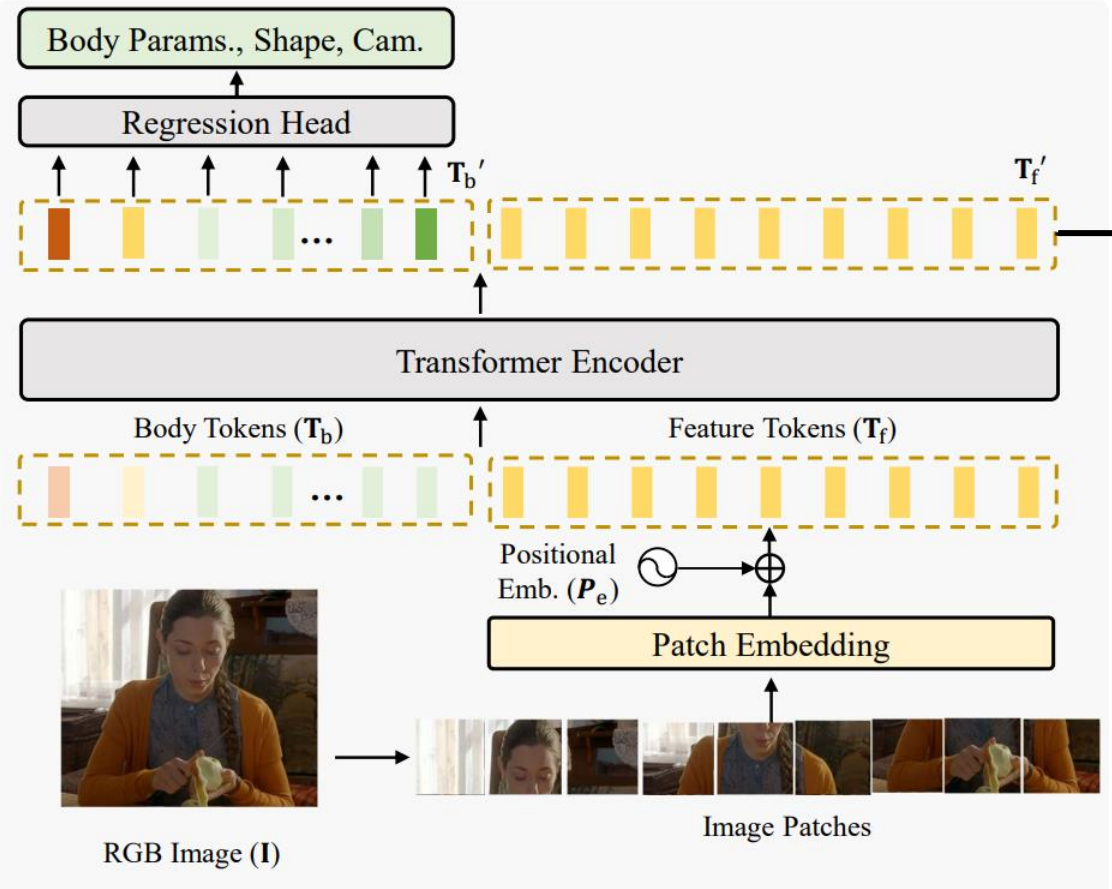


(a) Previous multi-stage pipeline

(b) Our one-stage pipeline

Proposed method

- ViT 구조의 인코더



Decoder

- Vit 구조 사용
- Body token 랜덤 초기화

그림. One-stage mesh 생성 인코더 구조

Proposed method

- Feature 토큰에 대해서 bbox 예측을 사용

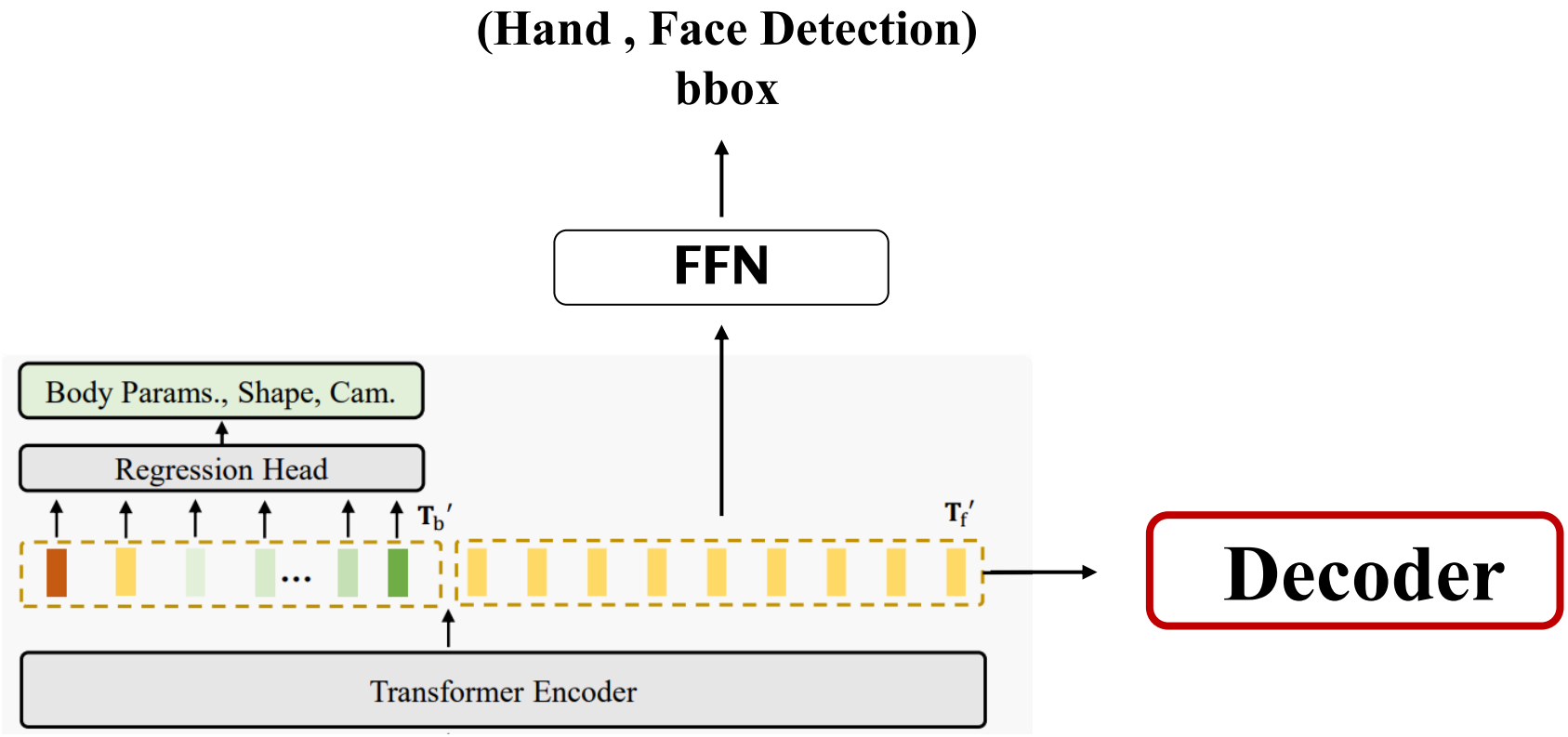


그림. Bbox 예측을 통해, 위치 정보에 관하여 token에 반영되도록 시도

Proposed method

- Hand 와 face 파라미터를 추출하는 디코더

- 정확도 개선
  - 입력 토큰 변경
    - (Face , hand 의 2d keypoint)
  - 다양한 크기의 feature map 을 사용
    - Feature token 을 업샘플링

- 빠른 추론 속도 측면
  - ROI align 으로 작은 feature map 을 생성

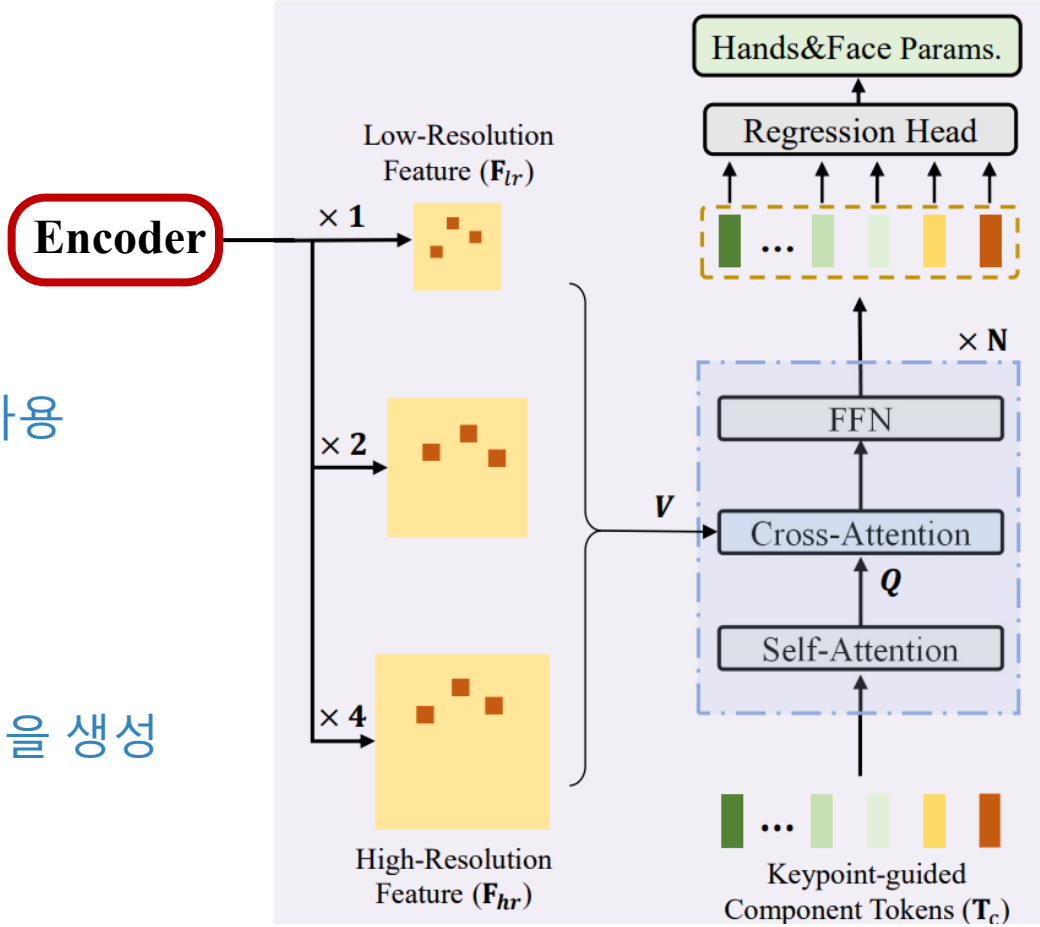


그림. 디코더 구조

Proposed method

- Loss 함수 정의

$$L = L_{smplx} + L_{kpt3D} + L_{kpt2D} + L_{bbox2D}.$$

① ② ③ ④

- SMPL-x 의 파라미터에 대한 L1 distance
  - SMPL-x 의 파라미터 : shape , 관절의 각도, 등
- Predicted 3D pose 와 GT 와의 L1 distance
- Predicted 2D pose 와 GT 사이의 L1 distance
- Predicted face , hands bbox 와 GT 사이의 L1 distance

Experiment result

- Comparison with SOTA models
  - 추론 속도 측면
    - 실험 기준 : A100 GPU , 동일 해상도 ,이미지에 한 사람만 존재,
    - AGORA – test 데이터 세트

NMJE-All (Normalized Mean Joint Error - All)

Method	ExPose [34]	PIXIE [13]	H4W [27]	PyMAF-X [48]	OSX
NMJE-All (mm)	263.3	230.9	141.1	140.0	127.6
Infer Time (ms)	120.2	192.0	73.3	209.3	54.6
Params (M)	135.8	192.9	77.9	205.9	102.9
FLOPS (G)	28.5	34.3	16.7	35.5	25.3

표. OSX 가 다른 multi-stage 모델보다 빠른 추론 속도를 보여주는 모습

Experiment result

- Comparison with SOTA models
  - 정확성 측면
    - MPVPE 지표 기준 9.5 퍼센트 낮은 에러를 보임 (H4W 기준)
    - AGORA – test 데이터 세트

Method	AGORA-test				
	MPVPE ↓			N-MPVPE ↓	
	All	Hands	Face	All	Body
ExPose [48]	217.3	73.1	51.1	265.0	184.8
FrankMocap [56]	-	55.2	-	-	207.8
PIXIE [16]	191.8	49.3	50.2	233.9	173.4
Hand4Whole [39]	-	-	-	-	-
Hand4Whole [39]×	135.5	47.2	41.6	144.1	96.0
OSX (Ours)	**122.8**↓9.5%	**45.7**	**36.2**	**130.6**	**85.3**

표. AGORA 데이터에 따른 3d mesh error 비교
× : 손과 얼굴에 대해 추가적인 학습 데이터 사용

Experiment result

- Comparison with SOTA models
  - 정확성 측면
    - MPVPE 지표 기준 7.8 퍼센트 낮은 에러를 보임 (H4W 기준)
  - 손과 얼굴의 상대적 높은 에러
    - Whole body 에 대해서는 SOTA를 달성
    - 손과 얼굴에 대해서 상대적 높은 에러 값 => OSX 가 해상도 문제를 완전히 해결하지 못한 것으로 해석

Method	EHF					
	MPVPE ↓			PA-MPVPE ↓		
	All	Hands	Face	All	Hands	Face
ExPose [48]	77.1	51.6	35.0	54.5	12.8	5.8
FrankMocap [56]	107.6	42.8	-	57.5	12.6	-
PIXIE [16]	89.2	42.8	32.7	55.0	11.1	**4.6**
Hand4Whole [39]	79.2	43.2	**25.0**	53.1	12.1	5.8
Hand4Whole [39]×	76.8	**39.8**	26.1	**50.3**	**10.8**	5.8
OSX (Ours)	**70.8**↓7.8%	53.7	26.4	48.7	15.9	6.0

표. EHF 데이터에 따른 3d mesh error 비교
× : 손과 얼굴에 대해 추가적인 학습 데이터 사용

Experiment result

- AGORA 데이터 세트에 대한 정성적 평가

그림. 여러 모델에 대한 정성적 평가 결과

Effective Whole-body Pose Estimation with Two-stages Distillation

(ICCVW 2023)

Background

- SimCC[1] 논문 기반 pose estimation 수행
  - Pose estimation 문제를 heat map 문제로 보지 않고, class 분류 문제로 이해

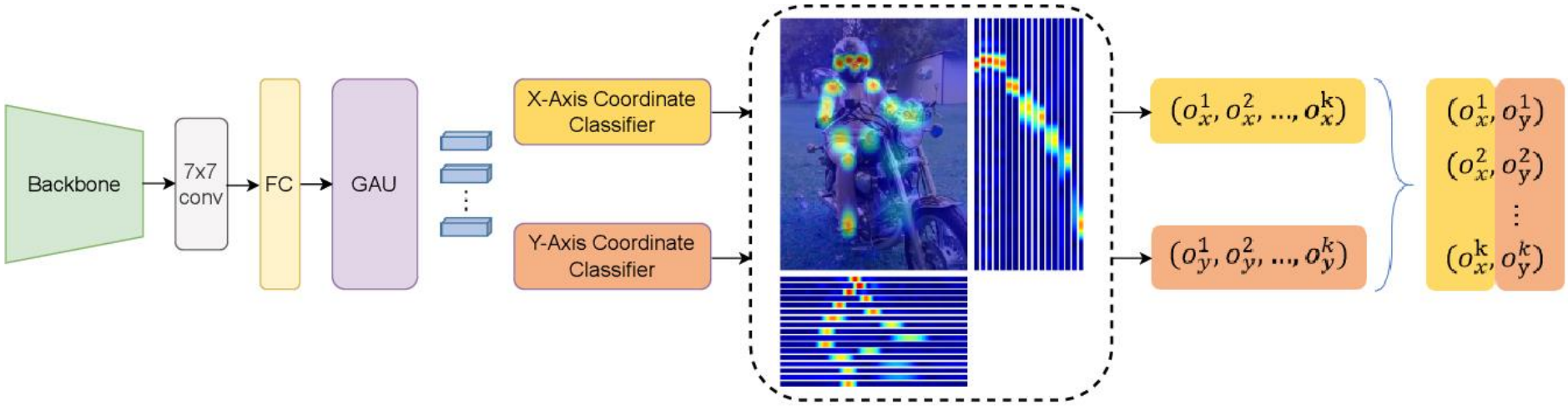


그림. SIMCC 모델을 채택한 모델 구조

Background

- Pose estimation 에 대한 Class 분류 문제 접근법
  - 사람이 있는 이미지를 일정한 크기 Bin 으로 x, y에 대해 나눔
    - 특정 관절이 x-Axis 에 대해서 어디에 존재할지 클래스 분류 (확률)
    - 뽑은 Feature 를 FFN 에 넣어서 확률을 계산
    - x-Axis 와 y-Axis 에 대해 각각 수행

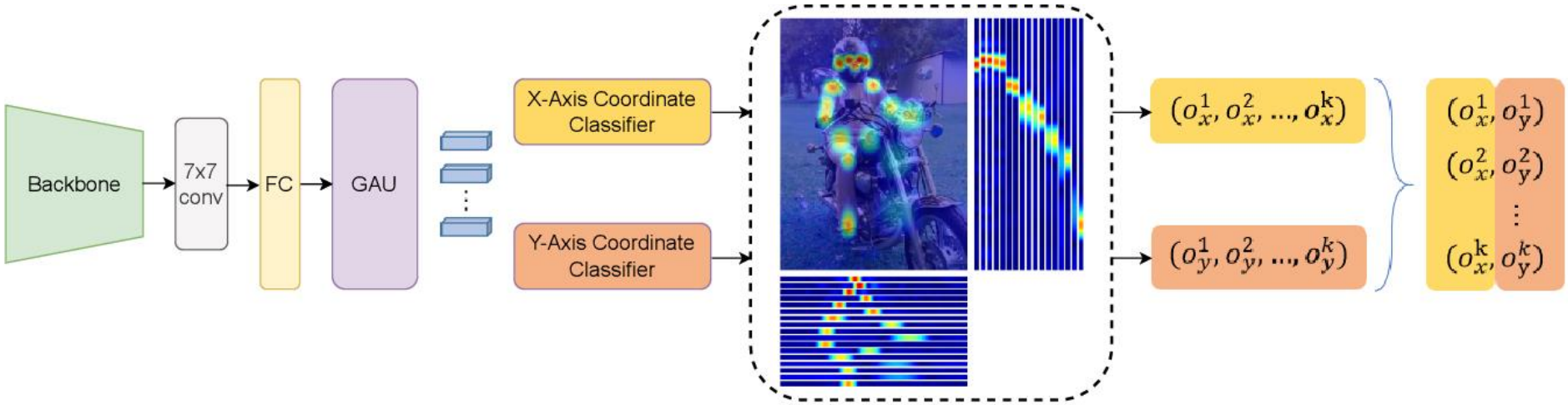


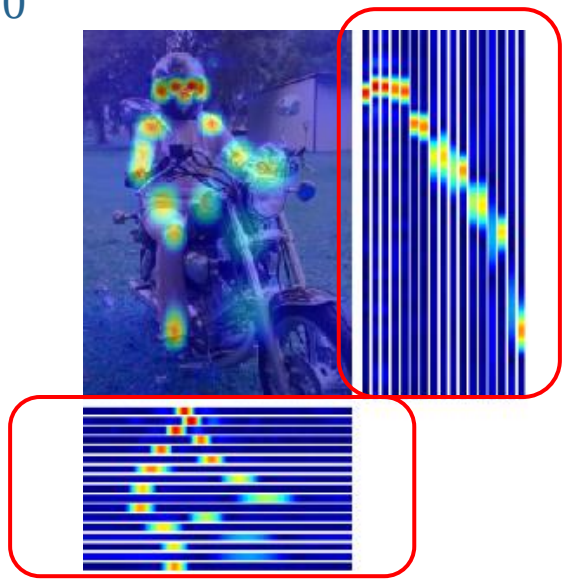
그림. SIMCC 모델을 채택한 SOTA RTM-Pose 모델 구조

Background

- Class 분류 문제에 대한 loss
  - 쿨백-라이블러 발산 (KL-divergence) 을 사용

$$D_{KL}(P\|Q) = \sum_{x \in \chi} P(x) \log_b \left(\frac{P(x)}{Q(x)} \right)$$

  - 두 이산 확률 분포 P,Q 의 유사도를 비교하는 방법으로, P와 Q가 완전히 동일하면 0



KL-divergence Loss ⟷ GT

그림. 관절 위치 가능성에 대한 X축,Y축 확률 분포 시각화

Abstract

- DWPose
  - 2D Whole-Body-Estimation 을 위한 새로운 두 단계 Knowledge-distillation 기법을 소개
  - Teacher Model , 2D Whole-Body-Estimation 의 SOTA 인 RTM-Pose 보다 높은 AP를 달성 (RTM-Pose 65.3% AP , DWPose 66.5% AP)
    - COCO-WholeBody 데이터 세트 기준

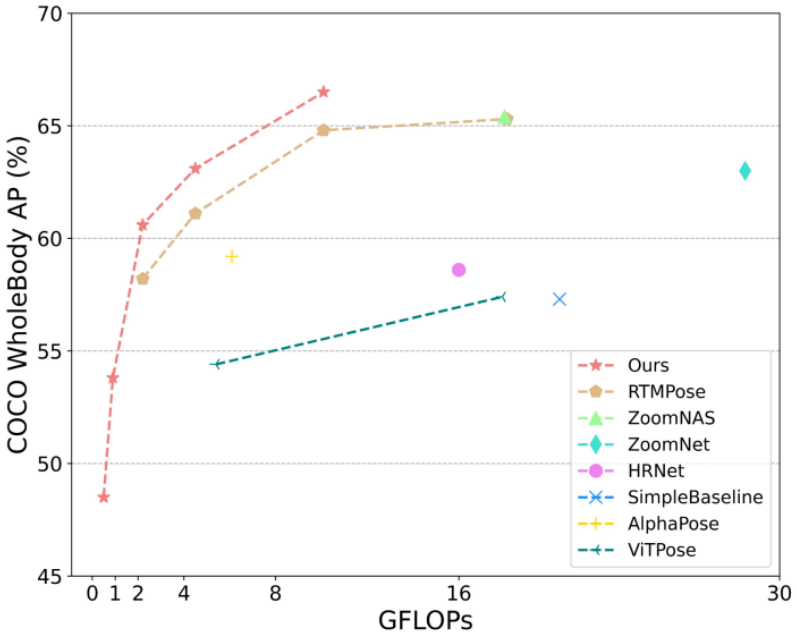
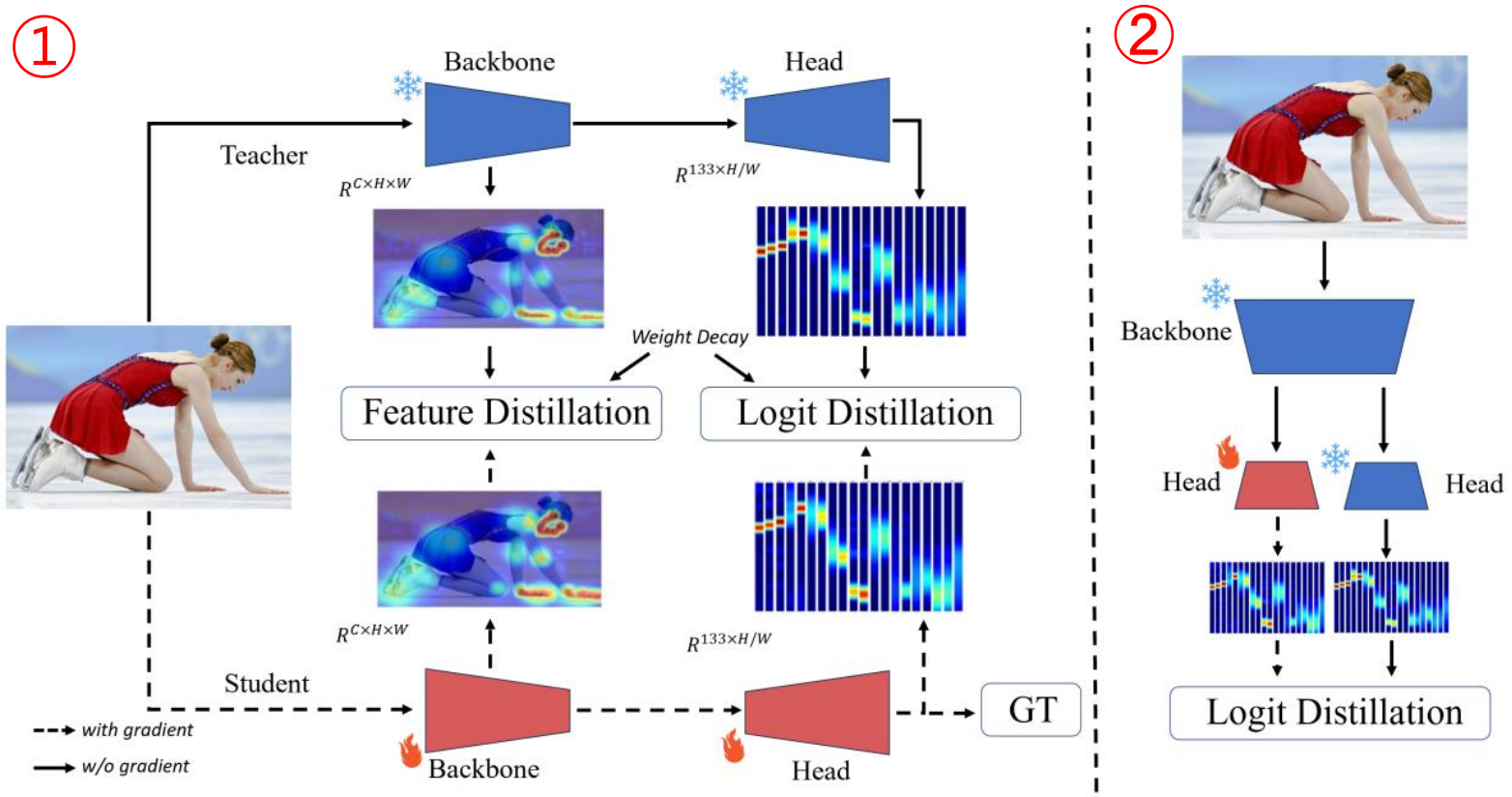


그림. GFLOPS 에 따른 모델 AP 비교

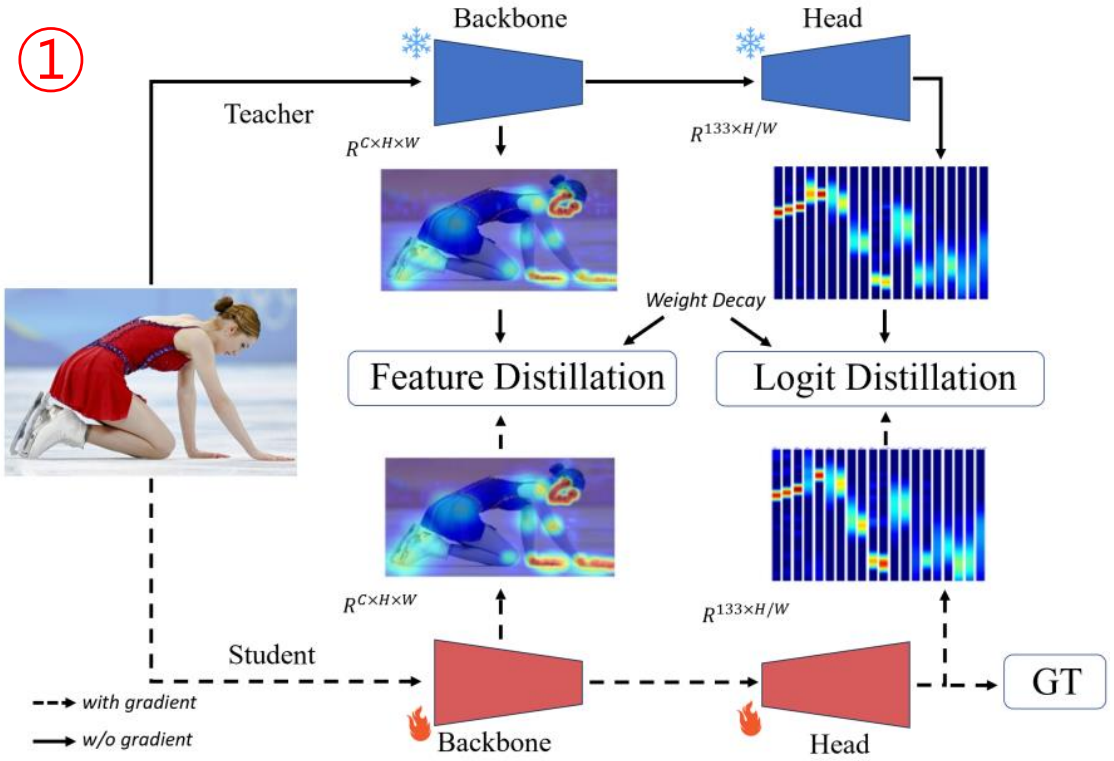
Proposed Method

- Two-Stages Pose Distillation
  - 두 단계를 거치는 Distillation 기법으로 학습이 끝난 학생 모델에도 사용 가능



Proposed Method

- First-Stage (전통적인 방법)
  - Teacher 백본 Feature 와 Student 백본 추출 Feature 의 차이를 줄이는 방법
  - Teacher Head 의 Logit 과 Student Head 의 Logit 분포의 차이를 줄이는 방법



Proposed Method

- First-stage loss
  - Feature-based distillation

$$L_{fea} = \frac{1}{CHW} \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{w=1}^{W} \left(F_{c,h,w}^{t} - f(F_{c,h,w}^{s}) \right)^2$$

수식. 피처맵에 대한 MSE

  - Logit-based distillation

$$L_{logit} = -\frac{1}{N} \cdot \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{i=1}^{L} T_i log(S_i).$$

수식. Logit 에 대한 KL-divergence Loss
N : 배치에 있는 사람의 수
K : 키포인트 넘버
L : X,Y 에 대해 나눈 Bin

$$D_{KL}(P \| Q) = \sum_{x \in \chi} P(x) \log_b \left(\frac{P(x)}{Q(x)} \right)$$

수식. KL-divergence

Proposed Method

- First-stage loss

$$L = L_{ori} + \alpha L_{fea} + \beta L_{logit}$$
$$\{\alpha = 0.00005, \beta = 0.1\}$$

  - Logit-based distillation loss (RTM-Pose uses)

$$L_{ori} = -\sum_{n=1}^{N}\sum_{k=1}^{K} W_{n,k} \cdot \sum_{i=1}^{L} \frac{1}{L} \cdot V_i log(S_i),$$

수식. RTM-Pose 의 Logit loss

- N : 배치에 있는 사람의 수
- K : 키포인트 넘버
- L : X,Y 에 대해 나눈 Bin
- W : 키포인트 보이는 유무 (0 , 1)
- V : GT 값 분포 Value

Proposed Method

- First-stage loss
  - Weight-decay strategy for distillation

$$r(t) = 1 - (t-1)/t_{max}$$

  - t_{max} : 전체 에포크
  - t : 현재 에포크
  - 0.3% AP 성능 향상

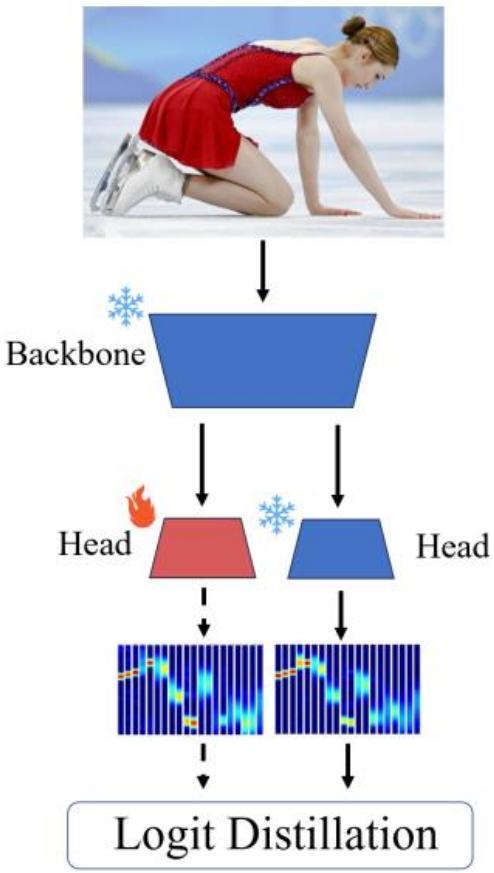
  - 최종 First-stage loss

$$L_{s1} = L_{ori} + r(t) \cdot \alpha L_{fea} + r(t) \cdot \beta L_{logit}$$

Proposed Method

- Second-stage
  - 훈련을 마친 student 를 사용, 추가적인 knowledge distillation 수행
    - 훈련을 마친 student 백본의 피처를 뽑음
    - Student head 와 teacher head 를 각각 사용하여 pose 예측
    - 예측한 확률 분포 map 에 대해서 logit distillation loss

②



  - Second stage loss

$$L_{logit} = -\frac{1}{N} \cdot \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{i=1}^{L} T_i log(S_i)$$

$$L_{s2} = \gamma L_{logit}.$$

수식. Second-stage loss

Experimental results

- RTM-Pose (SOTA, Teacher) 와 성능 비교

	Method	Input Size	GFLOPs	whole-body		body		foot		face		hand	
				AP	AR	AP	AR	AP	AR	AP	AR	AP	AR
Whole-body	SN† [13]	N/A	N/A	32.7	45.6	42.7	58.3	9.9	36.9	64.9	69.7	40.8	58.0
	OpenPose [2]	N/A	N/A	44.2	52.3	56.3	61.2	53.2	64.5	76.5	84.0	38.6	43.3
Bottom-up	PAF† [3]	512×512	329.1	29.5	40.5	38.1	52.6	5.3	27.8	65.6	70.1	35.9	52.8
	AE [34]	512×512	212.4	44.0	54.5	58.0	66.1	57.7	72.5	58.8	65.4	48.1	57.4
Top-down	DeepPose [43]	384×288	17.3	33.5	48.4	44.4	56.8	36.8	53.7	49.3	66.3	23.5	41.0
	SimpleBaseline [47]	384×288	20.4	57.3	67.1	66.6	74.7	63.5	76.3	73.2	81.2	53.7	64.7
	HRNet [40]	384×288	16.0	58.6	67.4	70.1	77.3	58.6	69.2	72.7	78.3	51.6	60.4
	PVT [44]	384×288	19.7	58.9	68.9	67.3	76.1	66.0	79.4	74.5	82.2	54.5	65.4
	FastPose50-dcn-si [9]	256×192	6.1	59.2	66.5	70.6	75.6	70.2	77.5	77.5	82.5	45.7	53.9
	ZoomNet [19]	384×288	28.5	63.0	74.2	74.5	81.0	60.9	70.8	88.0	92.4	57.9	73.4
	ZoomNAS [48]	384×288	18.0	65.4	74.4	74.0	80.7	61.7	71.8	88.9	93.0	62.5	74.0
	ViTPose+-S [51]	256×192	5.4	54.4	-	71.6	-	72.1	-	55.9	-	45.3	-
	ViTPose+-H [51]	256×192	122.9	61.2	-	75.9	-	77.9	-	63.3	-	54.7	-
	RTMPose-m	256×192	2.2	58.2	67.4	67.3	75.0	61.5	75.2	81.3	87.1	47.5	58.9
	RTMPose-l	256×192	4.5	61.1	70.0	69.5	76.9	65.8	78.5	83.3	88.7	51.9	62.8
	RTMPose-l	384×288	10.1	64.8	73.0	71.2	78.1	69.3	81.1	88.2	91.9	57.9	67.7
	RTMPose-x	384×288	18.1	65.3	73.3	71.4	78.4	69.2	81.0	88.8	92.2	59.0	68.5
	RTMPose-l + UBody	256×192	4.5	62.1	70.6	69.7	76.9	65.5	78.1	84.1	89.3	55.1	65.4
	RTMPose-l + UBody	384×288	10.1	65.4	73.2	71.0	77.9	68.6	80.2	88.5	92.2	60.6	69.9
	DWPose-t	256×192	0.5	48.5	58.4	58.5	67.0	46.5	63.6	73.5	80.7	35.7	49.0
	DWPose-s	256×192	0.9	53.8	63.2	63.3	71.3	53.3	69.0	77.6	84.1	42.7	54.9
	DWPose-m	256×192	2.2	60.6	69.5	68.5	76.1	63.6	77.2	82.8	88.1	52.7	63.4
	DWPose-l	256×192	4.5	63.1	71.7	70.4	77.7	66.2	79.0	84.3	89.4	56.6	66.5
	DWPose-l	384×288	10.1	66.5	74.3	72.2	78.9	70.4	81.7	88.7	92.1	62.1	71.0

Conclusion

- Teacher model (RTM-Pose) 보다 높은 성능 달성
  - 연산량 GFLOPS 18.1 기준 - RTM-Pose 65.3% AP
  - 연산량 GFLOPS 10.1 기준 - DWPose 66.5% AP

- Whole-body-estimation 에 two-stage-distillation 형식을 제시함
  - 추가 데이터 필요 없음
  - 0.2 AP 성능 향상

Method	RTMPose* x-l			
First-stage	-	✓	-	✓
Second-stage	-	-	✓	✓
body	69.7	**70.4**	69.7	**70.4**
foot	65.5	65.8	65.9	**66.2**
face	84.1	84.1	84.2	**84.3**
hand	55.1	56.4	55.4	**56.6**
whole-body	62.1	62.9	62.2	**63.1**

표. 새로운 형식을 통한 0.2 AP 향상