ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation

tag	Human Pose Estimation Vision
☑	☐
Github	https://github.com/ViTAE-Transformer/ViTPose
Paper Link	https://arxiv.org/abs/2204.12484
published	NeurIPS 2022

▼ references

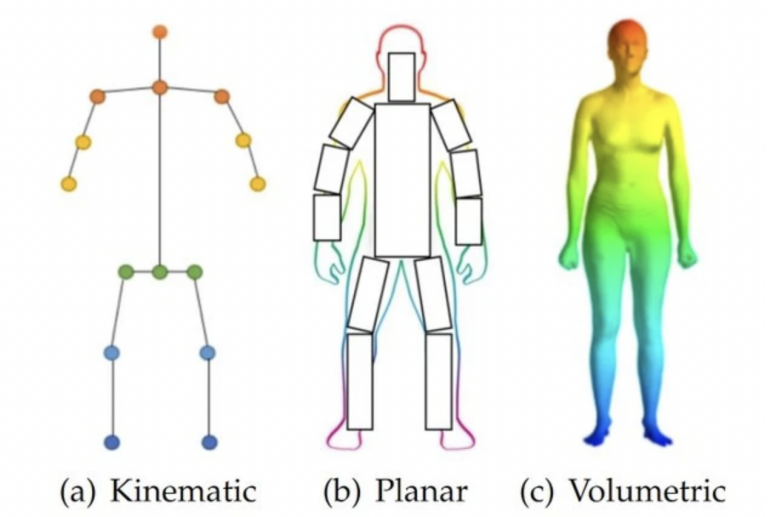
▼ 목차

1. Background

Human Pose Estimation

이미지나 비디오에서 인체 부분 중 머리, 몸체, 팔, 다리와 같은 관절의 위치를 올바르게 추정하는 것

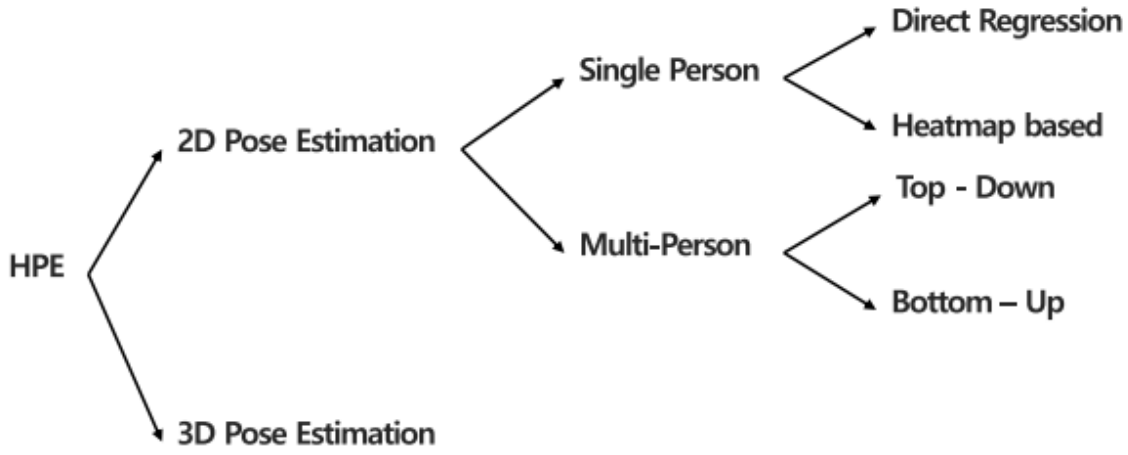
> **Human pose Estimation의 발전**



Human Pose Modeling: The three types of models for human body modeling – Source

- **Contour based model(Plannar)**
  - 이미지에서 객체의 외곽선(contour)를 찾으며, 고유한 특징을 기반으로 한 객체 인식에 사용
  - Ex. 사람의형상을인식, 물체를분리, 텍스트를 인식
- **Skeleton – based model (Kinematic model)**
  - 신체 골격 구조를 구성하는 관절(joint, key point)로 이루어진 모델
  - 2D pose에서는 (x,y)를 사용하고, 3D pose는 (x,y,z)를 사용
  - Part : (joint, key point) / Limb : (part pair, part connection) : 두 관절의 연결
    - 단, 코와 눈 같이 연결되더라도 관절의 연결로 볼 수 없는 pair도 존재.
    - texture나 shape 정보를 잘 표현하지 못함
- **Volume – based model**
  - 3D body shape 및 3D pose estimation에 사용
  - 3D mesh data를 활용해서 모델링.

HPE hierarchy

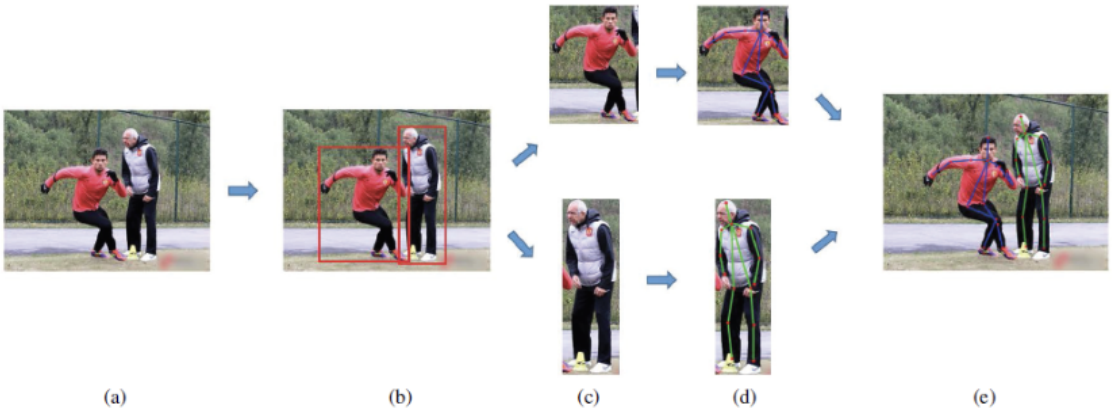


HPE Approaches

- **Bottom – Up** : 이미지에서 사람의 관절을 모두 추정 → 특정한 포즈 또는 하나의 사람 객체의 포즈로 그룹 지어주는 방식 (ex. DeepCut model

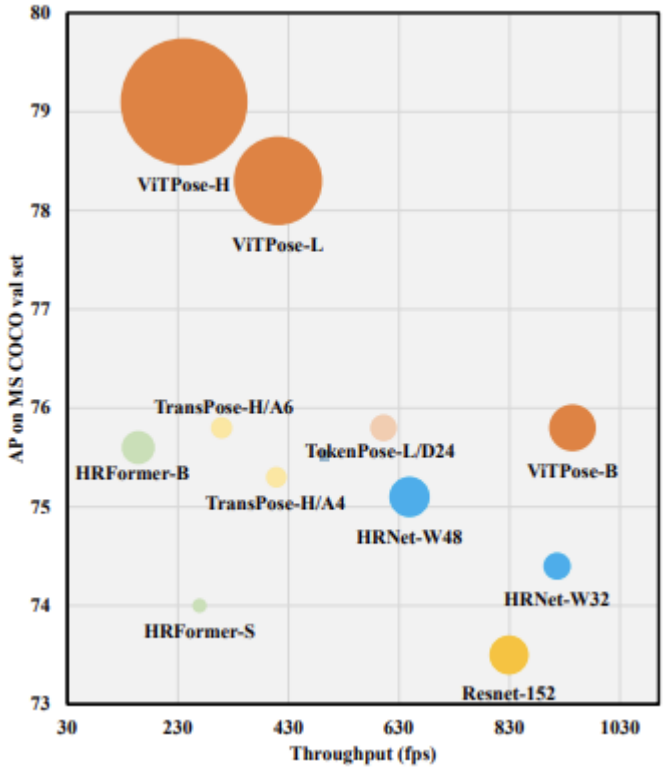
Bottom – Up

- **Top – Down** : 사람 detector를 이용해 사람 객체를 먼저 뽑아냄 → 각 사람 객체에서 관절 추정



Top – Down

2. Introduction



comparison of ViTPose and SOTA methods on MS COCO val set

- pose estimation 분야에서도 PRTR, TokenPose, TransPose 등 transformer의 구조를 적용한 모델이 제안되고 좋은 성능을 보여주고 있음. 하지만, backbone으로 **CNN을 필요로한다**는 공통적인 특징을 가지고 있음
- **오직 plain한 transformer를 backbone**으로 하여 pose estimation의 task를 수행하는 **ViTPose** 제안

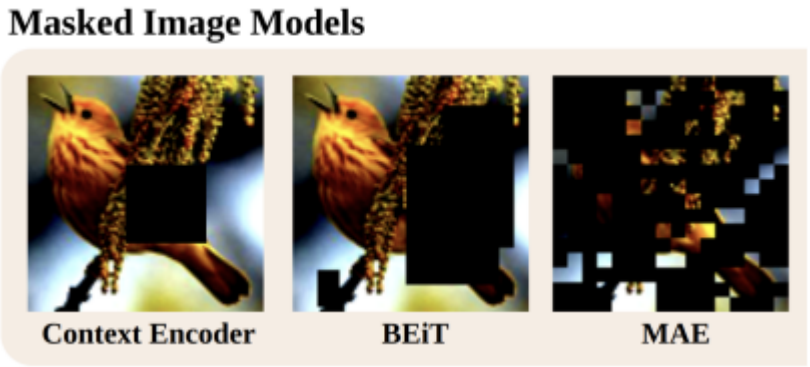
- **단순하고 정교하지 않은 model의 구조**를 통해서도 MS COCO 데이터셋에서 large 모델이 **80.9AP로 SOTA를 달성**

3. ViTPose

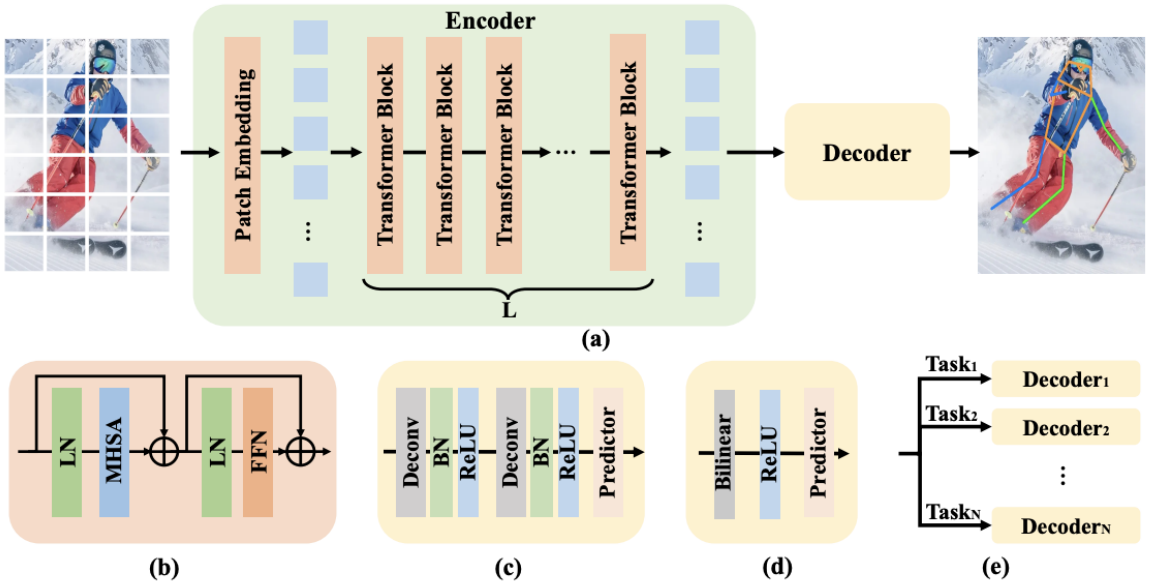
- 사람 인스턴스의 특성을 추출하기 위해 일반적이고 비계층적인 Vision Transformers를 backbone으로 사용하며, 이 backbone은 masked image modeling으로 pre-train됨

  ▼ **masked image modeling 이란?**

  - 이미지 데이터의 **일부 픽셀을 랜덤하게 Masking**하고 이를 복원하는 방법으로, BERT와 같은 Language modeling의 방법에서 착안
  - 자연어와 달리, 주변의 context 정보가 아니라 **local correlation**을 이용하여 복원



- 자세 추정을 위해서는 lightweight decoder를 사용하며, 이 디코더는 upsampling과 heatmap의 keypoints를 regressing함



(a) The framework of ViTPose

(b) The transformer block

(c) The classic decoder

(d) The simple decoder

(e) The decoders for multiple datasets

3.1 Simplicity

- ViTPose는 백본 인코더를 설계할 때 어떤 특정 도메인 지식도 필요하지 않음

> **백본에서 추출된 특성을 처리하는 두 가지 종류의 Lightweight decoder**

(c) The classic decoder : 2개의 deconvolution layers + 1개의 prediction layer

- feature map을 2배로 업샘플링 → 1x1 커널 크기를 갖는 컨볼루션 레이어를 사용 → key point의 localization heatmap(K)을 얻음

$$K = \mathrm{Conv}_{1\times1}(\mathrm{Deconv}(\mathrm{Deconv}(F_{out}))),$$

(d) The simple decoder

- bilinear interpolation을 통해 4배로 업샘플링 → 3 × 3의 컨볼루션 레이어 → 히트맵을 얻음

$$K = \text{Conv}_{3\times3}(\text{Bilinear}(\text{ReLU}(F_{out}))).$$

⇒ (d)의 non-linear capacity가 적을지라도, (c) 와 신중하게 설계된 트랜스포머 기반 디코더와 비교했을 때 경쟁력 있음

⇒ 간단한 구조로 ViTPose가 **더 나은 parallelism**을 가짐 → **추론 속도와 성능 측면에서 새로운 pareto front에 도달함**

3.2 Scalability

- 구조적 단순성은 다양한 배포 요구 사항에 맞게 추론 속도와 성능을 균형있게 조절할 수 있는 확장성을 가짐
- transformer의 layer를 쌓거나 feature 차원을 증가 또는 축소시킴으로써 **차원을 쉽게 늘리고 줄일 수 있음**
- ViT-B, ViT-L, ViT-H를 사용하여 추론 속도와 성능 요구사항을 맞출 수 있음
  - ViT-H와 ViTAE-G에서는 14x14 size의 patch embedding을 진행했는데, 이는 zero padding을 해주어 16x16으로 세팅

3.3 Flexibility

- 입력 이미지의 다양한 해상도와 특성 해상도에 잘 적응함.
- 다른 특성을 가진 dataset으로 학습해도 decoder만 수정하면 다른 특성에 대해서도 적용 가능
  - 단일 포즈 데이터셋 학습 외에도 디코더를 추가하여 다중 자세 추정 데이터셋도 학습할 수 있음

⇒ joint training pipeline을 생성하고 상당한 성능 향상을 이루고 상당히 가벼운 디코더로 인해 **아주 적은 추가 연산비용** 발생

3.4 Transferability

- 큰 ViTPose모델에서 학습한 지식을 전이시킴으로서 작은 ViTPose모델의 성능을 향상
- Knowledge token T를 teacher model의 **patch embedding 이후 visual token과 결합**
- Teacher model을 학습하고 **optimal token t***를 구함
- Optimal token t*를 student model의 visual token과 **결합**하여 학습을 진행

$$L_{t\to s}^{tod} = MSE(S(t*;X), K_t) + MSE(S(t*;X), K_{gt})$$

- **Student network의 loss = output distillation loss + teacher loss**

4. Ablation study and analysis

**일부 실험만 정리

- 기존 Vision Transformer와 동일한 size로 설정
- Top-down setting

4.1 Comparison with SOTA Methods

- 256 × 192 input resolution with multi-task training

Model	Backbone	Params (M)	Speed (fps)	Input Resolution	Feature Resolution	COCO val AP	COCO val AR
SimpleBaseline [42]	ResNet-152	60	829	256x192	1/32	73.5	79.0
HRNet [36]	HRNet-W32	29	916	256x192	1/4	74.4	78.9
HRNet [36]	HRNet-W32	29	428	384x288	1/4	75.8	81.0
HRNet [36]	HRNet-W48	64	649	256x192	1/4	75.1	80.4
HRNet [36]	HRNet-W48	64	309	384x288	1/4	76.3	81.2
UDP [18]	HRNet-W48	64	309	384x288	1/4	77.2	82.0
TokenPose-L/D24 [27]	HRNet-W48	28	602	256x192	1/4	75.8	80.9
TransPose-H/A6 [44]	HRNet-W48	18	309	256x192	1/4	75.8	80.8
HRFormer-B [48]	HRFormer-B	43	158	256x192	1/4	75.6	80.8
HRFormer-B [48]	HRFormer-B	43	78	384x288	1/4	77.2	82.0
ViTPose-B	ViT-B	86	944	256x192	1/16	75.8	81.1
ViTPose-B*	ViT-B	86	944	256x192	1/16	77.1	82.2
ViTPose-L	ViT-L	307	411	256x192	1/16	78.3	83.5
ViTPose-L*	ViT-L	307	411	256x192	1/16	78.7	83.8
ViTPose-H	ViT-H	632	241	256x192	1/16	79.1	84.1
ViTPose-H*	ViT-H	632	241	256x192	1/16	79.5	84.5

ViTPose	ViTAE-G	80.9	94.8	88.1	77.5	85.9	85.4
ViTPose$^+$	**ViTAE-G**	**81.1**	**95.0**	**88.2**	**77.8**	**86.0**	**85.6**

- ViTPose-L은 비슷한 inference 속도로 이전 CNN SOTA Model보다 좋은 성능을 보여줌
- ViTAE-G를 갖춘 단일 ViTPose 모델은 MS COCO 테스트 개발 세트에서 **80.9 AP로 최고 performance 달성**

5. Conclusion

- pose estimation에 간단하지만 효과적인 basemodel을 제안하였으며, 정교한 구조적인 설계나 복잡한 프레임워크를 사용하지 않고도 **MS COCO Keypoint 데이터셋에 SOTA를 달성**

- 추가적인 Mechanism과 FPN Structure와 같은 Complex한 Decoder를 사용하면 성능이 더 증가할 것으로 기대됨
- ViTPose의 Simplicity, Scalability, Flexibility, and Transferability을 증명함
- 하지만, Prompt-based Tuning을 통해 flexibility의 성능을 증명하지 못함
- 향후 Animal Pose Estimation, Face Keypoint Detection 등의 자세 추정 task에도 적용하는 것을 기대할 수 있음