DeepCap

Monocular Human Performance Capture Using Weak Supervision

2020 연구실 하계 세미나

김 기 남

Vision & Display Systems Lab.

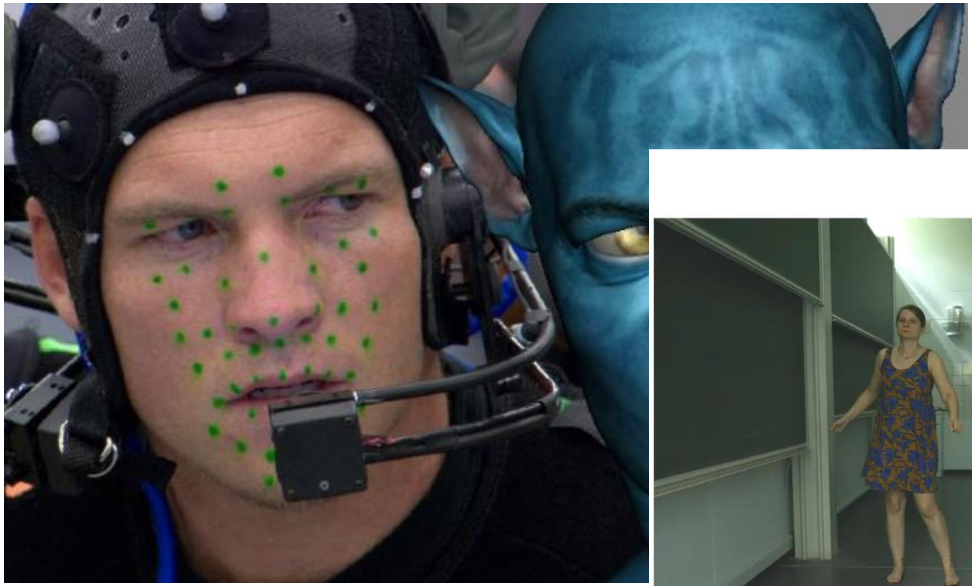
Dept. of Electronic Engineering, Sogang University

Outline

- GNN(Graph Neural Network)
- GCN(Graph Convolutional Network)
- GraphSAGE(GraphSAGE(SAmple and aggreGatE)
- How Powerful are Graph Neural Networks?
- References

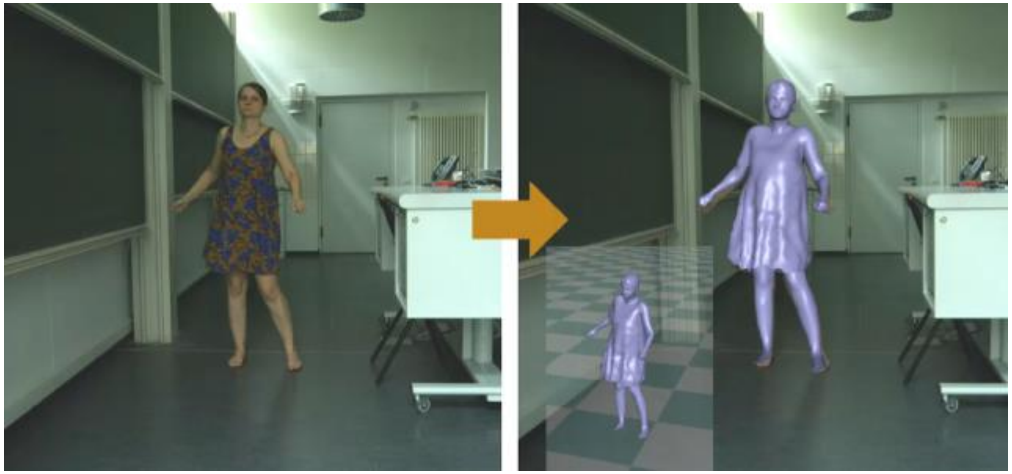
Introduction

- What is Human Performance Capture?
  - The space-time coherent 4D capture of full pose and non-rigid surface deformation of people in general clothing.



Introduction

- Challenges
  - Disadvantages of 3D data
    - In previous work, normally need 3D annotation(high cost)
    - High cost to inference model
      - Multi-view camera, Depth camera
  - High-dimensional problem
    - Input image: 2D
    - Output result: 3D

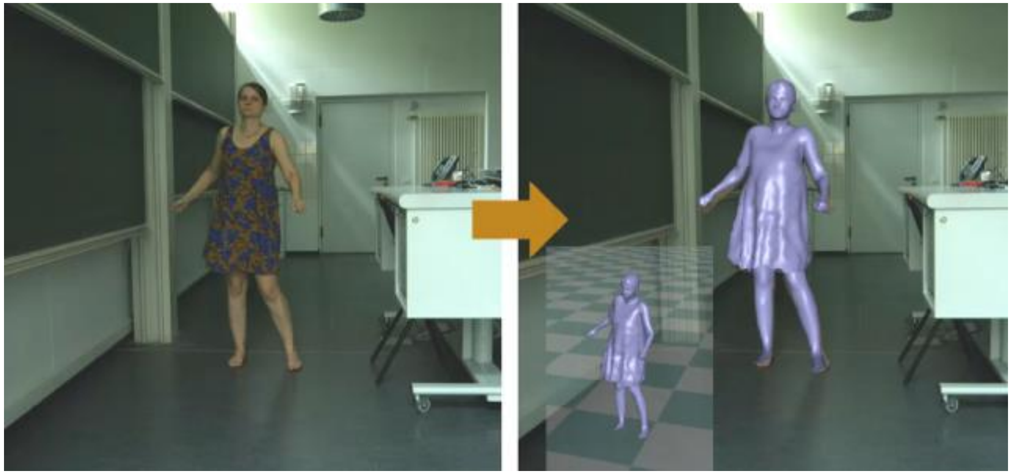


서강대학교 SOGANG UNIVERSITY

VDS LAB

Introduction

- Challenges
  - Disadvantages of 3D data
    - In previous work, normally need 3D annotation(high cost)
    - High cost to inference model
      - Multi-view camera, Depth camera
  - High-dimensional problem
    - Input image: 2D
    - Output result: 3D



Related Work

- Capture using parametric models
  - Pose estimation을 통해 추정된 Skeleton에 parameterize된 human body를 입히는 방식
    - 남성, 여성, 중성을 판단하고 각 성별에 맞는 parameter에 따라 body 생성
      - SMPL(Skinned Multi-Person Linear model)
    - 옷 등의 형태 및 질감 표현 불가능

Related Work

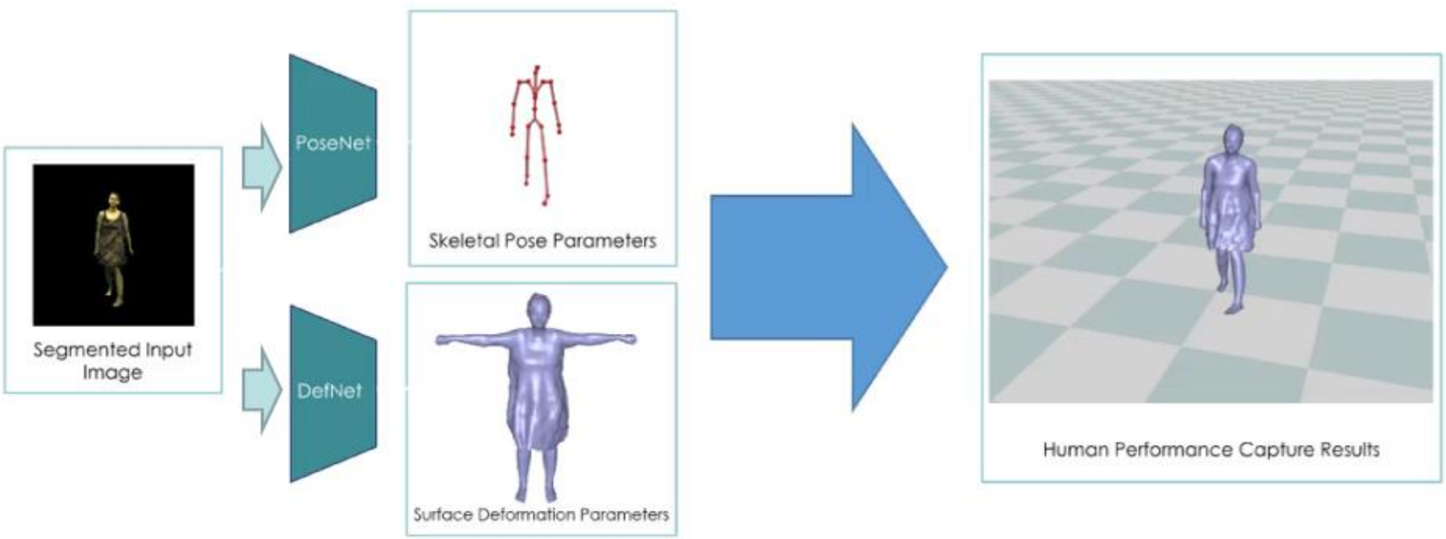
- Template-free capture
  - Depth-based Template-free Capture
    - 한 개 또는 여러 개의 depth sensor를 사용하여 얻어진 3D data를 이용하여 Human object 에 대해 reconstruction
    - Slow motion 및 변화가 크지 않은 motion 에 대해서만 사용가능
  - Monocular Template-free Capture
    - 2D image input 에서 voxel단위로 CNN을 통하여 reconstruction
    - Frame간의 correspondence 를 고려하지 않아 application level 에 부적합

Related Work

- Template-based capture
  - Template mesh를 사용하여 capture
    - Multi-view monocular camera setting 을 통해 template mesh 추출
      - multi-view setup 과정이 상당히 복잡함
      - input image 수가 너무 많아 computational cost가 상당히 높음

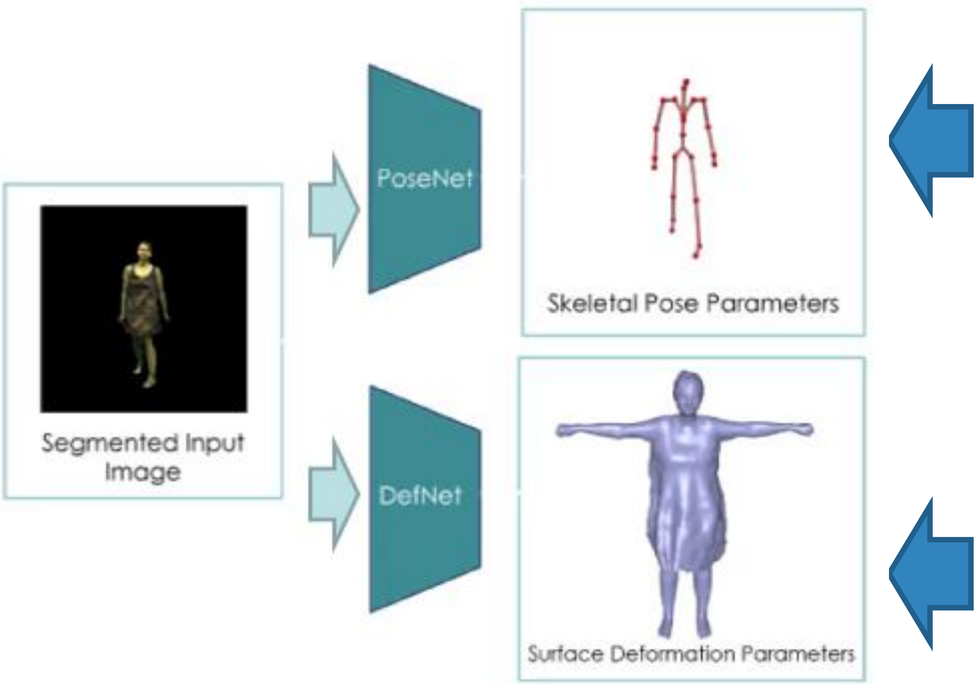
DeepCap

- Weak supervision 으로 학습하여 Single Monocular camera 를 이용한 inference 가능
- Input image의 skeleton 과 surface deformation parameter 를 estimation 하여 performance capture 수행
- Real-time 동작 가능(50ms/frame)



DeepCap

- Weak Supervision
  - Direct Supervision?



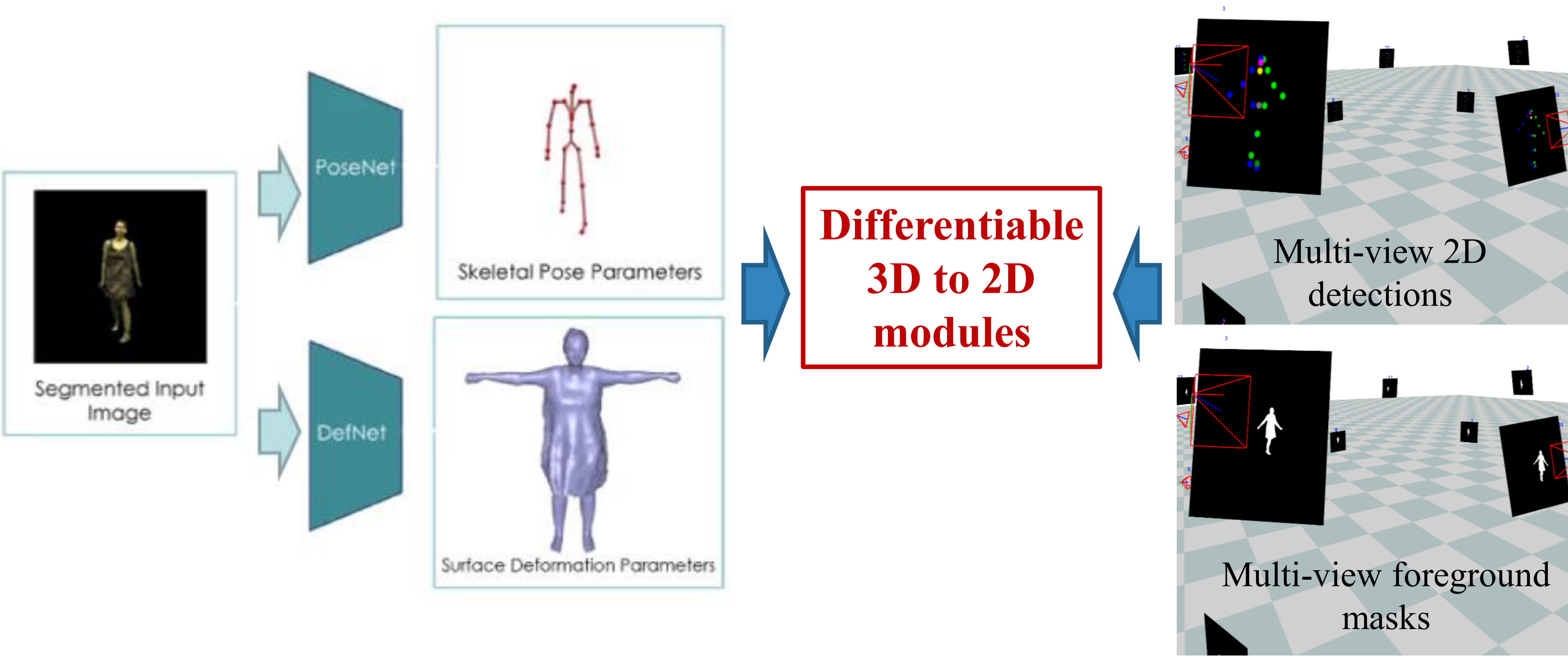
Ground truth 3D pose

Difficult to obtain

Ground truth 3D surface

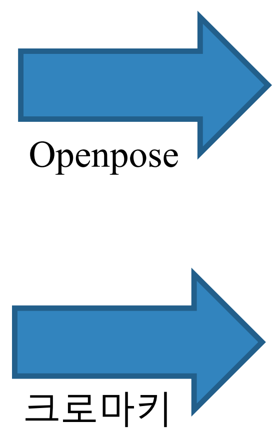
DeepCap

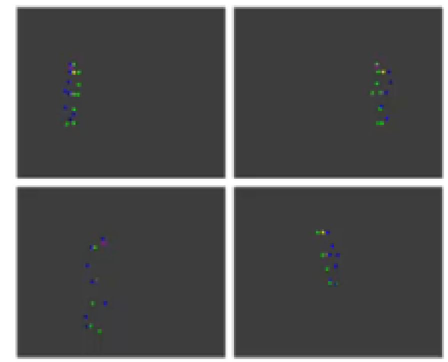
- Weak Supervision



DeepCap

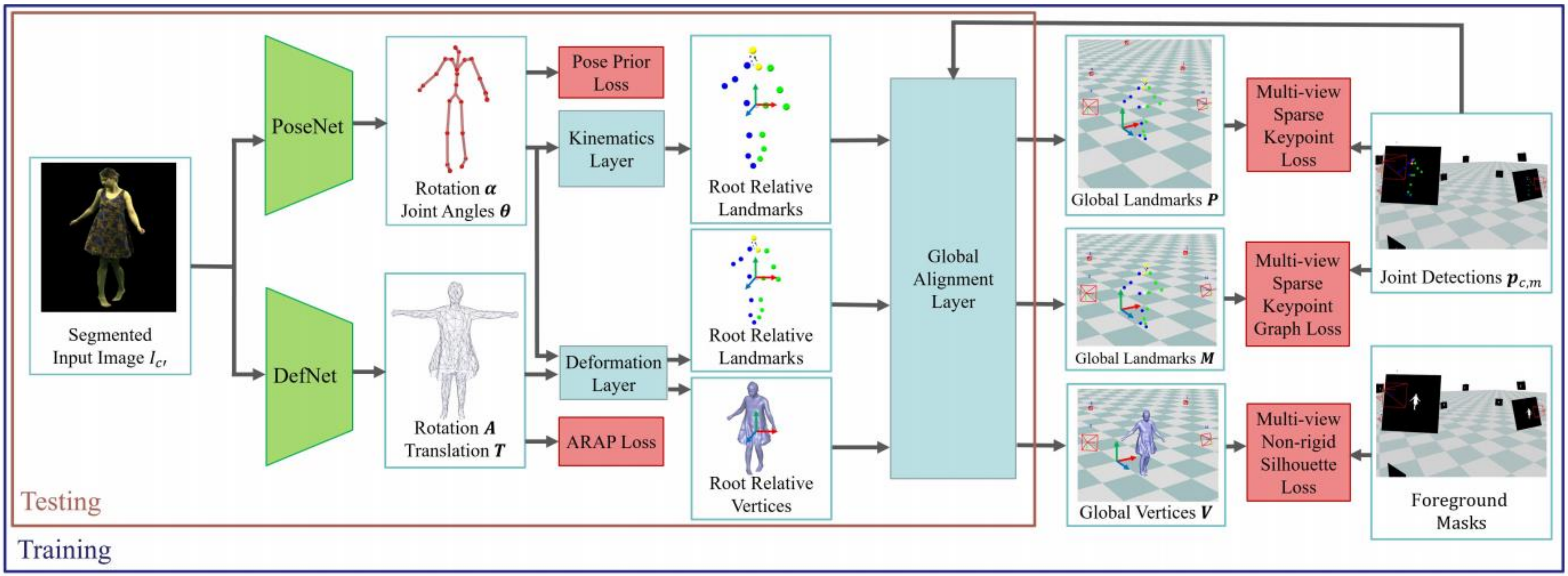
- Training Data
  - 학습시에만 2D multi-view images 사용
  - Openpose를 이용하여 GT Skeleton 추출
  - 크로마키 기법을 이용하여 GT Foreground mask 추출





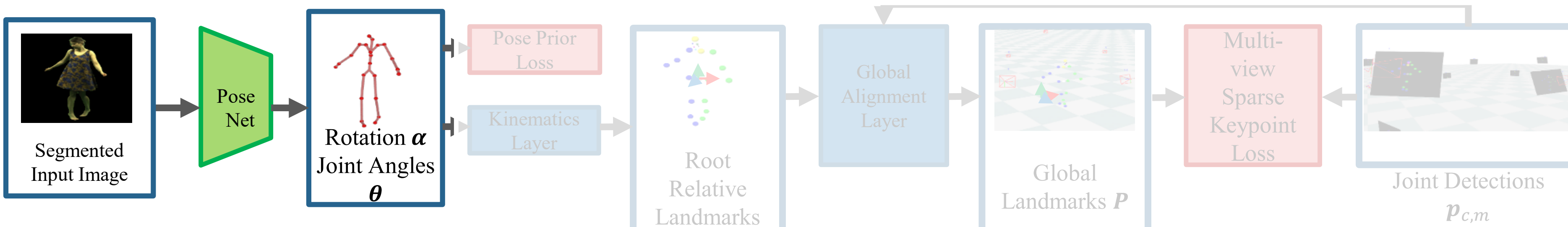
DeepCap

- Model Architecture



DeepCap

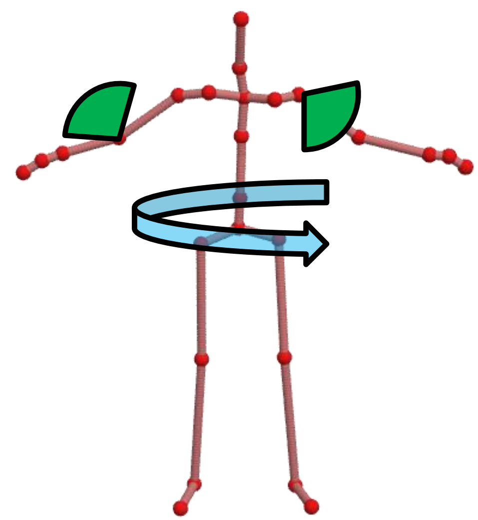
- Model Architecture
  - PoseNet



PoseNet

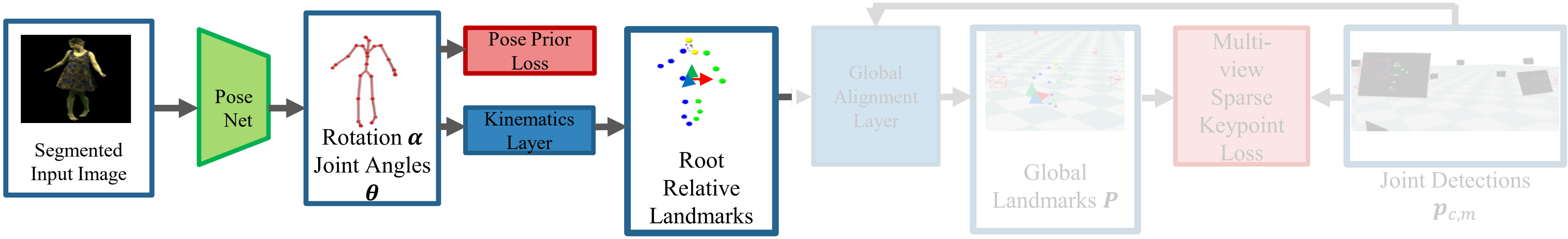
Root rotation $\boldsymbol{\alpha} \in \mathbb{R}^3$

Joint angles $\boldsymbol{\theta} \in \mathbb{R}^3$



DeepCap

- Model Architecture
  - PoseNet



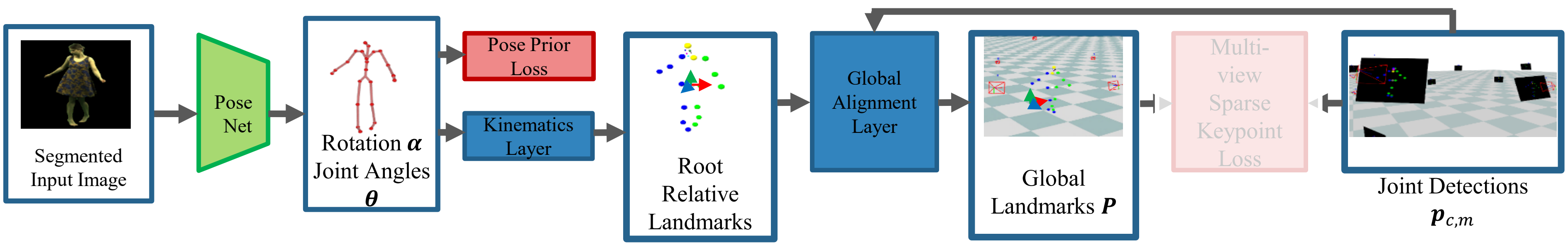
Kinematics Layer

Function $f_m(\alpha, \theta): \mathbb{R}^{30} \rightarrow \mathbb{R}^3$ per landmark m

Skeletal pose → **Camera** and **root relative** 3D landmark positions $P_{c',m}$

DeepCap

- Model Architecture
  - PoseNet



Rigid transform for landmark $\boldsymbol{P}_{c',m}$

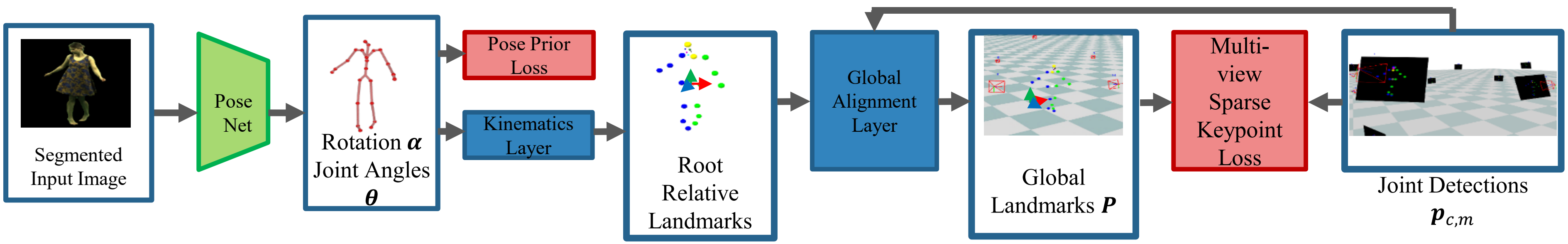
Camera and **root relative** 3D space → **Global** 3D space

$$\boldsymbol{P}_m = \boldsymbol{R}_{c'}^T \boldsymbol{P}_{c',m} + \boldsymbol{t}$$

$\boldsymbol{R}_{c'}^T$: Inverse extrinsic rotation of the input camera c' $\boldsymbol{t}$: Global translation

DeepCap

- Model Architecture
  - PoseNet



Multi-view Sparse Keypoint Loss

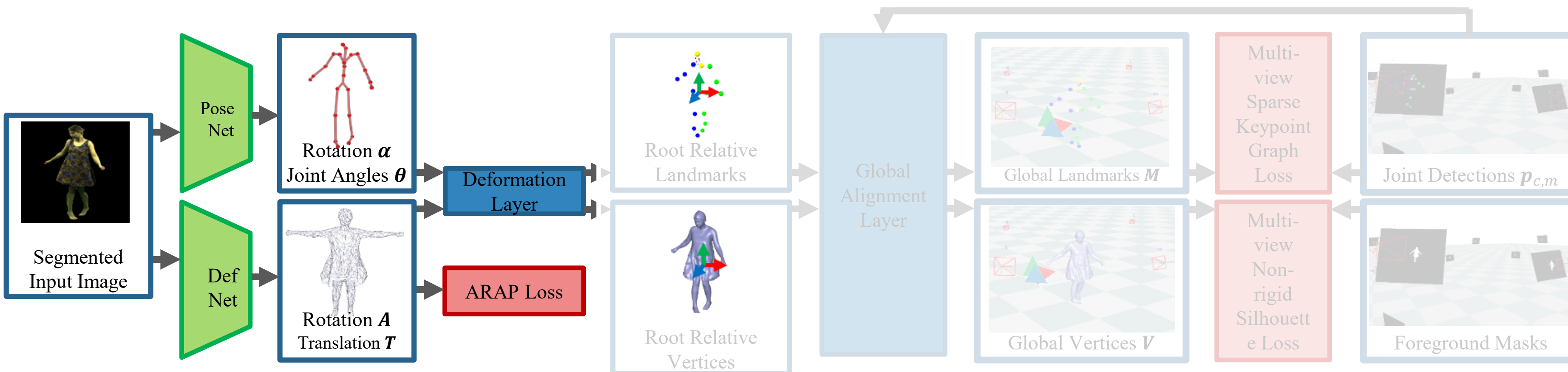
$$L_{kp}(\boldsymbol{P}) = \sum_{c} \sum_{m} \left\| \pi_c(\boldsymbol{P}_m) - \boldsymbol{p}_{c,m} \right\|_2^2$$

Projecting (π) 3D landmark $\boldsymbol{P}_m$ into camera view c

Comparing to 2D joint detection $\boldsymbol{p}_{c,m}$

DeepCap

- Model Architecture
  - DefNet



Pose

Deformation

Deformation Layer

Embedded deformation

Dual Quaternion Skinning

(Kavan et al. 2007)

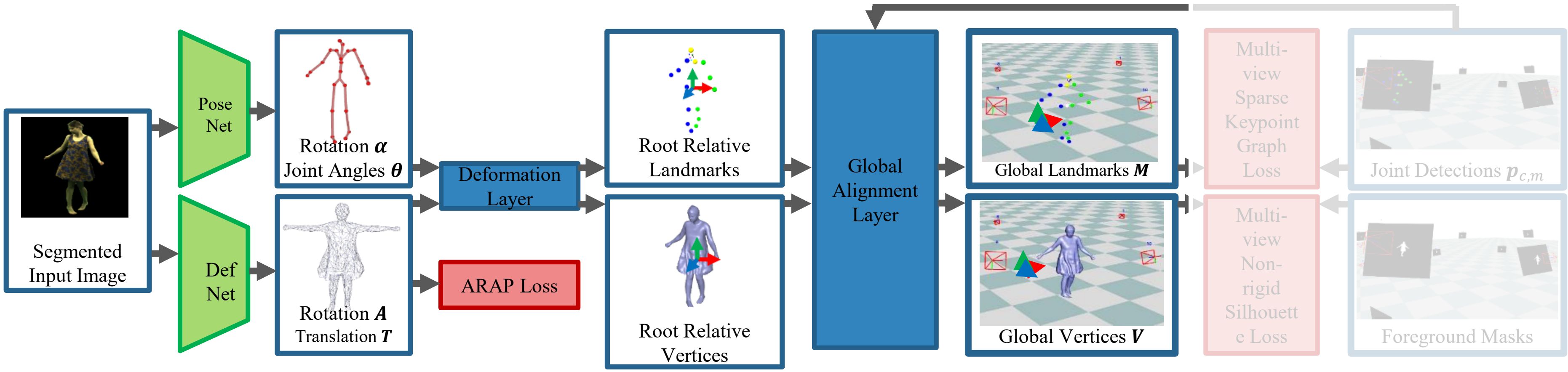
Posed and **deformed**

Landmarks $\boldsymbol{M}_{c',m}$

Vertices $\boldsymbol{V}_{c',i}$

DeepCap

- Model Architecture
  - DefNet

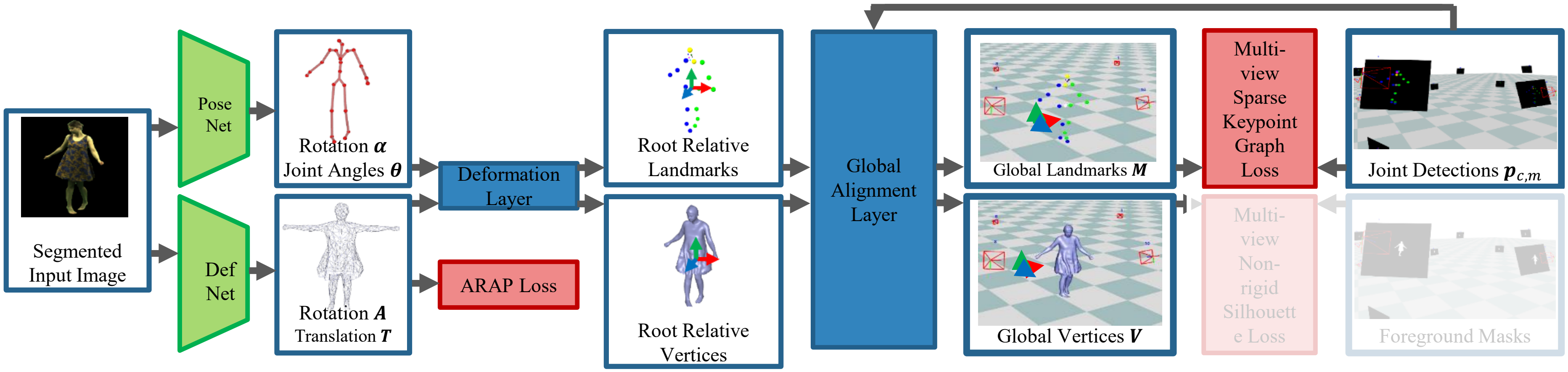


Rigid transform for landmark *m* and vertex *i*

Camera and root relative		Global
3D landmark $M_{c',m}$ and vertex $V_{c',i}$	→	3D landmark M_m and vertex V_i

DeepCap

- Model Architecture
  - DefNet



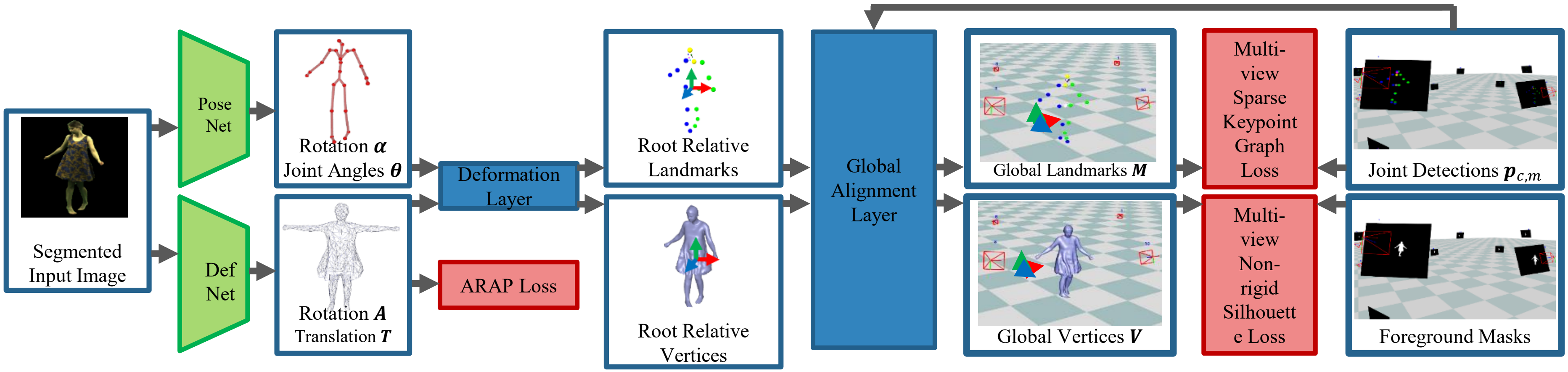
Multi-view Sparse Keypoint Graph Loss

$$L_{kpg}(\boldsymbol{P}) = \sum_{c} \sum_{m} \left\| \pi_c(\boldsymbol{M}_m) - \boldsymbol{p}_{c,m} \right\|_2^2$$

$\boldsymbol{M}_m$ **: Global** 3D landmark

DeepCap

- Model Architecture
  - DefNet



Non-rigid Silhouette Loss

$$L_{sil}(V) = \sum_{c} \sum_{i \in B_c} \left\| D_c\left(\pi_c(\mathbf{V}_i)\right) \right\|_2^2$$

B_c : Set of boundary vertices for camera c

D_c : Distance transform image

DeepCap

- Experimental result

Input Image | Our Result (overlayed) | Input Image | Our Result (overlayed) | 3D View

DeepCap

- Experimental result

3DPCK and AMVIoU (in %) on S4 sequence		
Method	**3DPCK↑**	**AMVIoU↑**
1 camera view	62.11	65.11
2 camera views	93.52	78.44
3 camera views	94.70	79.75
7 camera views	95.95	81.73
6500 frames	85.19	73.41
13000 frames	92.25	78.97
PoseNet-only	96.74	78.51
Ours(14 views, 26000 frames)	**96.74**	**82.53**

Thank you