# KTN: Knowledge Transfer Network for Multi-person DensePose Estimation


**Xuanhan Wang, Lianli Gao, Yixuan Zhou, Jingkuan Song and Heng Tao Shen**



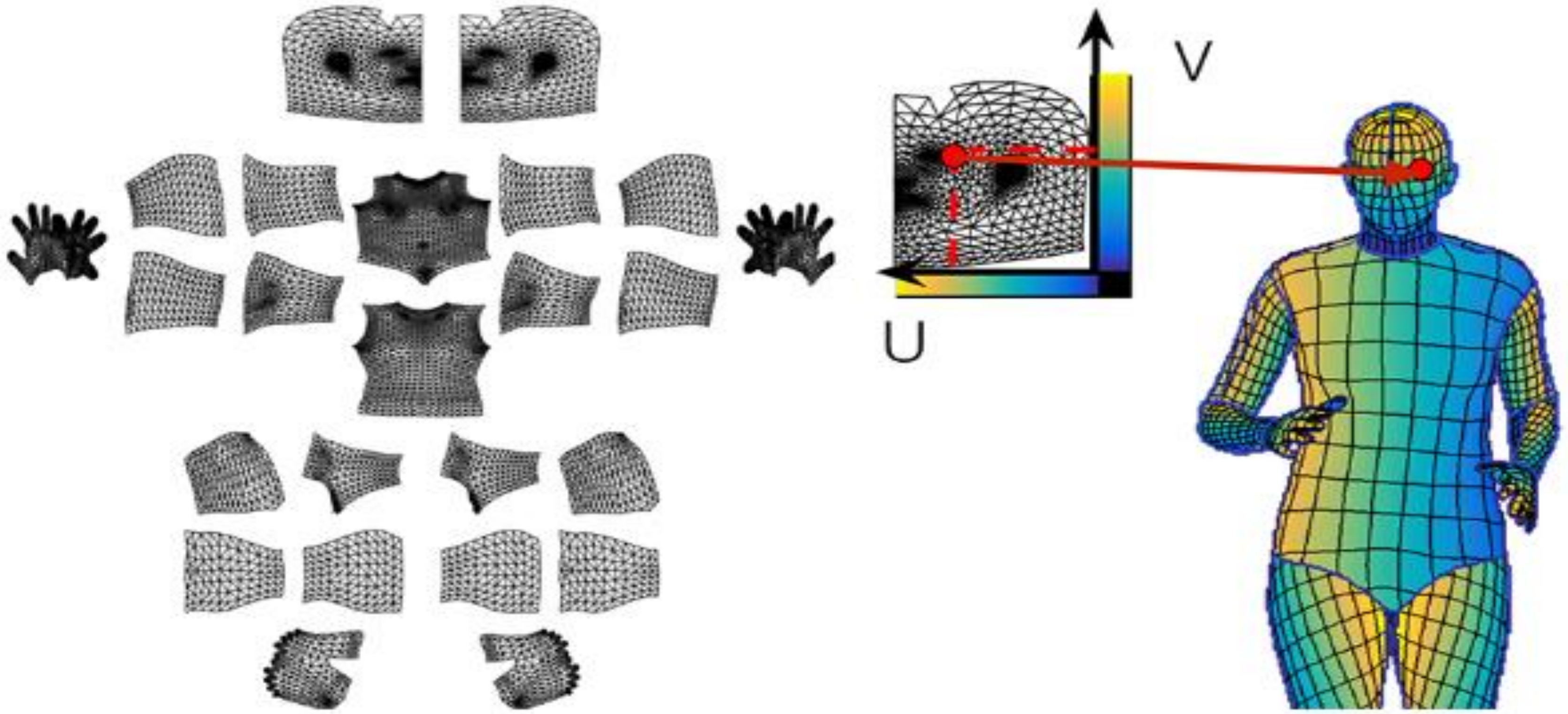
**Wang, X., Gao, L., Zhou, Y., Song, J., & Shen, H. T. (2020). KTN: Knowledge Transfer Network for Multi-person DensePose Estimation. In *28th ACM International Conference on Multimedia.***

# Outline

- Task Definition
- Motivation
- Method
- Experiments and Results
- Summary

# Task Definition

## Human DensePose Estimation

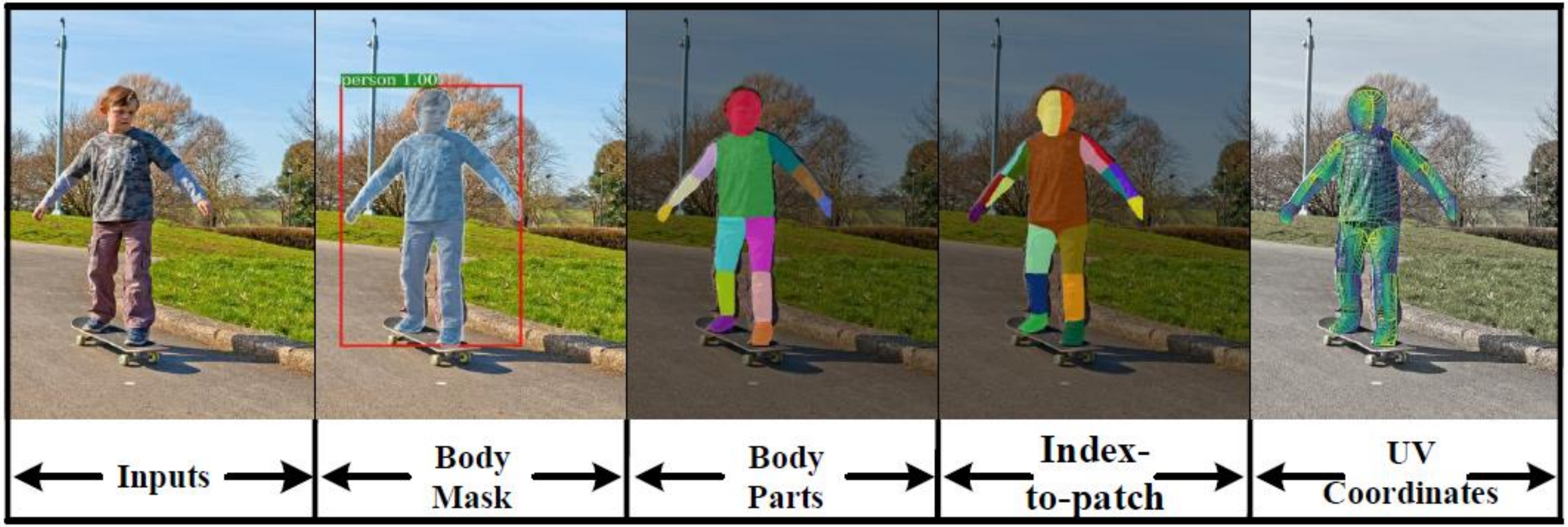



Mapping all human pixels of an RGB image to the 3D surface of the human body in challenging, uncontrolled conditions

**[1]** *Rıza Alp Guler, Natalia Neverova and Iasonas Kokkinos. DensePose: Dense Human Pose Estimation In The Wild. In CVPR 2018.*

# Task Definition

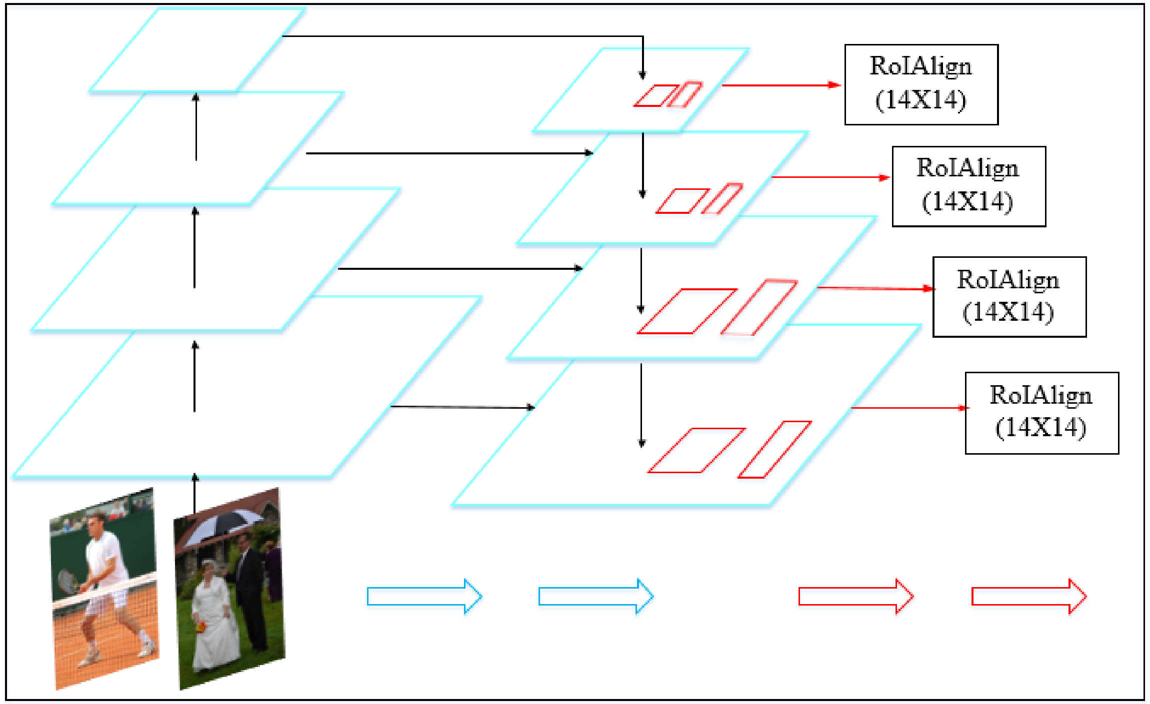
## Sub-tasks




Simultaneously detecting people, segmenting bodies or parts, and mapping body pixels to a standard 3D body template

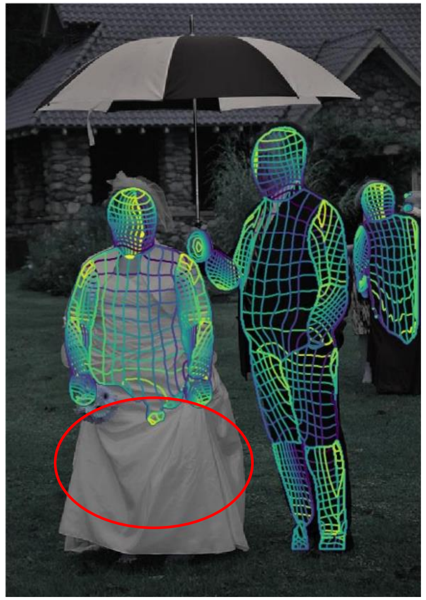
# Motivations

***How to design a simple yet effective pipeline for densepose estimation ?***




missing details

background interference

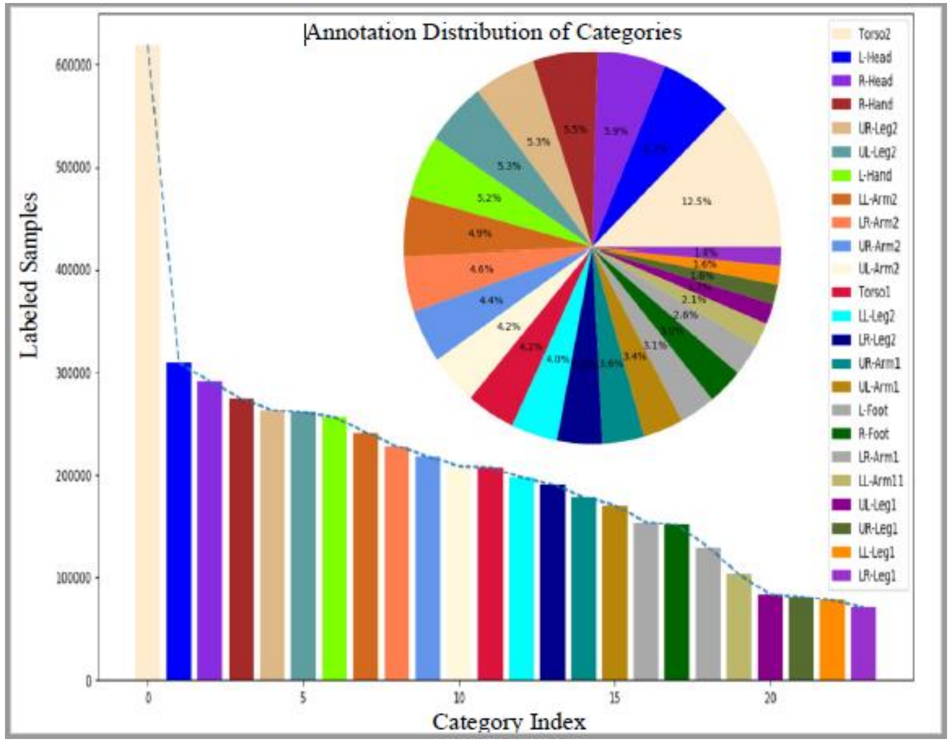


Pyramidal convolutional network → Incomplete Estimation

# Motivations

***How to handle the issue of limited annotations and class-imbalanced labels?***

Limited Annotations

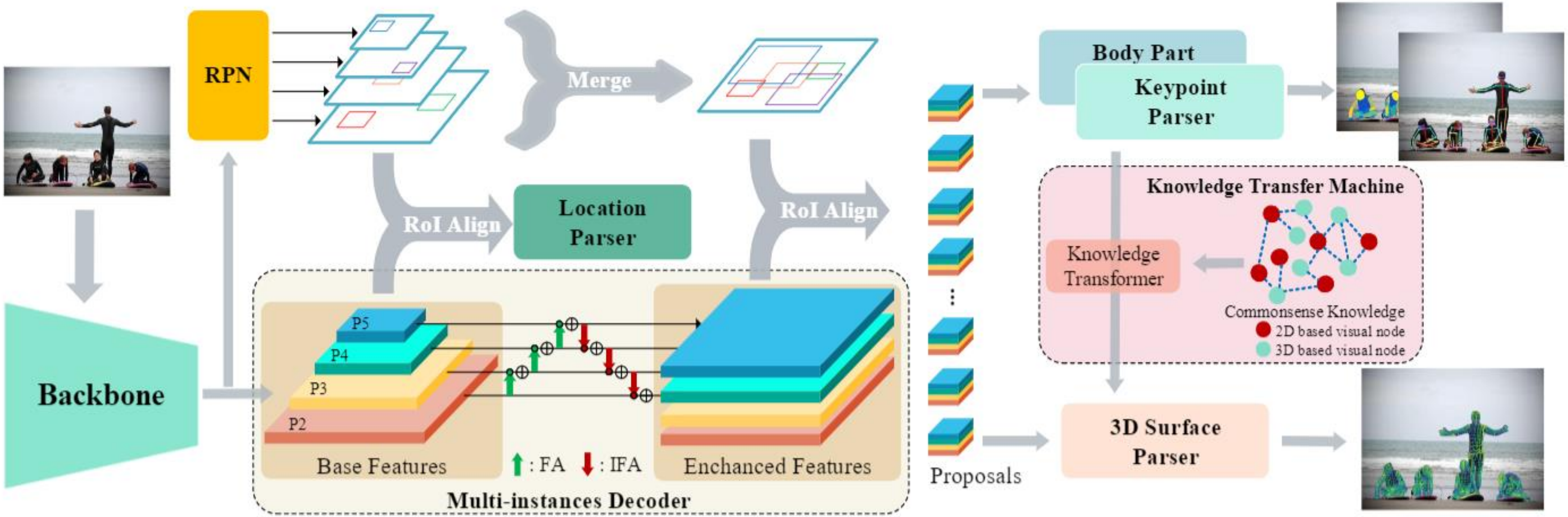



Class-imbalanced labels

# Method

## Knowledge Transfer Network

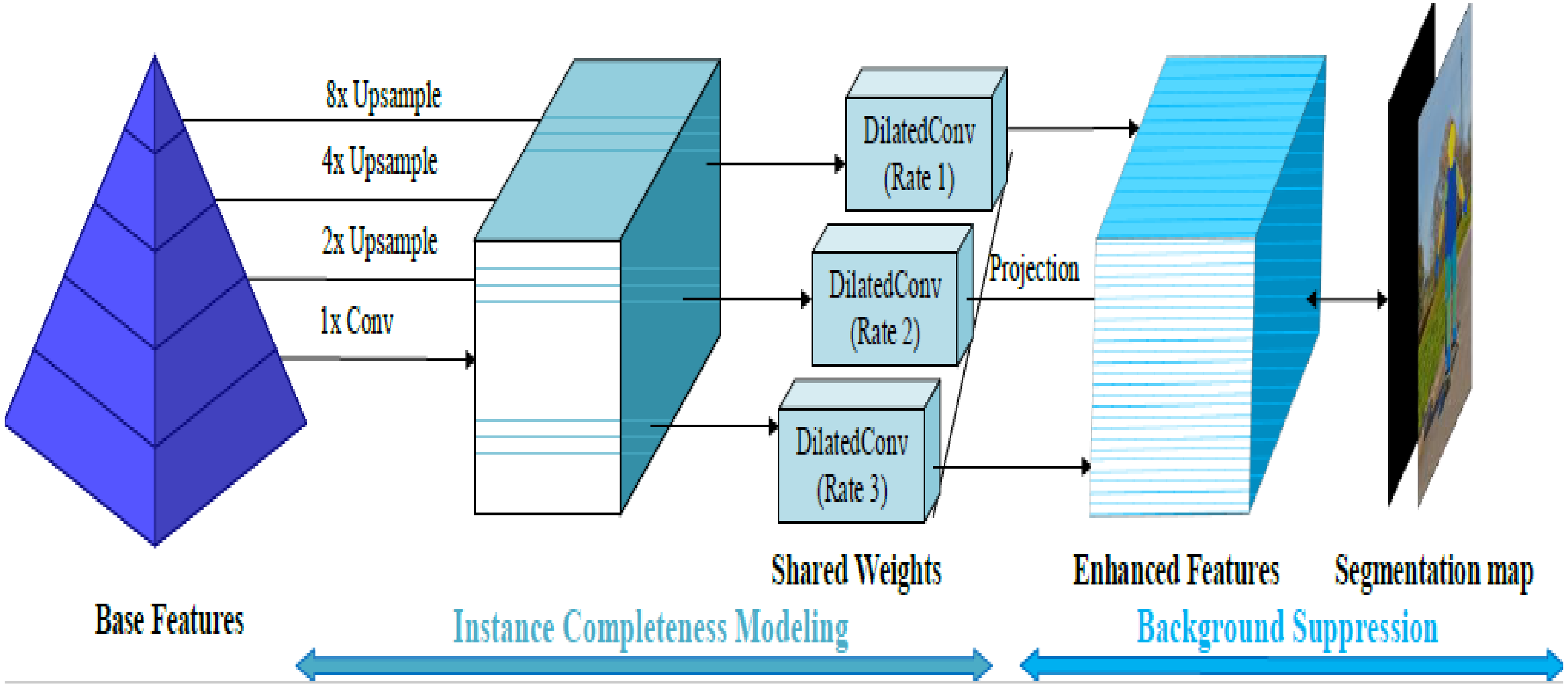
- Multi-instances Decorder (MID)
- Knowledge Transfer Machine (KTM)



# Method

## Multi-instances Decoder (V1)

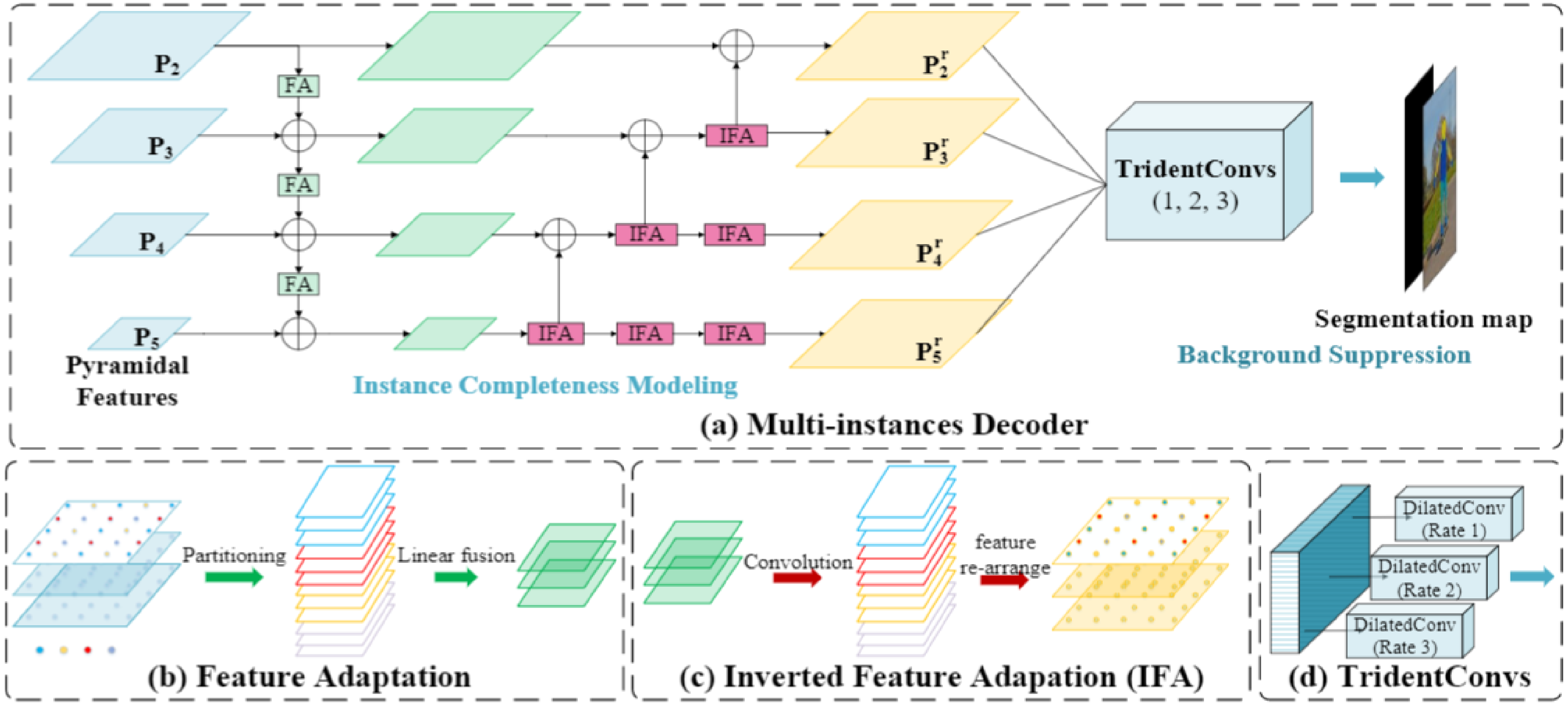
- Instance Completeness Modeling
- Background Suppression



# Method

## Multi-instances Decoder (V2)

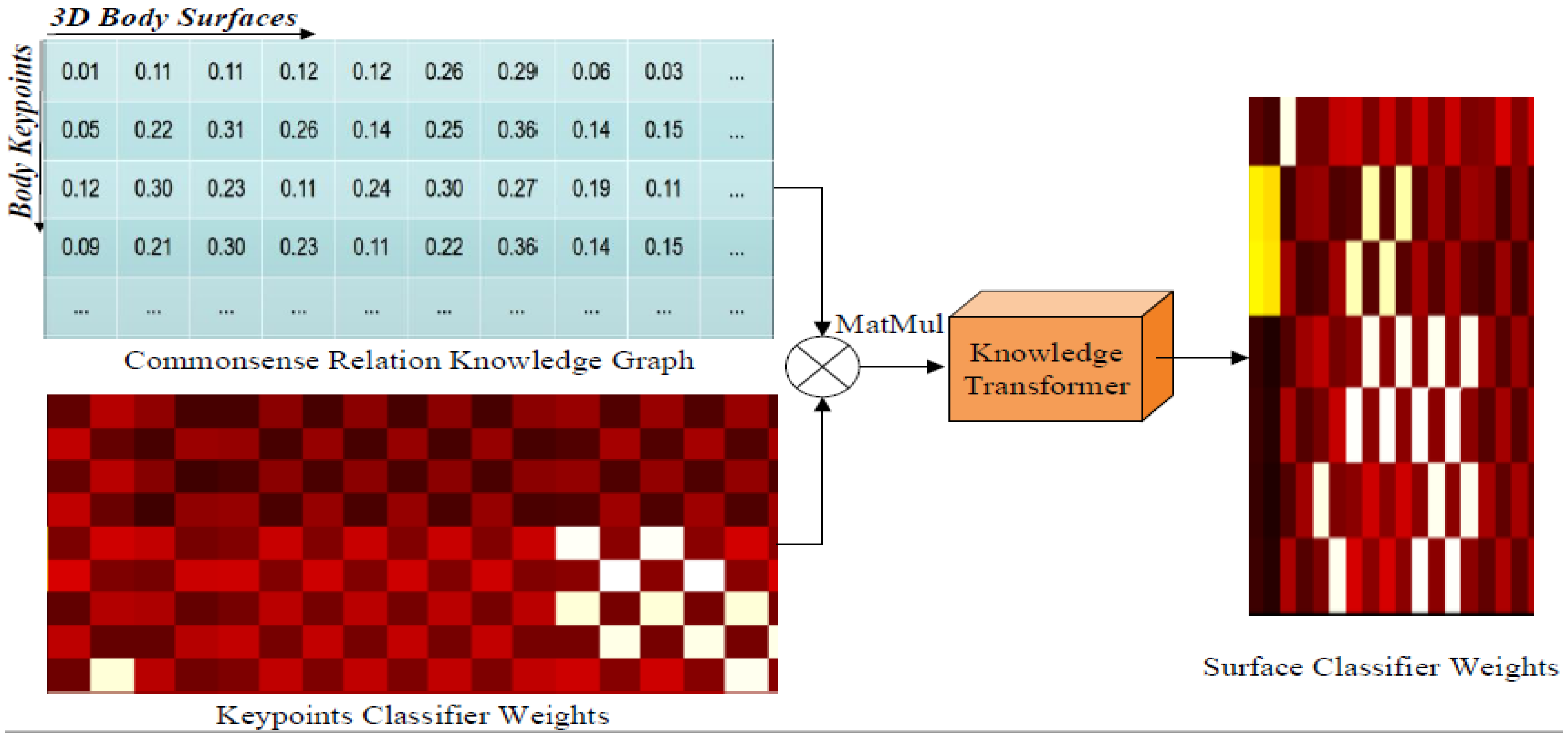
- Feature adjustment (Feature adaptation & Inverted Feature Adaptation)
- Background Suppression



# Method

## Knowledge Transfer Machine (V1)

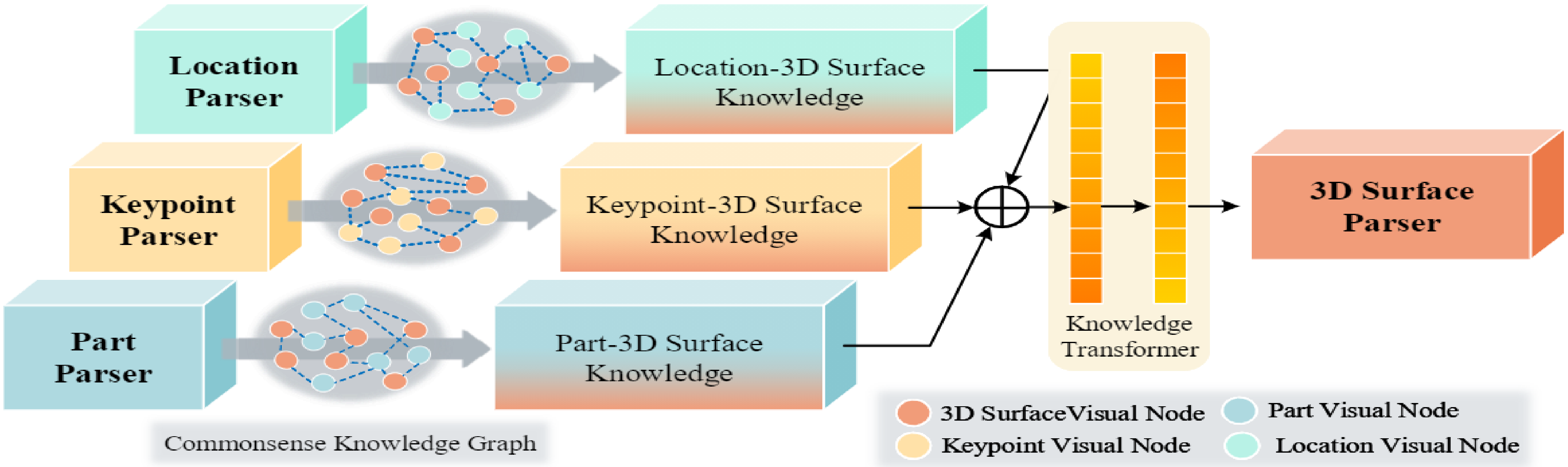
- Single-path knowledge graph (keypoint-to-surface)
- Parameter generation



# Method

## Knowledge Transfer Machine (V2)

- Multi-paths knowledge graph (location-to-surface, keypoint-to-surface and part-to-surface)
- Parameter generation



# Experiments and Results

## Dataset: DensePose-COCO

The DensePose-COCO dataset contains about 50K humans annotations, each of which is annotated with 100 UV coordinates in average. Moreover, it is split into two subsets: training set and validation set with 32K images and 1.5k images.

## Evaluation Metric: Geodesic Point Similarity (GPS)

$$GPS_j = \frac{1}{|P_j|}\sum_{p\in P_j} exp\left(\frac{-g(i_p, \hat{i}_p)^2}{2k^2}\right)$$

$P_j$: a set of ground truth points annotated on person instance $j$

$i_p$: the vertex estimated by a model at point $p$

$\hat{i}_p$: the ground truth vertex at point $p$

$k$ : 0.255

**mAP:** the mean of AP scores at a number of Geodesic Point Similarity (GPS) ranging from 0.5 to 0.95.

# Experiments and Results

## Ablation Study

| Baseline | MIDv1 | KTMv1 | $AP$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|
| √ | | | 58.8% | 55.0% | 60.2 |
| √ | √ | | 63.8% | 60.8% | 64.9% |
| √ | | √ | 61.9% | 58.0% | 63.3% |
| √ | √ | √ | **66.5%** | **61.9%** | **68.0%** |

**Vanilla version** of proposed method improves baseline model by 7.7% AP score.

Baseline + MIDv1: 58.8% -> 63.8% (+5.0%)

Baseline + KTMv1: 58.8% -> 61.9% (+3.1%)

| Baseline | MIDv2 | KTMv2 | $AP$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|
| √ | | | 58.8% | 55.0% | 60.2 |
| √ | √ | | 64.4% | 60.2% | 65.7% |
| √ | | √ | 63.4% | 61.0% | 64.8% |
| √ | √ | √ | **68.3%** | **63.8%** | **70.0%** |

**Improved version** of proposed method improves baseline model by 9.5% AP score.

Baseline + MIDv2: 58.8% -> 64.4% (+5.6%)

Baseline + KTMv2: 58.8% -> 63.4% (+4.6%)

# Experiments and Results

## The Generalizability of KTM

| Method | AP | $AP_{50}$ | $AP_{75}$ | $AP_M$ | $AP_L$ | AR | $AR_{50}$ | $AR_{75}$ | $AR_M$ | $AR_L$ |
|---|---|---|---|---|---|---|---|---|---|---|
| RCNN-based methods | | | | | | | | | | |
| DensePose R-CNN [11] | 51.9 | 85.5 | 54.7 | 39.4 | 53.9 | 61.1 | 89.7 | 65.5 | 42.0 | 62.4 |
| + KTM | $55.2^{+3.3}$ | $88.7^{+3.2}$ | $61.9^{+7.2}$ | $53.4^{+14.0}$ | $56.5^{+2.6}$ | $63.8^{+2.7}$ | $92.7^{+3.0}$ | $71.3^{+5.8}$ | $54.8^{+12.8}$ | $64.4^{+2.0}$ |
| Parsing R-CNN [5] | 58.3 | 90.1 | 66.9 | 51.8 | 61.9 | - | - | - | - | - |
| + KTM | $62.2^{+3.9}$ | $90.7^{+0.6}$ | $70.2^{+3.3}$ | $57.9^{+6.1}$ | $63.6^{+1.7}$ | 70.4 | 94.3 | 77.8 | 59.2 | 71.1 |
| AMA-net [12] | 64.1 | 91.4 | 72.9 | 59.3 | 65.3 | 71.6 | 94.7 | 79.8 | 61.3 | 72.3 |
| + KTM | $66.1^{+2.0}$ | $91.8^{+0.4}$ | $75.2^{+2.3}$ | $62.9^{+3.6}$ | $67.5^{+2.2}$ | $74.2^{+2.6}$ | $95.3^{+0.6}$ | $82.6^{+2.8}$ | $65.3^{+4.0}$ | $74.8^{+2.5}$ |
| Fully-convolutional methods | | | | | | | | | | |
| Simple [6] | 60.1 | 90.2 | 67.2 | 56.4 | 61.5 | 68.4 | 94.2 | 75.9 | 57.8 | 69.0 |
| + KTM | $62.9^{+2.8}$ | $92.5^{+2.3}$ | $73.6^{+6.4}$ | $60.7^{+4.3}$ | $63.8^{+2.3}$ | $70.2^{+1.8}$ | $95.8^{+1.6}$ | $80.5^{+4.6}$ | $62.6^{+4.8}$ | $70.7^{+1.7}$ |
| HRNet [10] | 65.1 | 92.9 | 76.8 | 62.4 | 66.2 | 72.3 | 96.1 | 83.4 | 64.5 | 72.8 |
| + KTM | $66.1^{+1.0}$ | $92.6^{-0.3}$ | $78.8^{+2.0}$ | $64.3^{+1.9}$ | $67.2^{+1.0}$ | $73.4^{+1.1}$ | $96.1^{+0.0}$ | $85.0^{+1.6}$ | $66.5^{+2.0}$ | $73.8^{+1.0}$ |

- RCNN-based methods can be improved with the help of KTM, at least 2% AP improvement.
- Fully-convolutinal methods can be improved with the help of KTM, at least 1% AP improvement.

# Summary

## Contributions

1. We propose an effective and end-to-end densepose estimation method named **knowledge transfer network (KTN)**, which addresses the issue of pyramidal representation and handles the problem of learning 2D-3D correspondences from insufficient and imbalanced labels.

2. **Multi-instances Decorder (MID)** that preserve instance details while suppressing the effect of backgrounds.

3. We are the first to introduce the knowledge to densepose estimation task. Our **Knowledge Transfer Machine (KTM)** can be easily embedded to any densepose estimation systems either from RCNN based methods or fully-convolutional frameworks.

# Summary

## Remaining challenges:

1. **Bottleneck of DensePose estimation system. (Surfaces & U coordinate)**
2. Highly Overlapping

$G_b$ denotes ground truth person mask, $G_{sp}$ is the ground truth surface mask, $G_V$ and $G_U$ are ground truth UV coordinates.

It indicates that the UV regression is the main bottleneck in densepose estimation, where the regression of V coordinates is main limitation

| Setting | | | | | Dense Pose Estimation | | |
|---|---|---|---|---|---|---|---|
| KTN-net | $G_b$ | $G_{sp}$ | $G_v$ | $G_u$ | $AP$ | $AP_{50}$ | $AP_{75}$ |
| ✓ | | | | | 68.3% | 92.1% | 77.4% |
| ✓ | ✓ | | | | 72.4% | 92.9% | 82.7% |
| ✓ | ✓ | | ✓ | ✓ | 72.7% | 93.1% | 83.1% |
| ✓ | ✓ | ✓ | | | 75.7% | 94.2% | 91.2% |
| ✓ | ✓ | ✓ | ✓ | | 80.1% | 94.3% | 92.2% |
| ✓ | ✓ | ✓ | | ✓ | 89.4% | 94.9% | 93.8% |
| ✓ | ✓ | ✓ | ✓ | ✓ | 92.1% | 94.9% | 93.8% |

# Summary

## Remaining challenges:

1. Bottleneck of DensePose estimation system. (Surfaces & U coordinate)
2. **Highly Overlapping**



# Thank you!

The code is released on GitHub:

https://github.com/stoa-xh91/HumanDensePose

If you have any questions, please e-mail us at:

wxuanhan@hotmail.com

yxzhou@std.uestc.edu.cn