Knowledge Distillation with Global Filters for Efficient Human Pose Estimation

902

***Kaushik Bhargav Sivangi, Fani Deligianni**

Bio-AIm Lab, School of Computing Science, University of Glasgow

***Contact:* Kaushik.Sivangi@glasgow.ac.uk**

Aim & Motivation

•**Challenge:** High-accuracy 2D-HPE models are resource intensive for real-time use, while lightweight models are faster but less accurate.

•**Need for Lightweight Models:** Devices with limited computational power require efficient 2D-HPE models that maintain performance.

•**Knowledge Distillation:** By transferring knowledge from complex to lightweight models, we can retain high accuracy while reducing computational demands.

•**Solution:** A distillation framework using Global Filter Layers (GFL) to reduce complexity, close the performance gap, and enhance speed.

Problem Formulation

- **Knowledge Distillation Setup:** The teacher network, a heatmap-based model, transfers knowledge to a lightweight student network[1], which can be either a coordinate classification or regression model.

- **Global Filter Layers (GFL):** The student network replaces self-attention modules with Global Filter Layers (GFL)[2], which operate in the frequency domain. The input tokens $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ are transformed via a 2D-FFT $\mathcal{T}(X)$ and a learnable filter K is applied in the frequency space: $\mathcal{G}(\mathcal{X}) = \mathcal{K} \odot \mathcal{T}(X)$ The output is then returned to the spatial domain using an inverse FFT: $\mathbf{X} = \mathcal{T}^{-1}(\mathcal{G}(\mathcal{X}))$ This reduces the computational complexity from $\mathrm{O}(\mathrm{H}^2\mathrm{W}^2)$ in traditional methods to $\mathrm{O}(\mathrm{HW}\log_2(\mathrm{HW}))$ significantly improving throughput.

- **Dynamic and Static Weighting Strategies:** We incorporate dynamic filters[3] in the GFL to reweight the low and high-frequency components of the input. This dynamic filter $\mathcal{K} \in \mathbb{R}^{H\times W\times F}$ is parameterized with weights learned from an MLP layer as : $\hat{\mathcal{K}} = \mathcal{K} \otimes \mathcal{M}(X)$ This allows for adaptive frequency reweighting based on the input image.

- **Loss Function:** The total loss function for the student model is computed as a combination of multiple terms, including the mean squared error (MSE) between the keypoint tokens of the teacher and student models (L_{kt}), the visual token loss (L_{vt}), and additional terms for regression or coordinate classification models:

$L_{total} = \alpha_1 L_{kt} + \alpha_2 L_{vt} + \alpha_3 L_{hm} + \alpha_4 L_{cs}$

Proposed Network & Methodology

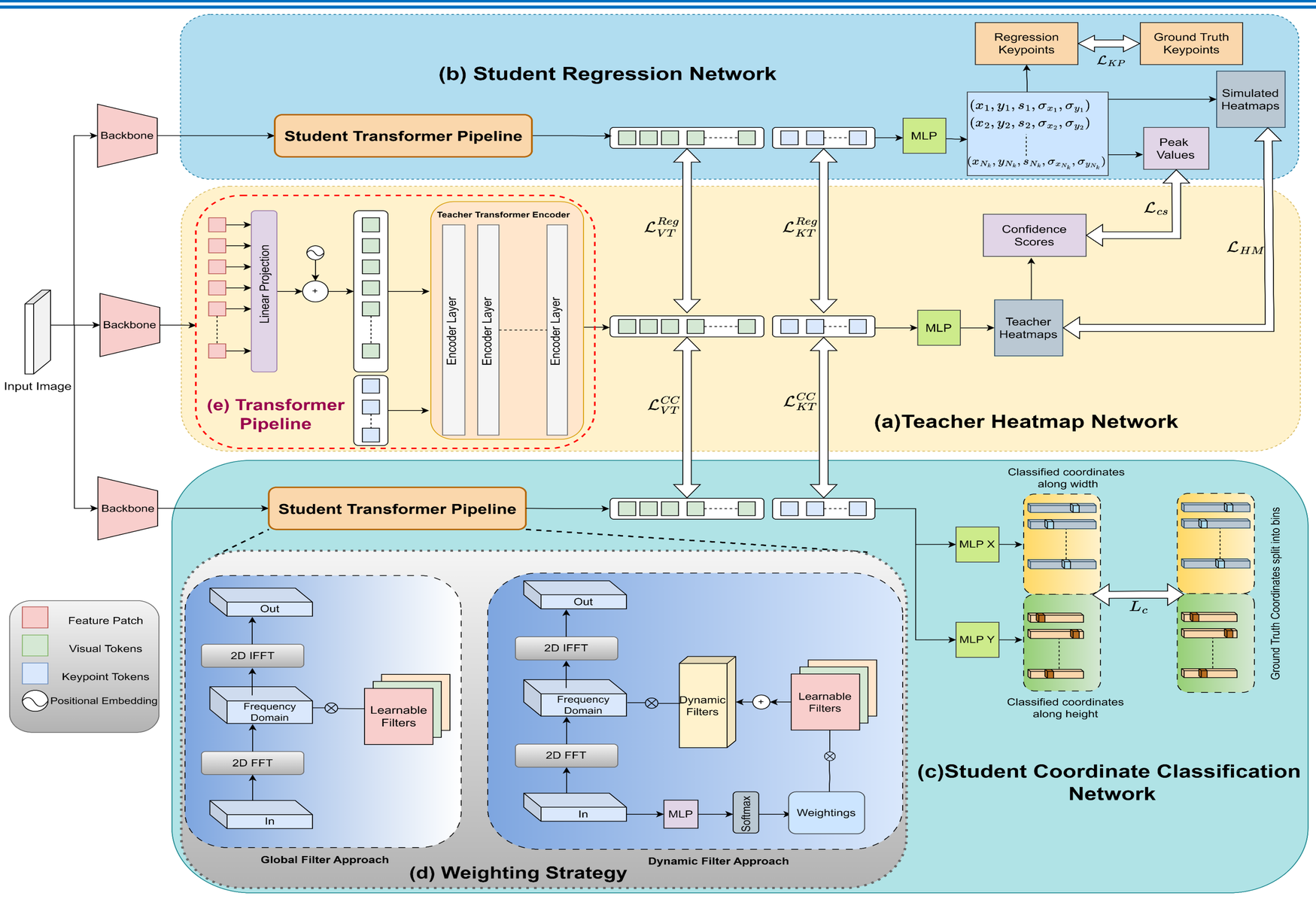

Figure A: Knowledge Distillation Approach (a) The pre-trained heatmap network acts as teacher providing visual and keypoint tokens as knowledge to student networks. (b) indicates the student regression network that uses the similar transformer pipeline as teacher.(c) represents the student coordinate classification method. (d) represents the global and dynamic weighting strategies used for the 2D-FFT token mixer which is applicable for both the students

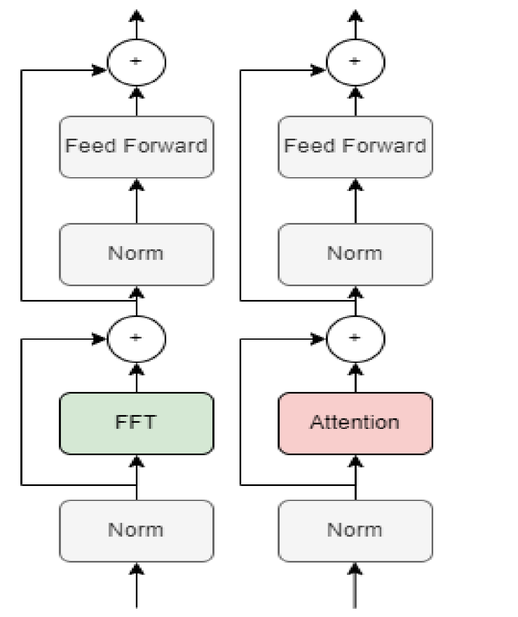

Figure B: representation of 2D-FFT and Attention based tokenmixer

Qualitative Results

Results

Model	Backbone	Weights	PCKh (%)	PCKh with Distillation(%)	Param.(M)(↓)	GFLOPs(↓)	Speed (fps)(↑)
FFT	-	GS	40.07	65.3(+25.23)	3.61	0.77	3325
		GD	48.97	76.0 (+27.03)	6.39	1.28	1560
	HRNet-W32	**GS**	**71.90**	**87.7 (+15.8)**	**11.25**	**7.13**	**645**
		GD	67.81	87.7 (+19.85)	14.03	7.64	525
	HRNet-W48	GS	71.0	87.9(+16.9)	20.3	12.9	425
Attention	-	-	46.8	79.30(+32.5)	4.72	1.6	1481
	HRNet-W32		68.39	88.73 (+20.34)	12.36	7.95	506
	HRNet-W48		72.8	88.61(+9.81)	21.4	13.8	367

Table A: Regression with Distillation on MPII. PCKh denotes PCKh@0.5(%)

Model	Backbone	Weights	PCKh	PCKh with Distillation	Param.(M)(↓)	GFLOPs(↓)	Speed (fps)(↑)
FFT	-	GS	69.2	77.87(+8.67)	3.73	0.78	3668
		GD	77.48	80.63 (+3.15)	6.51	1.28	1854
	HRNet-W32	**GS**	**88.13**	**89.40 (+1.27)**	**11.37**	**7.13**	**717**
		GD	87.67	89.25 (+1.58)	14.14	7.63	606
	HRNet-W48	GS	88.87	89.5(+0.63)	20.4	13	456
Attention	-	-	78.42	81.94(+3.52)	4.84	1.60	1831
	HRNet-W32		88.83	89.97 (+1.14)	12.48	7.96	597
	HRNet-W48		89.07	90.3(+1.23)	21.5	13.8	396

Table B: Coordinate Classification with Distillation on MPII. PCKh denotes PCKh@0.5(%)

Discussion

• **Accuracy and Speed**: The FFT-B-GS model achieves 89.4% PCKh@0.5 with an 18.26% increase in speed, demonstrating GFL's balance of accuracy and efficiency.

• **Dynamic Weighting:** FFT-NB-GD shows a 27.03% accuracy boost, proving dynamic filters adapt well to varying inputs while maintaining fast inference.

• **Real-World Use**: With a 120 FPS improvement, the framework enhances scalability, making it ideal for real-time use on resource-constrained devices without sacrificing accuracy.

Conclusion

• Our approach establishes an effective method for making high-performance 2D-HPE models more scalable and deployable in real-world settings.

• The results demonstrate that lightweight models can be enhanced without sacrificing significant accuracy, opening opportunities for future improvements in real-time applications.

1. *Ye, Suhang, et al. "DistilPose: Tokenized pose regression with heatmap distillation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.*
2. *Rao, Y., Zhao, W., Zhu, Z., Lu, J., & Zhou, J. (2021). Global filter networks for image classification. Advances in neural information processing systems, 34, 980-993.*
3. *uki Tatsunami and Masato Taki. Fft-based dynamic token mixer for vision. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 15328–15336, 2024*