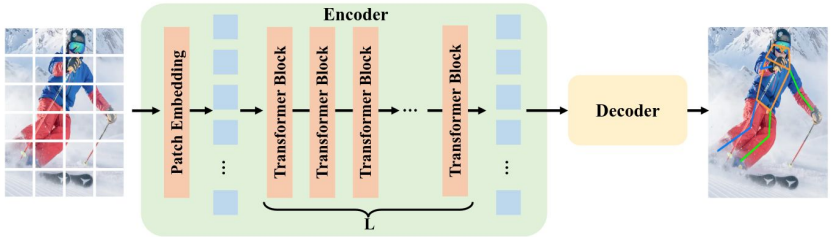
Research progress briefing

Revisiting intermediate supervision for human pose estimation

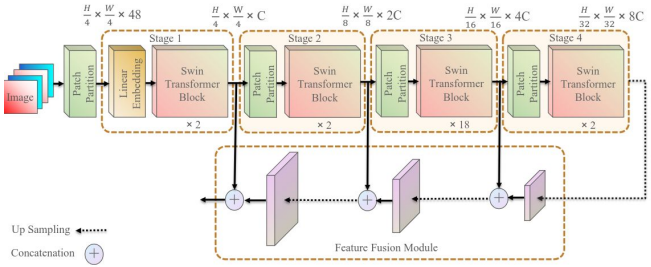
Presenter: 林保羅 (Paulo L.)
指導教授 : 傅楸善 博士
E-mail: d12922028@ntu.edu.tw

Intermediate supervision on pose estimation

Applying supervision at the very end part of a pose estimation network-- as done for ViTPose and SwinPose --has shown to reduce the expressiveness of the stages close to the input images.



ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation (NeurIPS ‘22)



Swin-Pose: Swin Transformer Based Human Pose Estimation (MIPR ‘22)

Previous work suggest that weak gradient flow may cause this underperforming scenario.

Intermediate supervision or Intermediate representation: Supervising features extracted from intermediate layers to diminish the gradient flow issue, enhance expressiveness and improve overall performance.

MIPR: IEEE International Conference on Multimedia Information Processing and Retrieval (Acceptance rate < 20%)

Intermediate supervision on pose estimation

Stacked Hourglass Networks (2016): One of the early works adopting intermediate supervision:

- A single Hourglass block (stage) plays the role of a lightweight feature extractor.
- Placing them in cascade while supervising the features at each stage revealed the power of intermediate supervision.

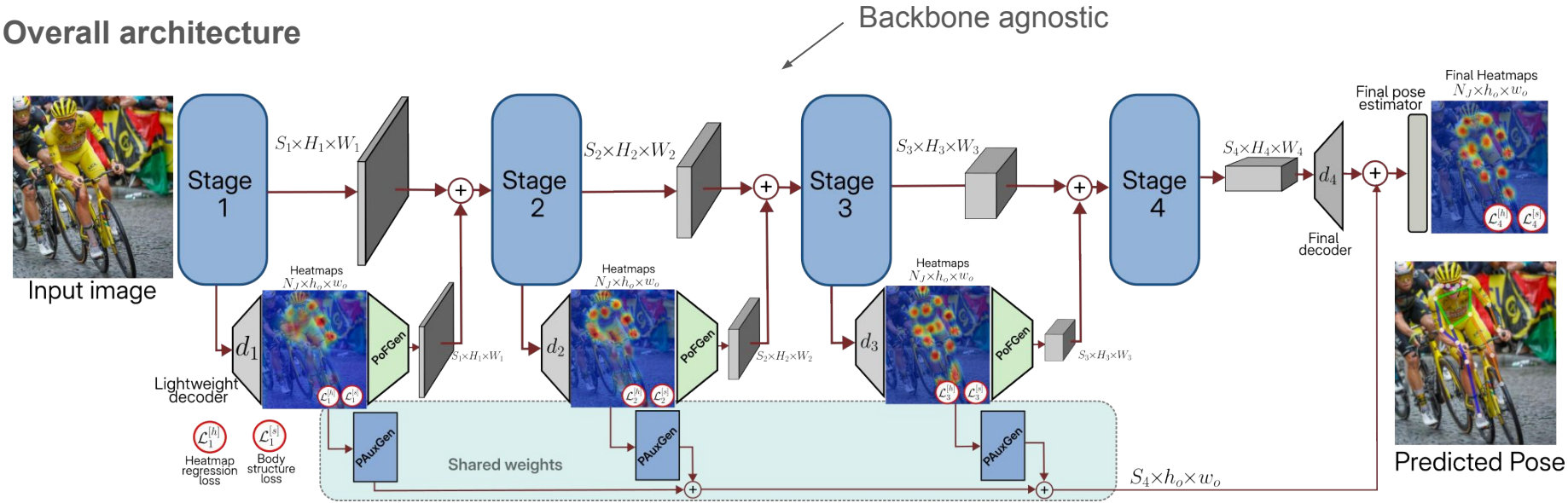
D-Pose(2024) proposed to include depth maps regression and body part segmentation blocks as a means of intermediate supervision. Thus, regression and cross-entropy based losses complement the main pose estimation objective.

PGCN (2022) introduces an intermediate dense supervision (IDS) module comprising dense classification layers to each channel of a 4D spatio-temporal volume. This module is placed right after the stem to increase its expressiveness.

Few works have been conducted on extending this strategy to other pose estimation backbones neither on supervising intermediate features within a vision backbone.

Proposed methodology

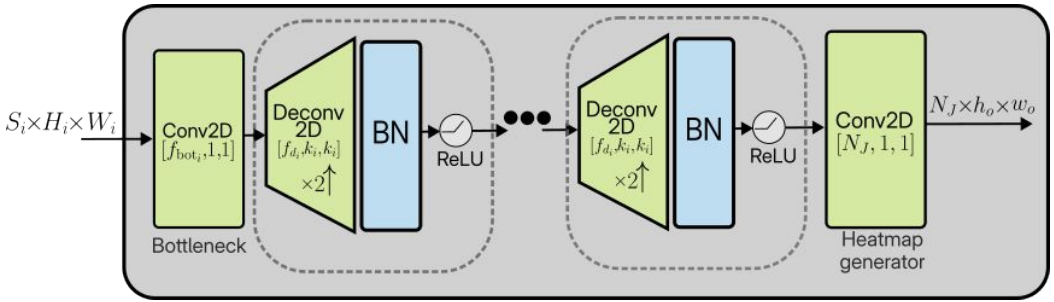
Overall architecture



- **Lightweight decoders**: Bottleneck + lightweight deconvolutions (learned upsampling) d1,d2,d3
- Pose-guided Feature Generator (**PoFGen**): Heatmaps -> Backbone feature space
- Pose-guided Auxiliary Feature Generator (**PAuxGen**): Heatmaps -> Final decoder feature space

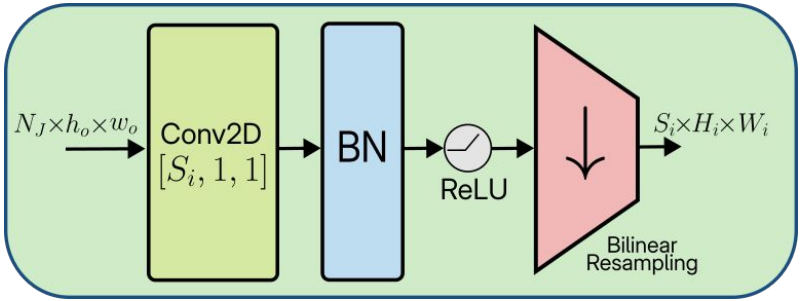
Proposed methodology

Key blocks proposed in this work

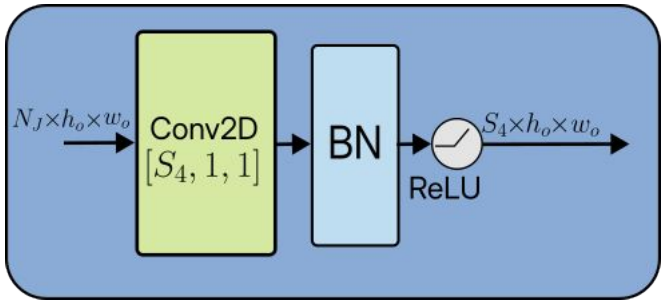


Lightweight decoder

$$S_i \gg f_{\text{bot}}$$
$$S_i \gg f_d$$



Pose-guided Feature Generator (**PoFGen**)



Pose-guided Auxiliary Feature Generator (**PAuxGen**)

Shared for all the backbone stages

Experimental results: Backbones, datasets and metrics

Baseline	Type
Stacked HG	CNN
HRNet	CNN
ResNet	CNN
VitPose	Transformer
SwinPose v1	Transformer
Uniformer	Hybrid CNN-Transformer

Backbones adopted in this work

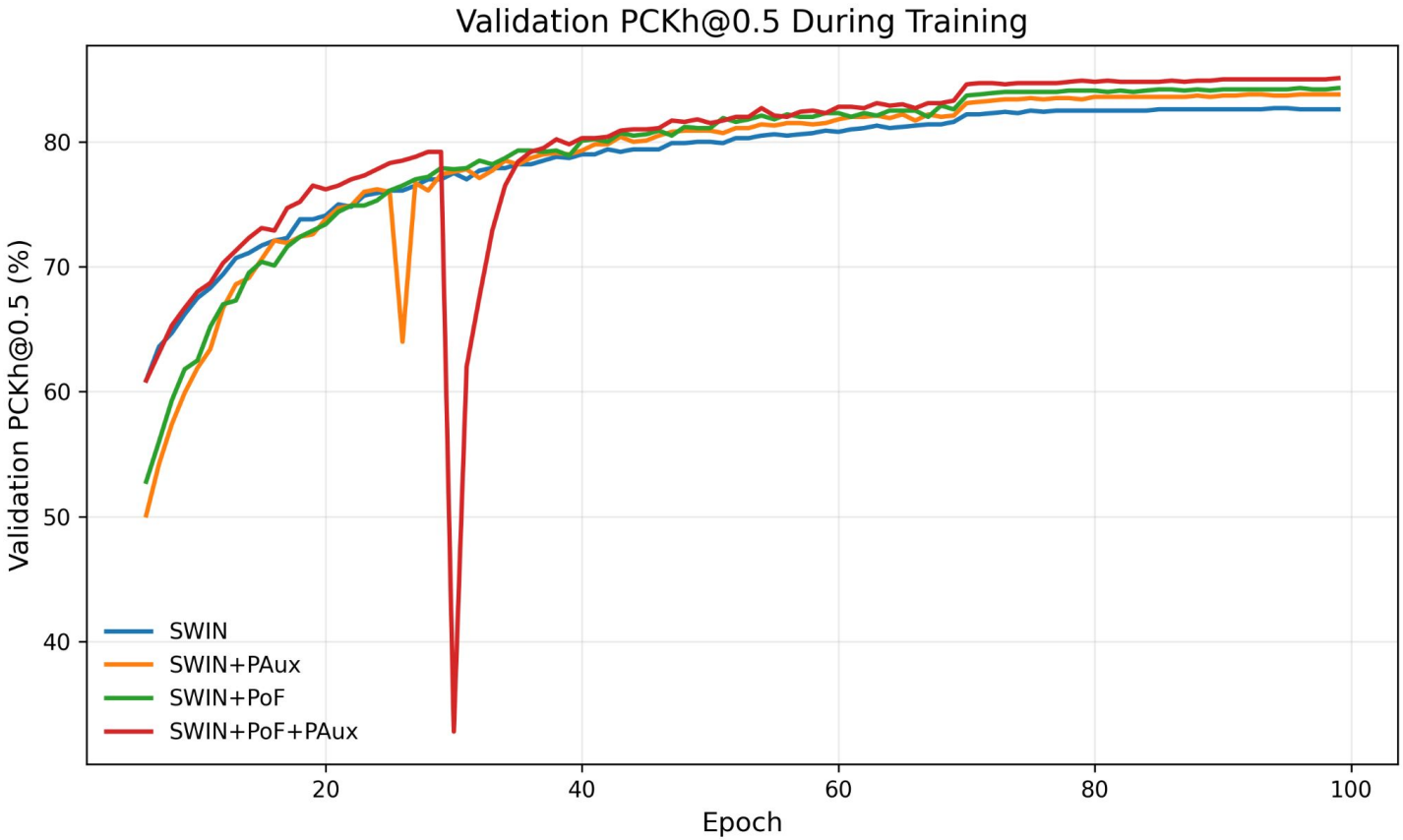
Benchmark	Modalities	Metrics	Setup	N_{joints}	Purpose
MSCOCO	RGB	mAP,OKS	Regular HuPE	17	RGB training
MPII	RGB	PCK , MPJPE	Regular HuPE	16	RGB testing
~~Human3.6M~~	~~RGB-D~~	~~PCK, MPJPE~~	~~Regular HuPE~~	–	–
MKV	RGB-D	PCK,MPJPE	Regular HuPE	18	Depth training
~~mRI~~	~~RGB-D~~	~~PCK, MPJPE~~	~~Regular HuPe~~	–	–
ITOP	Depth	PCK,MPJPE	Regular HuPE	15	Depth testing
NTU RGB+D	RGB-D	PCK,MPJE,mAP	Regular HuPE	25	Depth testing
SLP	RGB-D	PCK, MPJPE	In-bed HuPE	14	In-bed RGB-D,Depth training / testing

Selected 5 benchmarks and their metrics

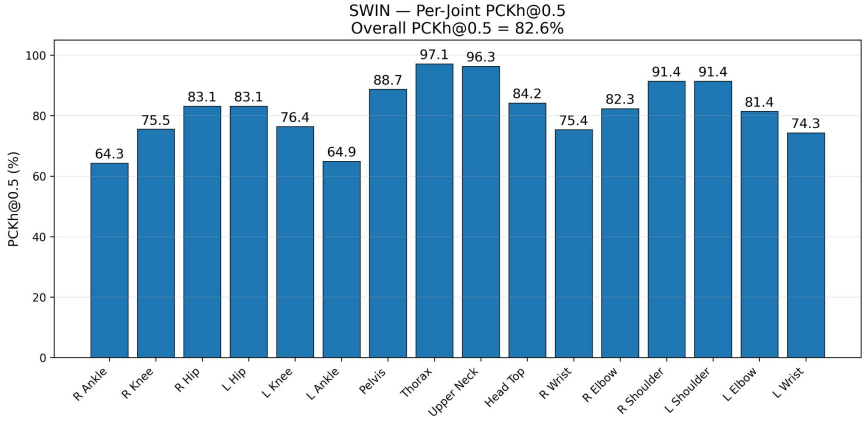
Experimental results: Ablations

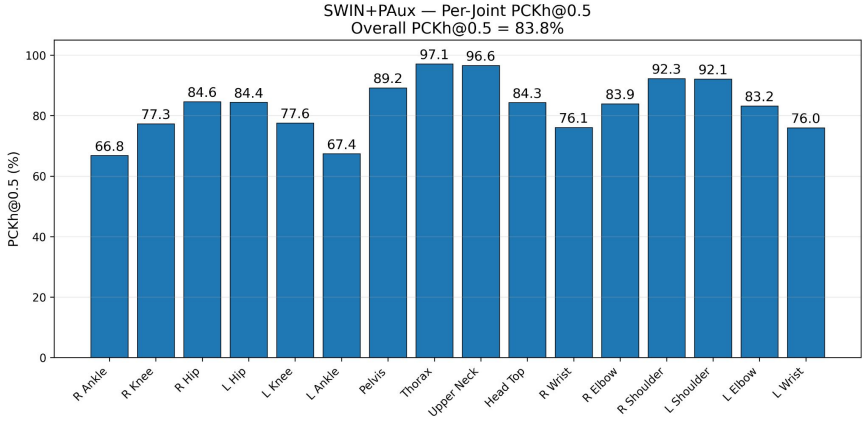
Model	Description
Backbone	Vision Backbone with a single decoder at the last stage
Backbone + PoFGen	Vision backbone with lightweight decoders and pose-guided backbone features.
Backbone + PAuxGen	Vision backbone with lightweight decoders and pose-guided auxiliary features.
Backbone + PoFGen + PAuxGen	Vision backbone with lightweight decoders, pose-guided backbone features and pose-guided auxiliary features.

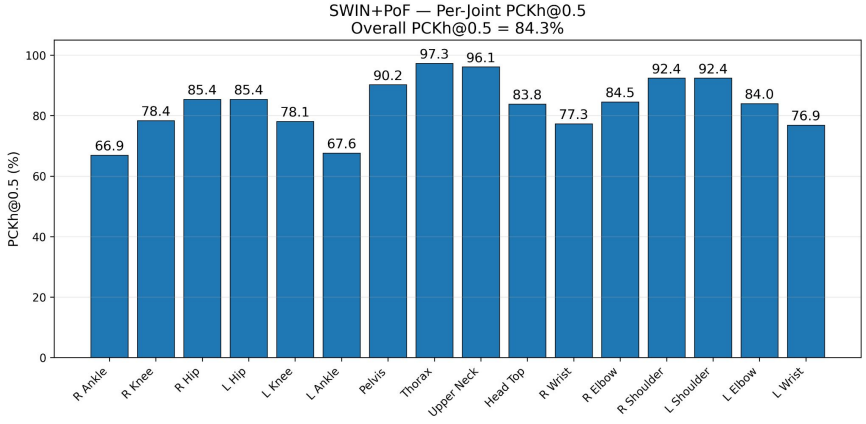
Experimental results: Ablations

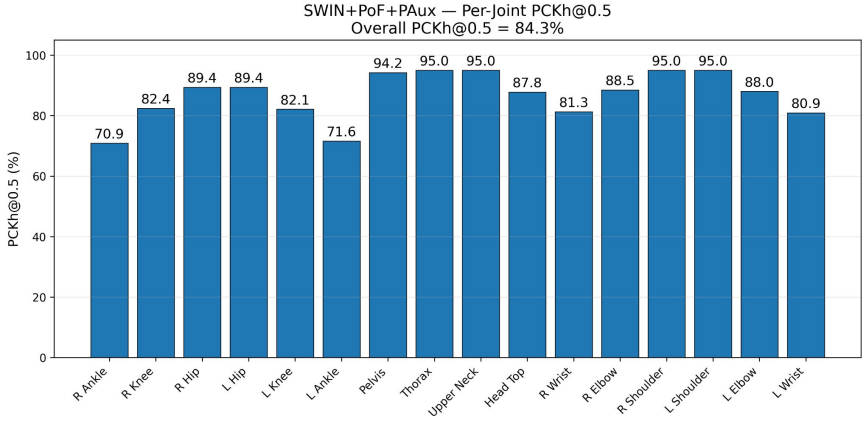


Experimental results: Ablations









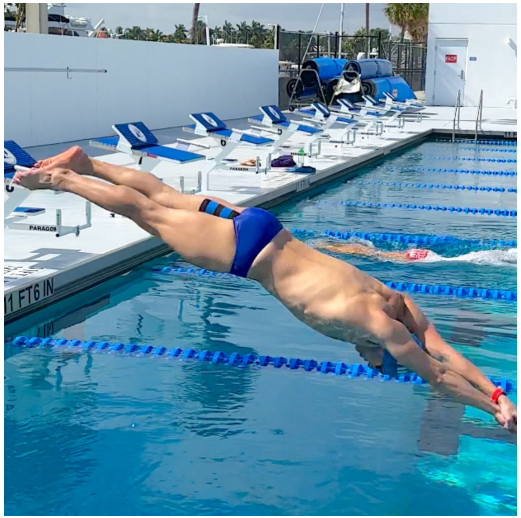
Experimental results: Qualitative results



Cyclist



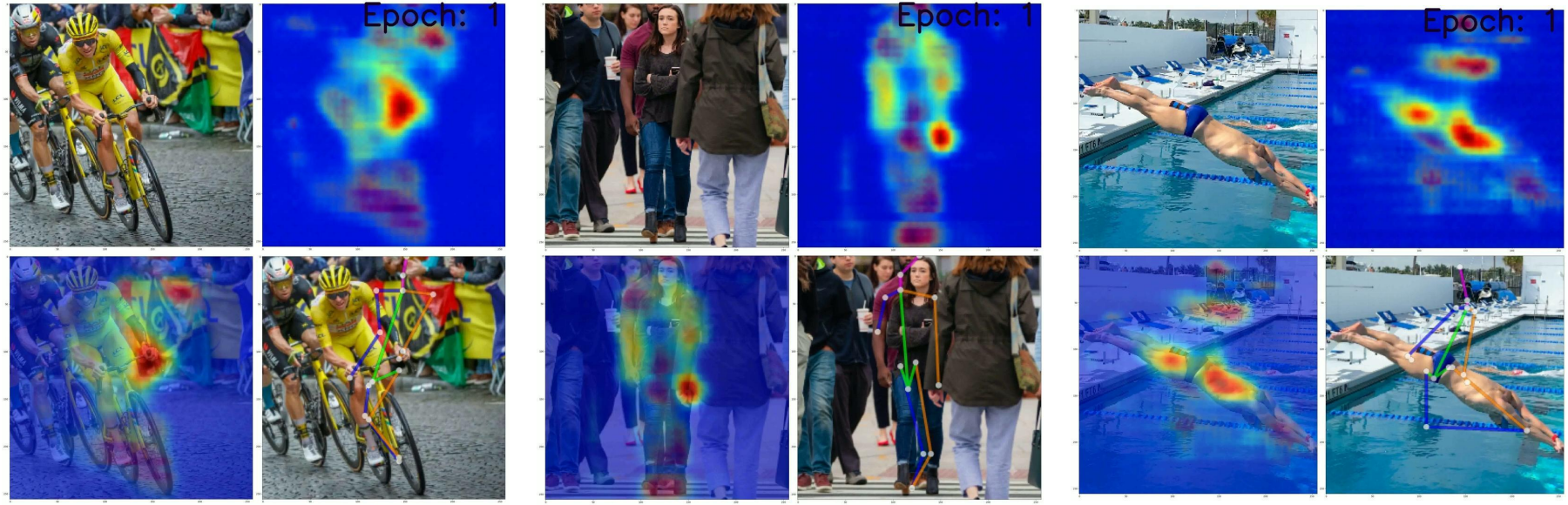
Pedestrian



Swimmer

Experimental results: Qualitative results

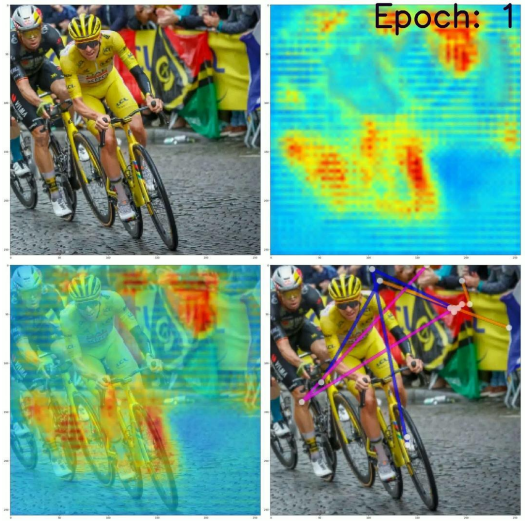
SwinPose



Cyclist Pedestrian Swimmer

Experimental results: Qualitative results

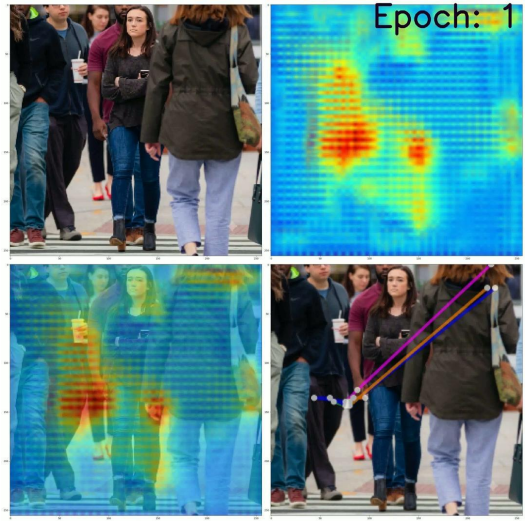
SwinPose + PoFGen + PAuxGen



Cyclist



Pedestrian



Swimmer

Conclusion

- **Two novel blocks** have been introduced to enhance the performance of HuPE backbones (backbone-agnostic) through intermediate supervision.
- The **proposed Intermediate supervision** operates along the stages comprising a vision backbone. Different from Stacked Hourglass Networks.
- **Ablations** on 6 different backbones demonstrate that the proposed blocks and intermediate supervision scheme do improve the overall performance on HuPE.
- **Experimental results** on 5 benchmarks show that the proposed methodology has a comparable performance with heavier SOTA models.

Conclusion (cont.)

Method: Incorporate intermediate supervision into vision backbones to improve performance on human pose estimation (HuPE).

Steps:

1. Train and evaluate HuPE backbones on different benchmarks (COCO, MPII).
2. Incorporate intermediate supervision blocks into the backbones.
3. Perform ablations to assess the impact of intermediate supervision blocks.
4. Compare the obtained results with heavier SOTA models.

Result: More accurate and coherent body joint localization than HuPE models using a single decoder.

Please click on the following link to see this presentation on Google slides

https://docs.google.com/presentation/d/1yIlQ5WqhtD5E6_TGJIPwK5Gb5ewS6hq90Dghk64OCQ0/edit?usp=sharing