METRO: End-to-End Human Pose and Mesh Reconstruction with Transformers

School of Industrial and Management Engineering, Korea University

Jin Hyeok Park

KOREA UNIVERSITY

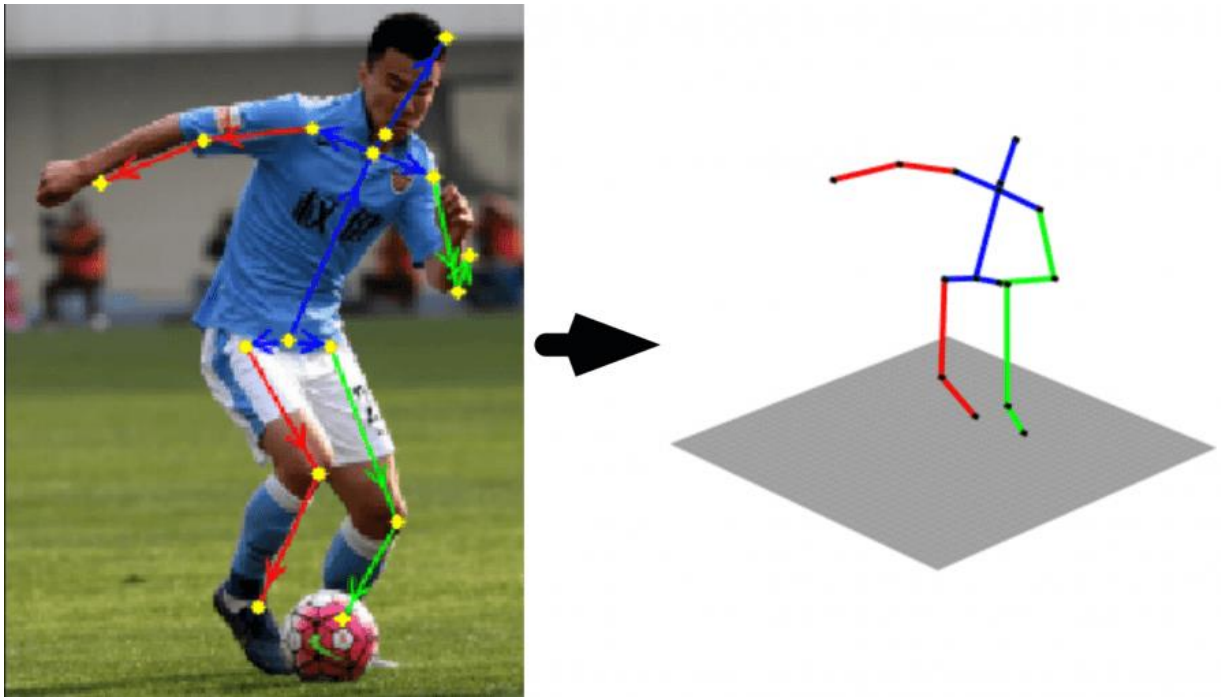
DMQA

Contents

Introduction

3D Human Pose & Mesh Reconstruction Task

- ❖ VR, 스포츠 동작 분석 등 다양한 Application에 이용 가능해 많은 관심을 받고있는 Task
- ❖ 관절 운동의 복합성과, Occlusion 문제 때문에 Challenging한 Task



<3D Human Pose>

Overview of METRO

METRO

- Microsoft
- 2021년 11월 28일 기준 37회 인용
- Transformer를 human pose에 적용한 연구

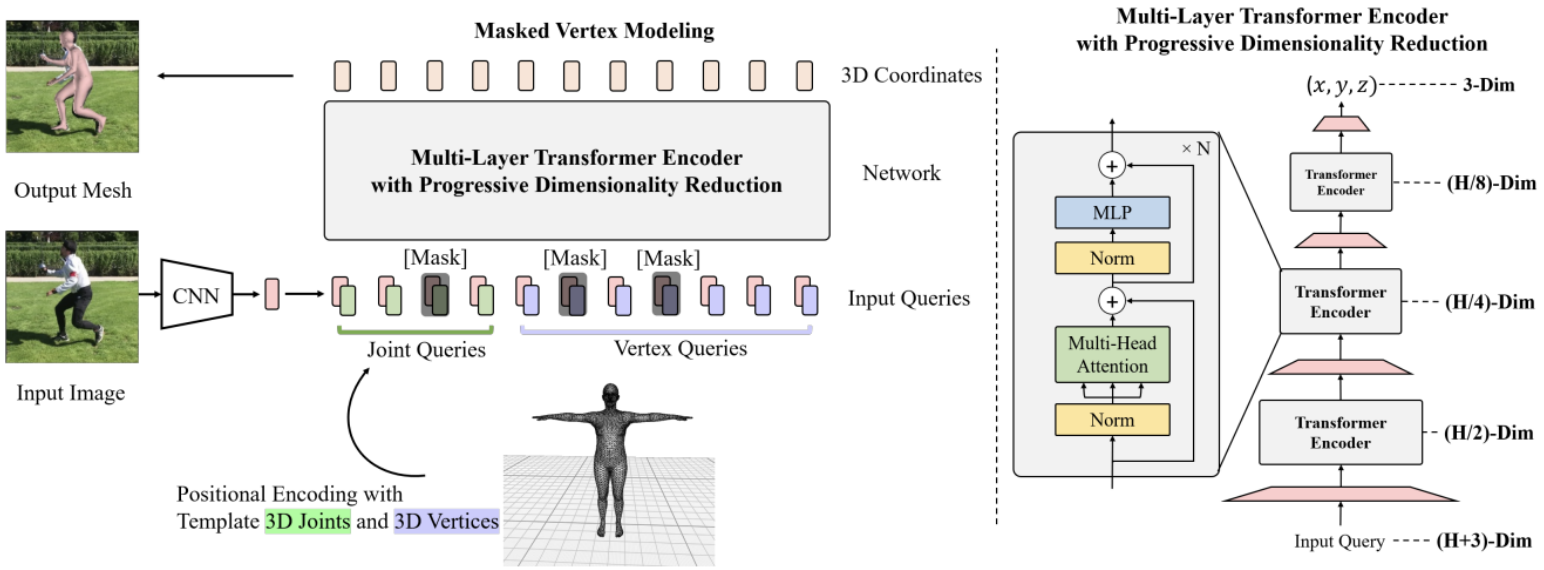
End-to-End Human Pose and Mesh Reconstruction with Transformers

Kevin Lin Lijuan Wang Zicheng Liu
Microsoft
{keli, lijuanw, zliu}@microsoft.com

Overview of METRO

METRO

- ❖ 단일 이미지로부터 3D Human Pose와 Mesh Vertex를 추출하기 위한 방법론
- ❖ Transformer의 Encoder 구조를 활용
- ❖ Transformer 구조를 이용해 간단하지만 효율적인 Global Interaction Modeling을 구현

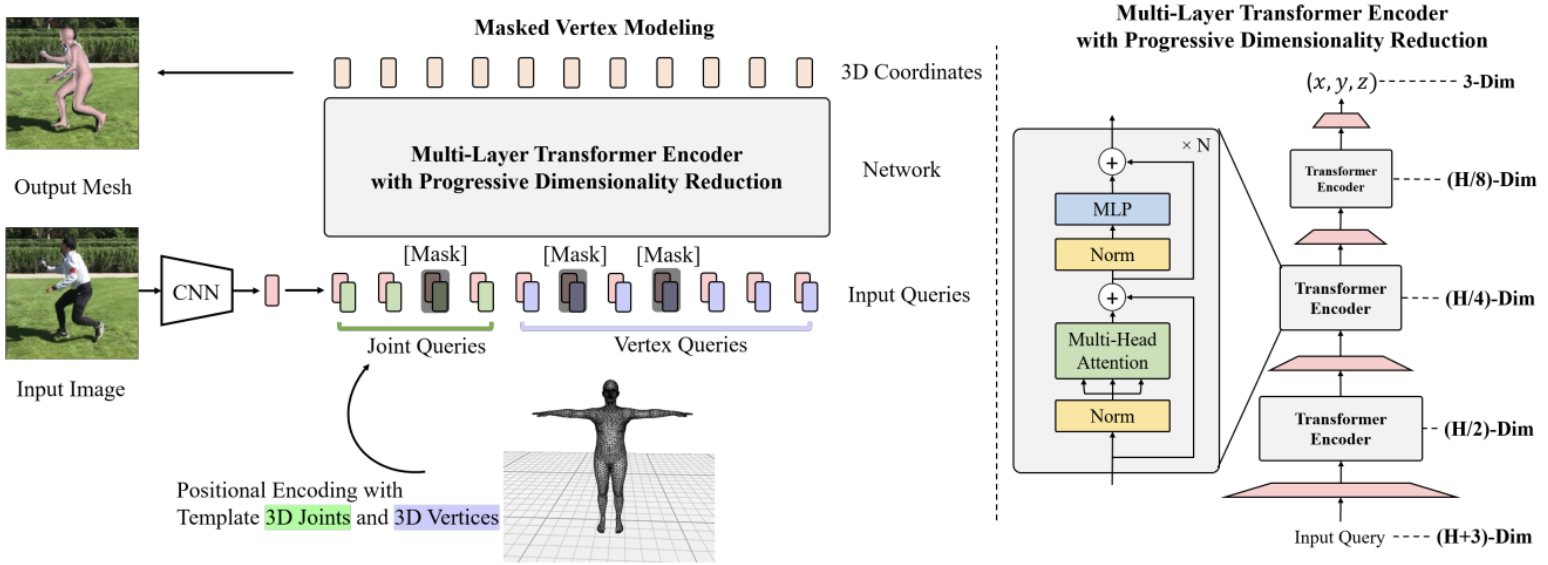


<METRO Architecture >

DMQA

METRO Architecture

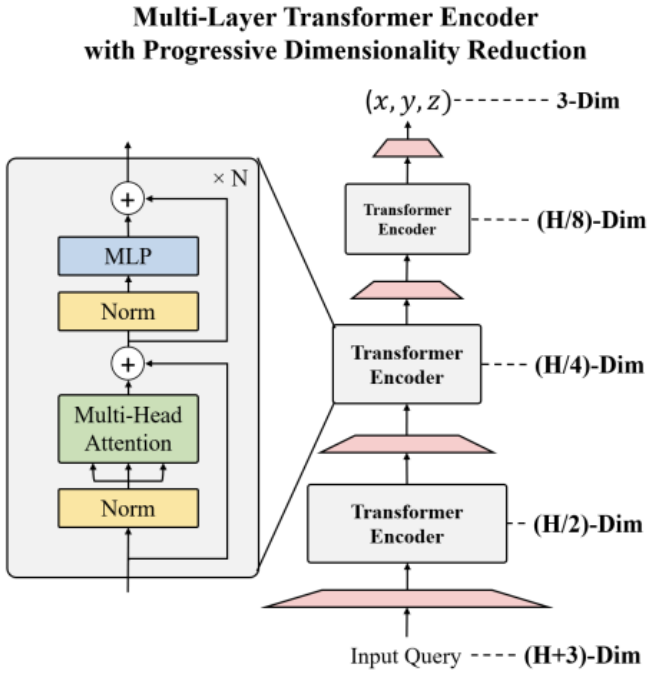
- CNN으로 Input이미지에서 Image Feature를 추출함
- Template joint와 Vertex를 Concat한 뒤 Positional Encoding을 진행
- Joint, Vertex Query Set이 주어지면 병렬적으로 3D Coordinate Value를 Regresstion함
  - ImageNet Classification Pretrain CNN을 사용
  - 마지막 Hidden Layer로 부터 Feature Vecter(X)를 얻음



DMQA

METRO Architecture

- 각 Token은 Layer를 거치면서 차원축소되며 3D Coord에 도달
- Encoder는 Progressive Dimensionality Reduction 구조 사용
- 각 Block은 4개의 Layer와 4개의 Head를 가지고 있음
- Dimension Reduction을 위해 Ecoder끝에 Embedding을 통해 Linear Projection을 사용
- 3D Coordinates에서의 Joint, Vertex 좌표가 Oupt으로 나옴

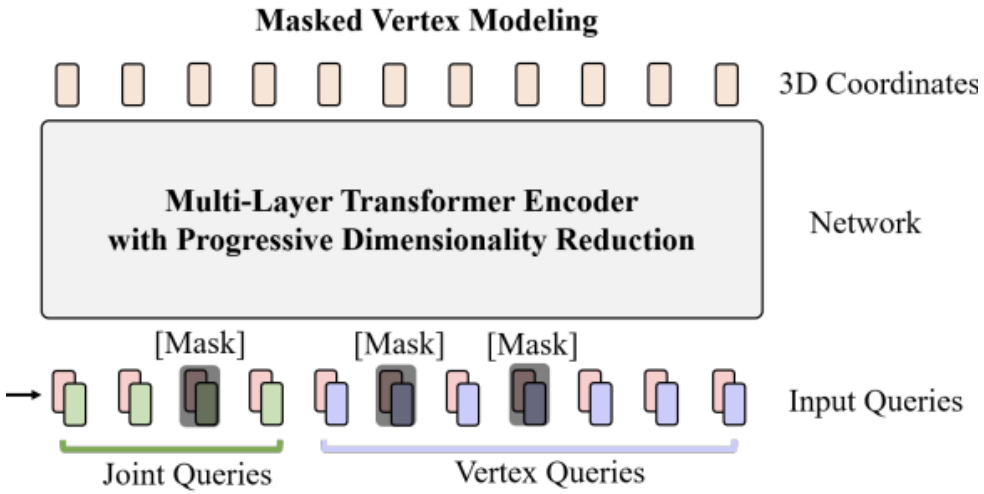


DMQA

METRO Architecture

MLM vs MVM

- MLM
  - MLM(Masked Language Modeling)을 통해 Bi-Direction을 극대화 하여 언어적 특성을 학습
  - 기존 MLM은 Input Recovering에 초점이 맞춰져 있어 3D Regression Task에 활용하기 어려움
- MVM
  - Input Query의 일부를 Masking함
  - Input을 Recovery하는 대신, Query로 Joint와 Vertex를 Regress하도록 학습
  - 결론적으로 Transformer가 가변적으로 필요한 Joint에 Attention 진행



DMQA

METRO Architecture

- L_1 Loss를 사용하여 3D Vertices와 3D Joints 최적화 진행
- 3D Joints를 mesh vertices 데이터에서 추출하는 방법도 존재하기 때문에 Pre-Defined Regression Matrix를 활용해서 3D Joints를 추론한 값을 L_1 Loss로 최적화
- Camera Parameter를 이용하여 3D Joints를 2D Joints로 Re-Projection해서 L_1 Loss 사용

$$\mathcal{L}_V = \frac{1}{M}\sum_{i=1}^{M}\left|\left|V_{3D} - \bar{V}_{3D}\right|\right|_1$$

$$\mathcal{L}_J^{reg} = \frac{1}{K}\sum_{i=1}^{K}\left|\left|J_{3D}^{reg} - \bar{J}_{3D}\right|\right|_1$$

$$\mathcal{L}_J = \frac{1}{K}\sum_{i=1}^{K}\left|\left|J_{3D} - \bar{J}_{3D}\right|\right|_1$$

$$\mathcal{L}_J^{proj} = \frac{1}{K}\sum_{i=1}^{K}\left|\left|J_{2D} - \bar{J}_{2D}\right|\right|_1$$

Experiments

$$MPVPE = \frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{V}_i - \boldsymbol{V}_i^*\|_2$$
$$MPJPE = \frac{1}{J}\sum_{j=1}^{J}\|\boldsymbol{P}_j - \boldsymbol{P}_j^*\|_2$$

- 3DPW: 아웃도어 이미지(2D, 3D) , 22000장의 Train, 35000장의 Test Dataset이 존재
- UP-3D: 아웃도어 이미지 데이터 셋 7000장, Annotation은 Model Fitting으로 생성
- H3.6M: 3D Mesh가 존재하지 않아 SMPLify-X로 Pseudo Data를 만들어서 사용
- 평가지표 식에서 P_j는 j번째 관절에 대한 실제 값, V_i는 i번째 정점에 대한 실제값

Method	3DPW			Human3.6M	
	MPVE ↓	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
HMR [22]	–	–	81.3	88.0	56.8
GraphCMR [25]	–	–	70.2	–	50.1
SPIN [24]	116.4	–	59.2	–	41.1
Pose2Mesh [8]	–	89.2	58.9	64.9	47.0
I2LMeshNet [32]	–	93.2	57.7	55.7	41.1
VIBE [23]	99.1	82.0	51.9	65.6	41.4
METRO (Ours)	**88.2**	**77.1**	**47.9**	**54.0**	**36.7**

Conclusion

❖ METRO: End-to-End Human Pose and Mesh Reconstruction with Transformers

- Non-Local Interaction을 위해 Masked Vertex Modeling을 제안
- METRO는 Input에 의존적이여서 고정된 Mesh Topology와 관계없이 Non-Local Interaction가능
- METRO는 다양한 Domain의 Reconstruction으로 확장 될 수 있음


Thank You