Supplementary Material: Knowledge Distillation with Global Filters for Efficient Human Pose Estimation

Kaushik Bhargav Sivangi
2836578s@student.gla.ac.uk
Fani Deligianni
fani.deligianni@glasgow.ac.uk

School of Computing Science
University of Glasgow
Glasgow, UK

1 Discrete Fourier Transform

2D Discrete Fourier Transform. The Discrete Fourier Transform (DFT) in the signal processing domain converts the discrete data from time domain into the frequency domain. This is crucial as it provides underlying characteristics of the data by analyzing the signals in terms of frequency components. Let $f(n)$ represent a 1-Dimensional discrete sequence of length N, where n represents the sample points in discrete time-domain and $0 \leq n \leq N-1$. Let the set of basis vectors be given as $b_k(n)$ then, the forward transformation is given as:

$$\hat{f}(k) = \sum_{n=0}^{N-1} f(n) b_k(n) : 0 \leq k \leq N-1 \tag{1}$$

where the basis vectors is given by the complex sinusoid $b_k(n) = e^{(j2\pi \frac{k}{N} n)}$, and j represents the imaginary unit. Since the transformation is one-to-one, the inverse transform (IDFT) can be recovered as:

$$f(n) = \frac{1}{N} \sum_{k=0}^{N-1} \hat{f}(k) e^{(j2\pi \frac{k}{N} n)} : 0 \leq n \leq N-1 \tag{2}$$

This can be further generalized to a 2D signal (2D-DFT) $f(h,w)$ as:

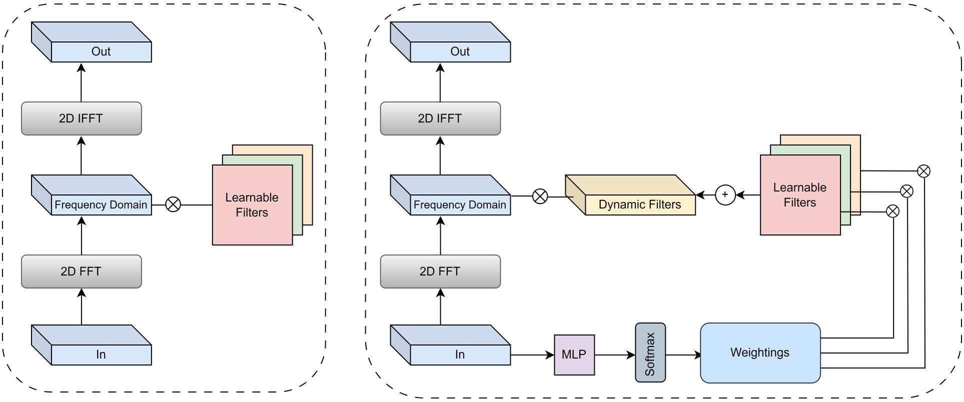
$$\hat{f}(\overline{h}, \overline{w}) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} f(h,w) e^{j2\pi(\frac{h\overline{h}}{H} + \frac{w\overline{w}}{W})} \tag{3}$$

where $h, \overline{h} \in \{z \in \mathbb{Z} \mid 0 < z < H\}$;$w, \overline{w} \in \{z \in \mathbb{Z} \mid 0 < z < W\}$ and $H, W \in \mathbb{N}$[3]. For our experiments we specifically use the 2D-Fast Fourier Transform (2D-FFT) to calculate the 2D-DFT, owing to its lower complexity.

2 Dynamic Filters

As noted in the main paper, we introduce the concept of dynamic filters in contrast to the static filters utilized by the global filter layer depicted in Figure 1.

© 2024. The copyright of this document resides with its authors.



(a) Global Filter Approach

(b) Dynamic Filter Approach

Figure 1: Weighting Strategies for the Global Filter Layer

While the global filter layer relies on fixed filters during inference, making it image agnostic, the dynamic filter approach enables the reweighting of both low and high-frequency components of the input image. Similar to global filter approach, the input tokens $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ are first transformed along spatial dimensions to the frequency domain. The resulting tensor is then used to perform element-wise multiplication with a dynamic filter $\hat{\mathcal{K}}$. The input $\mathbf{X}$ is reshaped along the spatial dimension by performing a mean operation, resulting in a tensor of shape $\mathbb{R}^{C}$. This tensor is further reshaped by decomposing the dimension C into filters F and spatial dimension S, resulting in $\hat{X} \in \mathbb{R}^{F\times S}$. This is propagated through an MLP layer $\mathcal{M}(.)$ with an output shape of C, finally resulting in a tensor $\mathcal{M}(\hat{X})$ of shape $\mathbb{R}^{F\times C}$ with softmax layer in the end, representing the dynamic spectrum weights as follows:

$$\hat{X} = Reshape(mean\_spatial\_dimension(X)) \in \mathbb{R}^{F\times S} \tag{4}$$

$$\mathcal{M}(\hat{\mathcal{X}}) = softmax(W_2 \cdot StarReLU(W_1(\hat{X}))) \in \mathbb{R}^{F\times C} \tag{5}$$

Similar to the previous case of GFL with static weights, the learnable filters $\mathcal{K} \in \mathbb{R}^{H\times W\times F}$ are generated. These learnable filters are weighted with the generated dynamic spectrum weights to generate dynamic filter $\hat{\mathcal{K}} \in \mathbb{R}^{C\times H\times W}$ formulated as:

$$\hat{\mathcal{K}} = \mathcal{K} \otimes \mathcal{M}(\hat{\mathcal{X}}) \in \mathbb{R}^{C\times H\times W} \tag{6}$$

Finally, the dynamic filter is coupled with the transformed input tokens through element-wise multiplication and is reverted back to spatial domain through inverse transform which is given by:

$$\mathcal{D}(\mathcal{X}) = \hat{\mathcal{K}} \odot \mathcal{X} \tag{7}$$

$$\mathbf{X} = \mathcal{T}^{-1}(\mathcal{D}(\mathcal{X})) \tag{8}$$

3 Implementation Details

Training. For our experiments We set the image resolution to 256x256. In case of experiments without backbone, the patch size $P_H \times P_W$ for visual tokens is 16×16; when a

It may be distributed unchanged freely in print or electronic forms.

backbone is employed, the feature maps are of size 64×64 and patch size is 4×4. The embedding dimension d is set to 192. The number of keypoint tokens N_k is set to 16, corresponding to the total number of joints. It is to be noted that the flip test has been used during evaluation.

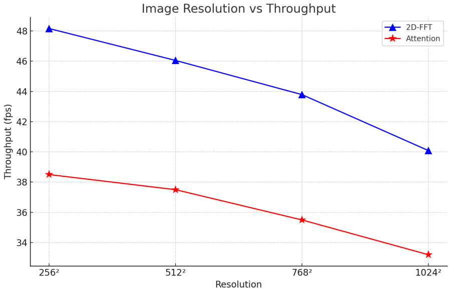
We mainly follow the training settings from [2]. We train our models for 300 epochs with an Adam optimizer starting with an initial learning rate of 1e-3 with rate reduction at 200th and 260th epochs to 1e-4 and 1e-5 respectively. We use dual Nvidia 3090 GPUs with a batch size of 32 per GPU. For both the experiments we adopt $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = 0.0005$, $\alpha_3 = \beta_3 = 1$, and $\alpha_4 = 0.01$ [4].

4 Throughput Across Various Resolutions

Theoretical Complexity: The complexity of the 2D-FFT layer is given as $\mathcal{O}(HWClog_2HW)$. Additionally, in the case of GFL-Static, there is a complexity of $\mathcal{O}(HWC)$ from element-wise multiplication with the filters, which results in the total complexity of $\mathcal{O}(HWClog_2HW + HWC)$. In case of GFL-Dynamic, we should also consider the use of two feed-forward layers with weights $W_1 \in \mathbb{R}^{l \times F}$ and $W_2 \in \mathbb{R}^{F \times l}$, resulting in an increased complexity of $\mathcal{O}(lF^2C + l^2FC)$ where l is the hidden dimension of the feed-forward layer.

We further calculate the throughput of transformer only models on a CPU (without backbone) on CPU to observe the effect of using 2D-FFT and Attention based tokenmixer. Our calculations are based on batch size of 1 on an Intel Xeon Silver 4210R CPU. Figure 2 illustrates the impact of image resolution on the throughput of FFT-based and attention-based methods in the context of coordinate-classification models. As the resolution increases from 256×256 to 1024×1024, the 2D-FFT-based models consistently outperform the attention model. At the lowest resolution, the 2D-FFT method processes images at approximately 48 frames per second (fps), while the attention model processes at 38 fps. As the resolution increases, the performance of the attention method declines sharply, highlighting the efficiency of the 2D-FFT model in handling larger resolutions and maintaining higher fps.

Figure 2: Resolution vs Throughput for Coordinate Classification models



5 Additional Experimental Results with Discrete Cosine Transform

We further perform our experiments by using Discrete Cosine Transform (DCT) as a token mixer[1]. Table 1 represents the performance metrics of the DCT model variants evaluated on MPII dataset. Overall, distillation improves the accuracy in comparison to non-distilled versions. DCT-NB-GS achieved a PCKh of 67.64% which significantly improved to 76.21% when distillation is applied. This model also has the lowest GFLOPs (0.79) and highest speed of 1337fps. However, in contrast, we observe DCT-NB-GD has lesser improvement with distillation through its base performance is higher than DCT-NB-GS. Variants with HRNet-W32 backbone, specifically, DCT-B-GS produced a high PCKh of 88.93%, marginally improving to 89.16% with distillation. DCT-GD-B achieves the highest performance of 89.58% with more parameters and computation.

Model	**Backbone**	**Weights**	**PCKh (%)**	**PCKh with Distillation(%)**	**Param.(M)(↓)**	**GFLOPs(↓)**	**Speed (fps)(↑)**
DCT	-	GS	67.64	76.21	3.69	0.79	1337
		GD	69.36	72.77	6.47	1.29	711
	HRNet-W32	GS	88.93	89.16	11.33	7.14	547
		GD	88.0	89.58	14.11	7.64	400

Table 1: Coordinate Classification with Distillation and DCT token mixer.

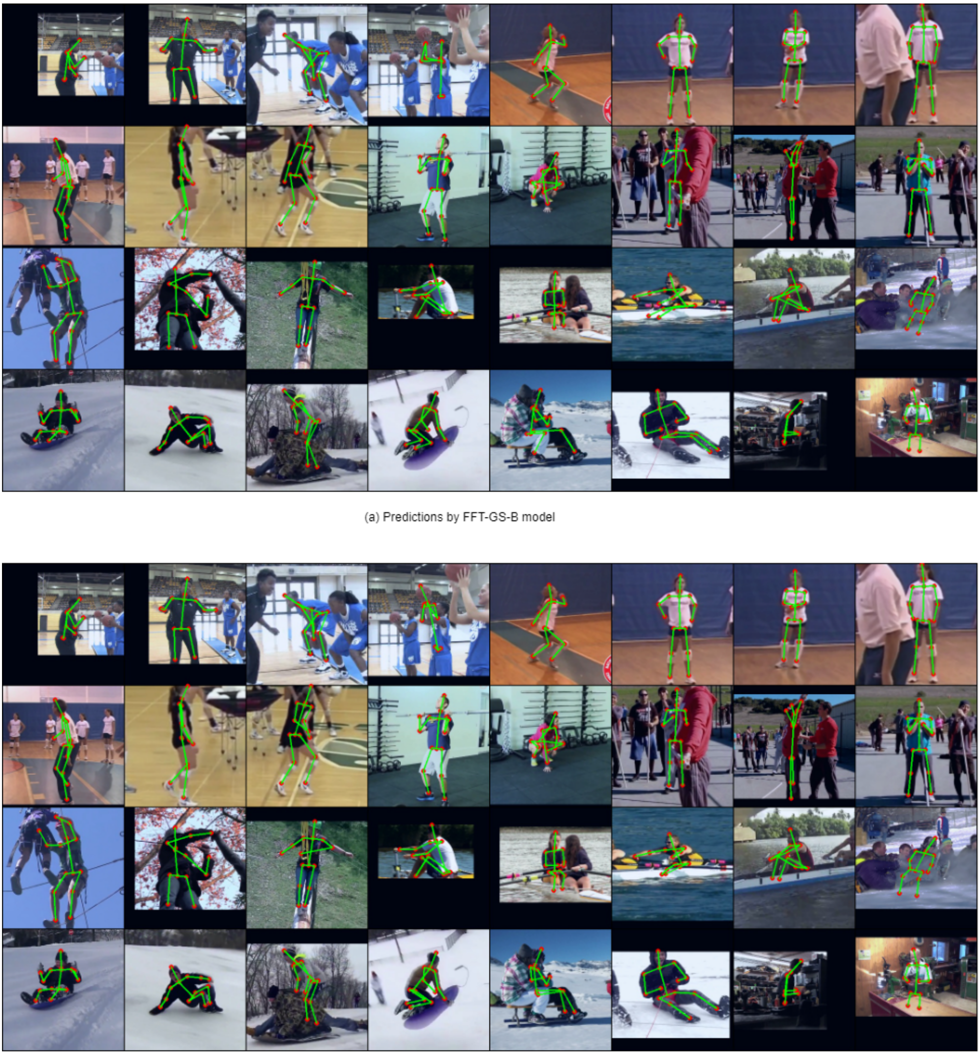
Similarly, Table 2 represents the usage of DCT module for regression based models. Overall, distillation has shown performance improvement compared to training the student independently. However, the speed is relatively low. Unlike the cases of GFL-Dynamic layer in FFT based token mixer models where the accuracy significantly increases in no distillation cases, in case of DCT (DCT-GD-* models) declination is observed. In other words, the PCKh values drop significantly when GFL-Dynamic layer is incorporated. This suggests that the efficiency of GFL-Dynamic is limited to FFT-based token mixers.

Model	**Backbone**	**Weights**	**PCKh (%)**	**PCKh with Distillation(%)**	**Param.(M)(↓)**	**GFLOPs(↓)**	**Speed (fps)(↑)**
DCT	-	GS	49.82	64.45	3.58	0.78	1160
		GD	36.47	67.3	6.35	1.28	590
	HRNet-W32	GS	70.58	87.73	11.21	7.14	470
		GD	65.94	87.09	13.99	7.64	338

Table 2: Regression with Distillation and DCT token mixer.

6 Qualitative Results

We represent the prediction of our proposed models on a sample of images from MPII validation set. Figure 3 shows the predictions by coordinate classification models by both FFT-GS-B and ATTN-B models both being the distilled versions. Similarly Figure 4 represents the predictions made by regression based models. Additionally, Figure 5 represents the ground truth keypoints.



(a) Predictions by FFT-GS-B model

(b) Predictions by ATTN-B model

Figure 3: Predictions of Coordinate Classification based model. (a) representation of keypoint predictions by 2D-FFT based token mixer. (b) representation of keypoint predictions by Attention based token mixer

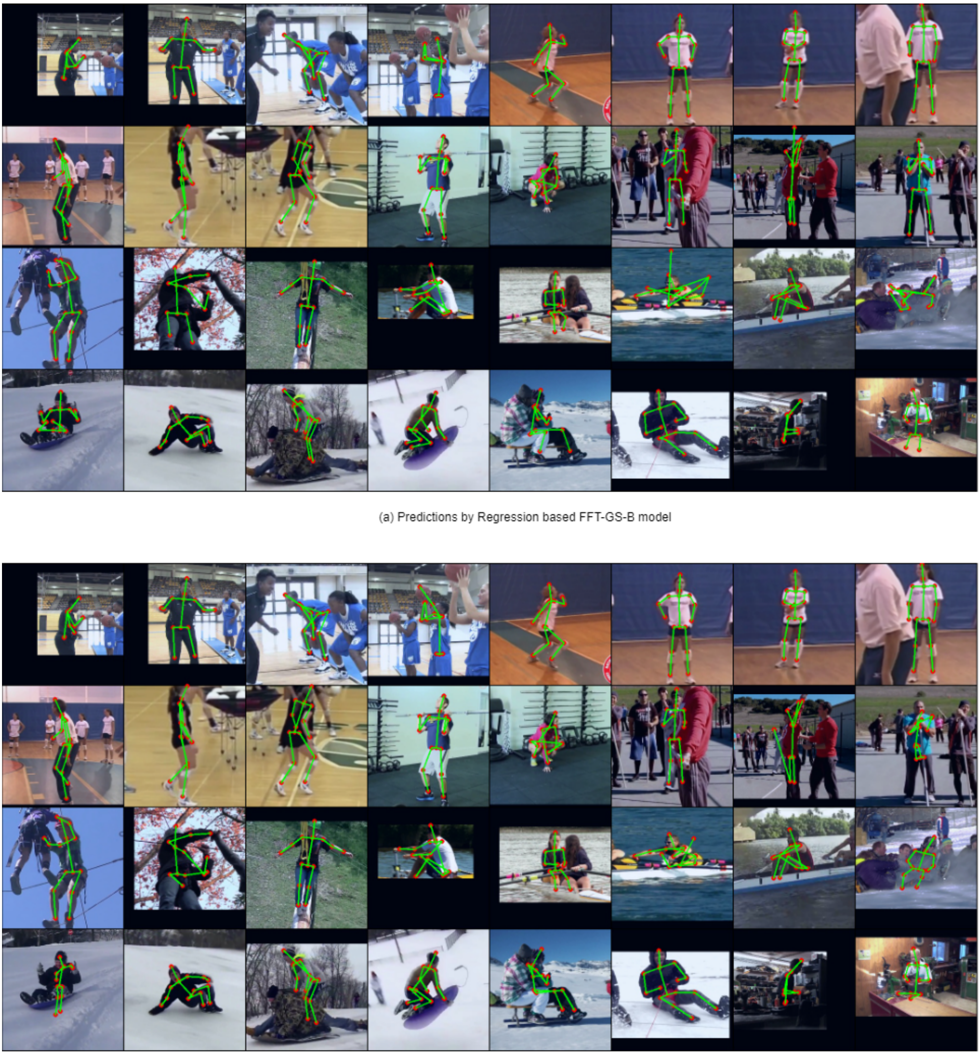
References

[1] Zhiqiang Hu and Tao Yu. Dynamic spectrum mixer for visual recognition. *arXiv preprint arXiv:2309.06721*, 2023.

[2] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 11313–11322, 2021.

[3] Yuki 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.

[4] Suhang Ye, Yingyi Zhang, Jie Hu, Liujuan Cao, Shengchuan Zhang, Lei Shen, Jun Wang, Shouhong Ding, and Rongrong Ji. Distilpose: Tokenized pose regression with heatmap distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2163–2172, 2023.



(a) Predictions by Regression based FFT-GS-B model

(b) Predictions by Regression based ATTN-B model

Figure 4: Predictions of Regression based model. (a) representation of keypoint predictions by 2D-FFT based token mixer. (b) representation of keypoint predictions by Attention based token mixerl

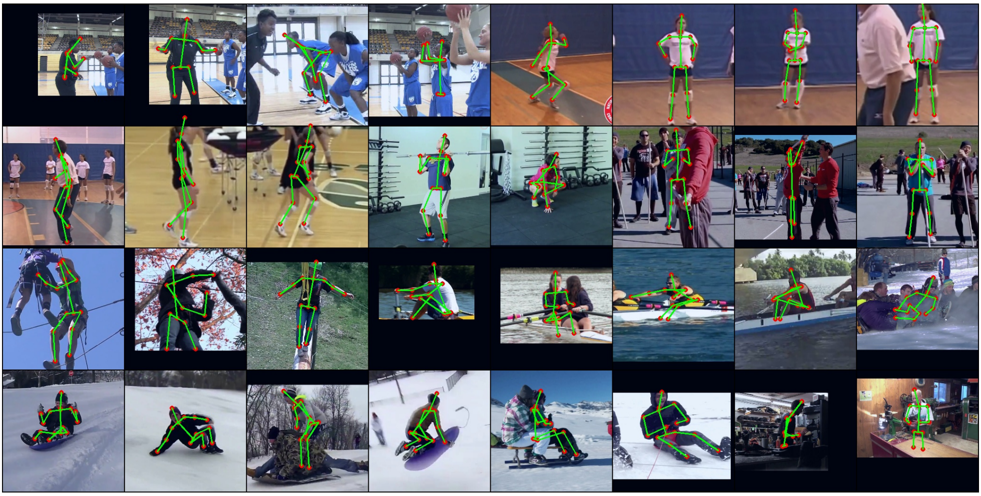


Figure 5: Ground Truth Keypoints