SUPPLEMENTARY MATERIAL

1 Fingertip Tracking Model Evaluation

To evaluate the performance of the model, we calculated the End Point Error (EPE) between predicted point position and real point position, on the testing dataset for different neural network architectures. We already showed comparisons between ResNet50, EfficientNetB0 and MobileNetV2 in the main paper. We evaluated a range of models, in addition to the three presented in the main paper, by adding in different versions of ResNet, EfficientNet and MobileNet for comparison. We show fingertip tracking results of ResNet50, ResNet101, ResNet152 [1], EfficientNetB0, EfficientNetB3, EfficientNetB6 [2], MobileNetV1 and MobileNetV2 [3] in Table 1. We also showed the training loss error vs epoch graph in Figure 1.

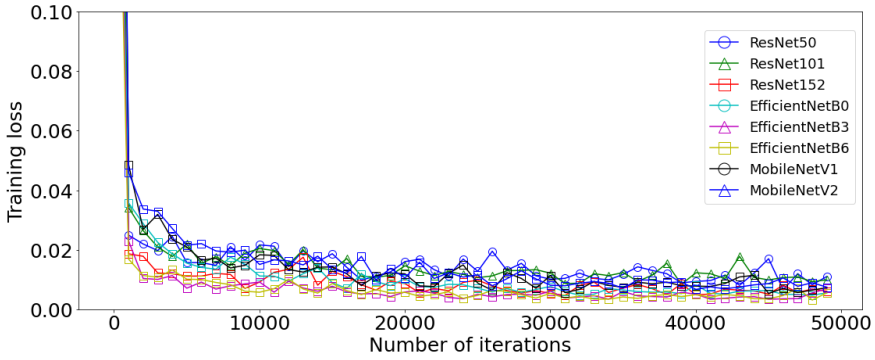

Figure 1: The training loss vs number of iterations during the training process.

Table 1: End Point Error (EPE) for the three deep learning neural networks.

Deep learning neural network	EPE on training dataset	EPE on testing dataset
ResNet50	2.98 pixels	3.00 pixels
ResNet101	2.46 pixels	2.50 pixels
ResNet152	1.53 pixels	1.60 pixels
EfficientNetB0	1.52 pixels	1.53 pixels
EfficientNetB3	1.14 pixels	1.15 pixels
EfficientNetB6	1.16 pixels	1.15 pixels
MobileNetV1	2.31 pixels	2.35 pixels
MobileNetV2	2.21 pixels	2.30 pixels

The fingertip tracking results of different neural network architectures are overall very good (with EPE in a range from 1.14 pixels to 3.00 pixels).

To evaluate the complexity of the model, we show the EPE vs number of parameters plot for different networks in Figure 2 to see the efficiency of different models.

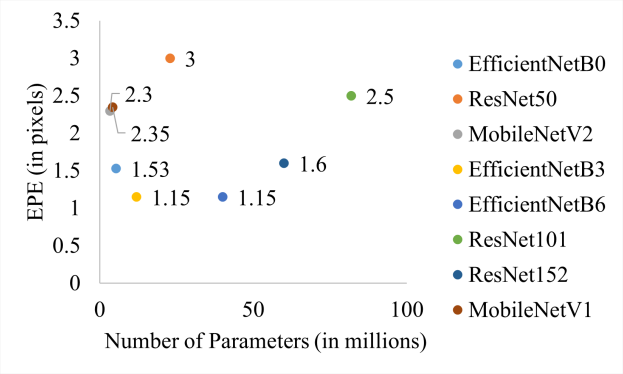

Figure 2: The performance (measured in EPE) vs number of parameters for different networks.

2 Feature Extraction Results

2.1 Validation of Computer Vision Methods Compared to The Gold Standard System

The comparison between features (M-TF and Var-TF) extracted from a range of computer vision methods and from the Optotrak method is summarized in Table 2. The computer vision methods progressively under-estimated the tapping frequencies with fast movements, giving falsely low measures of frequency compared to the benchmark.

The M-TF and Var-TF measures extracted from the Optotrak method compared to a range of computer vision methods' are presented as scatter plots in Figure 3 (the Optotrak vs ResNet101/ResNet152/MobleNetV1) and Figure 4 (the Optotrak vs EfficientNetB3/EfficientNetB6). Figure 5 and 6 are Bland Altman plots comparing the Optotrak measures of M-TF and Var-TF compared to the computer vision measures using ResNet101/ResNet152/MobileNetV1 and EfficientNetB3/EfficientNetB6 respectively.

2.2 Reliability of Computer Vision Methods at Two Different Distances from Camera

We show the comparison between features (M-TF and Var-TF) extracted in Near-To-Laptop condition and Far-From-Laptop condition by different computer vision methods and the Optotrak method in Table 3, and the Bland Altman plots in Figure 7 (ResNet101/ResNet152/MobileNetV2) and Figure 8 (EfficientNetB3/EfficientNetB6).

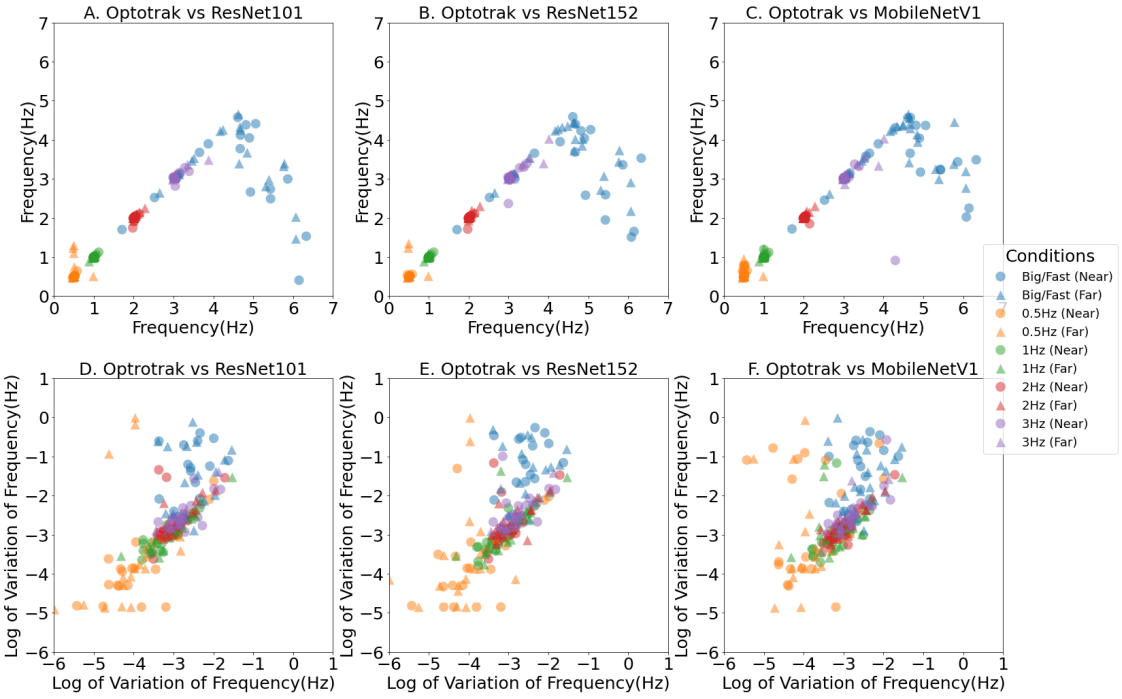

Figure 3: **Tapping Frequency x-y scatter plot between ResNet101/ResNet152/MobileNetV1 (y axis) and the Optotrak method (x axis).** A, B and C show the scatter plots of mean tapping frequency. D, E and F show the scatter plots of the logarithm of variation in the tapping frequency. The colored marks represent the different finger tapping conditions with blue denoting the 'Big/Fast' self-paced conditions, and yellow, green, red and purple the externally paced conditions at frequencies of 0.5Hz, 1Hz, 2Hz and 3Hz respectively. Circles are for conditions performed Near-To-Laptop (60cm) and triangles are for conditions performed Far-From-Laptop (80cm).

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[2] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[3] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

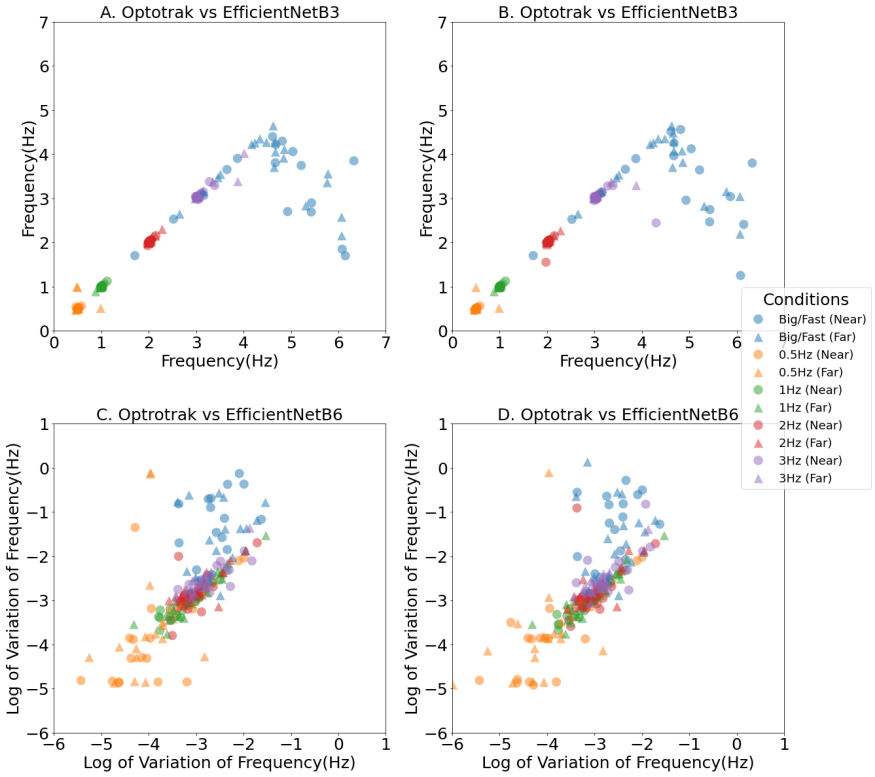

Figure 4: **Tapping Frequency x-y scatter plot between EfficientNetB3/EfficientNetB6 (y axis) and the Optotrak method (x axis).** A and B show the scatter plots of mean tapping frequency. C and D show the scatter plots of the logarithm of variation in the tapping frequency. The colored marks represent the different finger tapping conditions with blue denoting the 'Big/Fast' self-paced conditions, and yellow, green, red and purple the externally paced conditions at frequencies of 0.5Hz, 1Hz, 2Hz and 3Hz respectively. Circles are for conditions performed Near-To-Laptop (60cm) and triangles are for conditions performed Far-From-Laptop (80cm).

Table 2: Accuracy of each computer vision method compared with the Optotrak measures

Methods	M-TF t-value (p-value)	Var-TF t-value (p-value)
'As fast as possible' condition with **frequency < 4Hz**		
ResNet50	0.52 (0.48)	1.48 (0.24)
ResNet101	1.55 (0.22)	1.34 (0.26)
ResNet152	0.73 (0.48)	1.39 (0.20)
EfficientNetB0	0.72 (0.47)	2.60 (0.01)
EfficientNetB3	0.17 (0.87)	1.57 (0.56)
EfficientNetB6	0.15 (0.88)	1.07 (0.32)
MobileNetV1	0.01 (0.99)	1.41 (0.17)
MobileNetV2	0.39 (0.54)	1.89 (0.19)
'As fast as possible' condition with **frequency > 4Hz**		
ResNet50	107.61 (0)	59.40 (0)
ResNet101	94.9 (0)	44.50 (0)
ResNet152	7.94 (0)	7.10 (0)
EfficientNetB0	7.44 (0)	6.57 (0)
EfficientBetB3	24.44 (0)	5.72 (0)
EfficientNetB6	7.52 (0)	5.74 (0)
MobileNetV1	7.76 (0)	6.36 (0)
MobileNetV2	109.67 (0)	41.56 (0)
0.5Hz condition		
ResNet50	0.39 (0.53)	6.41 (0.02)
ResNet101	2.11 (0.15)	4.11 (0.05)
ResNet152	0.95 (0.35)	1.64 (0.11)
EfficientNetB0	0.84 (0.40)	1.81 (0.08)
EfficientNetB3	0.65 (0.52)	1.64 (0.12)
EfficientNetB6	0.04 (0.97)	1.09 (0.28)
MobileNetV1	2.22 (0.03)	3.33 (0)
MobileNetV2	0.02 (0.88)	7.63 (0.01)
1Hz condition		
ResNet50	0.02 (0.90)	3.40 (0.70)
ResNet101	0 (0.97)	0.44 (0.51)
ResNet152	0.31 (0.76)	1.21 (0.23)
EfficientNetB0	0.02 (0.99)	0.58 (0.56)
EfficientNetB3	0 (1)	0.76 (0.45)
EfficientNetB6	0.01 (0.99)	0.71 (0.48)
MobileNetV1	1.16 (0.25)	1.78 (0.08)
MobileNetV2	0.05 (0.82)	3.63 (0.06)
2Hz condition		
ResNet50	0.02 (0.89)	6.10 (0.02)
ResNet101	0.48 (0.49)	6.25 (0.02)
ResNet152	0.31 (0.75)	2.06 (0.04)
EfficientNetB0	0.93 (0.36)	1.43 (0.16)
EfficientNetB3	0.09 (0.93)	1.59 (0.11)
EfficientNetB6	0.67 (0.51)	1.60 (0.12)
MobileNetV1	0.13 (0.90)	1.60 (0.11)
MobileNetV2	0.01 (0.92)	4.94 (0.03)
3Hz condition		
ResNet50	4.26 (0.04)	3.92 (0.05)
ResNet101	2.85 (0.1)	4.64 (0.04)
ResNet152	0.80 (0.43)	2.43 (0.02)
EfficientNetB0	1.61 (0.11)	1.79 (0.08)
EFficientNetB3	0.91 (0.37)	1.93 (0.06)
EfficientNetB6	2.09 (0.04)	2.00 (0.05)
MobileNetV1	1.24 (0.22)	2.08 (0.04)
MobileNetV2	2.16 (0.15)	3.63 (0.06)

Table 3: Difference between Near-To-Laptop and Far-From-Laptop computer measures compared with the Optotrak

Methods	M-TF t-value (p-value)	Var-TF t-value (p-value)
'As fast as possible' condition		
Optotrak	0 (1)	0.43 (0.52)
ResNet50	0.55 (0.46)	0.36 (0.55)
ResNet101	0.55 (0.46)	0 (0.97)
ResNet152	1.57 (0.13)	0.92 (0.37)
EfficientNetB0	1.86 (0.07)	0.45 (0.66)
EfficientNetB3	0.51 (0.62)	0.32 (0.75)
EfficientNetB6	1.60 (0.12)	0.60 (0.55)
MobileNetV1	1.50 (0.14)	0.33 (0.74)
MobileNetV2	3.63 (0.07)	0.94 (0.34)
0.5Hz condition		
Optotrak	0.48 (0.5)	2.64 (0.12)
ResNet50	1.07 (0.31)	0.64 (0.43)
ResNet101	3.31 (0.08)	3.09 (0.09)
ResNet152	1.30 (0.21)	0.92 (0.37)
EfficientNetB0	1.50 (0.15)	1.33 (0.20)
EfficientNetB3	0.69 (0.50)	1.65 (0.11)
EfficientNetB6	0.82 (0.42)	0.66 (0.52)
MobileNetV1	1.18 (0.24)	0.98 (0.33)
MobileNetV2	1.07 (0.31)	0.13 (0.72)
1Hz condition.		
Optotrak	2.87 (0.1)	1.71 (0.2)
ResNet50	2.16 (0.15)	0.29 (0.6)
ResNet101	1.58 (0.22)	1.22 (0.28)
ResNet152	0.74 (0.46)	1.40 (0.17)
EfficientNetB0	1.39 (0.17)	1.30 (0.20)
EfficientNetB3	1.59 (0.12)	1.39 (0.18)
EfficientNetB6	1.31 (0.20)	1.41 (0.17)
MobileNetV1	0.49 (0.62)	1.12 (0.27)
MobileNetV2	1.76 (0.19)	0.45 (0.51)
2Hz condition.		
Optotrak	0.84 (0.37)	0.52 (0.48)
ResNet50	1.66 (0.21)	0.03 (0.85)
ResNet101	1.9 (0.18)	0.02 (0.9)
ResNet152	1.43 (0.16)	0.27 (0.79)
EfficientNetB0	1.39 (0.18)	0.26 (0.80)
EfficientNetB3	0.79 (0.44)	0.76 (0.45)
EfficientNetB6	1.02 (0.32)	0.24 (0.81)
MobileNetV1	1.84 (0.08)	0.66 (0.51)
MobileNetV2	0.62 (0.44)	0.32 (0.57)
3Hz condition.		
Optotrak	0 (0.93)	1.86 (0.18)
ResNet50	0.98 (0.34)	1.33 (0.26)
ResNet101	0.16 (0.69)	0.63 (0.44)
ResNet152	1.40 (0.17)	1.24 (0.22)
EfficientNetB0	1.30 (0.20)	1.63 (0.12)
EfficientNetB3	0.09 (0.93)	1.41 (0.17)
EfficientNetB6	0.73 (0.47)	1.08 (0.29)
MobileNetV1	1.27 (0.22)	0.99 (0.33)
MobileNetV2	1.28 (0.27)	2.18 (0.16)

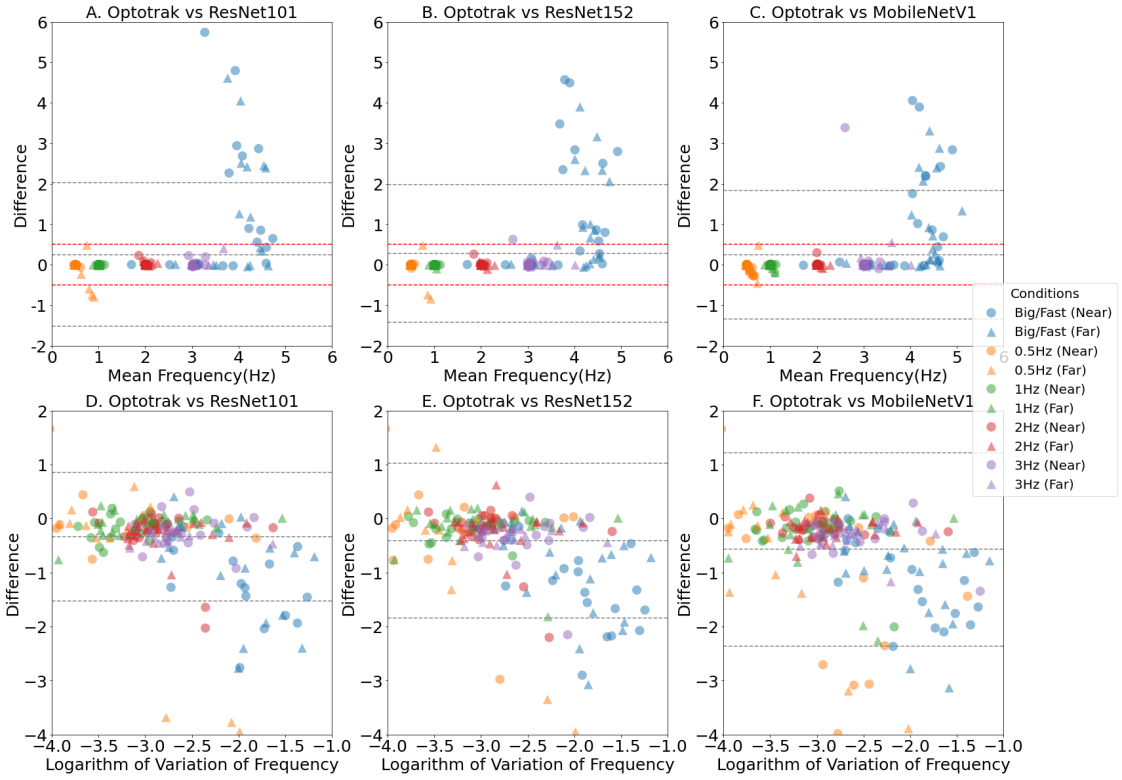

Figure 5: **Validity of computer measures of Tapping Frequency, demonstrated by the Bland Altman Plots, with the representation of the limits of agreement (red dashed lines), and from -1.96 standard deviation to +1.96 standard deviation (lower and upper grey dashed lines).** A, B and C show the mean tapping frequency comparison between the Optotrak system and ResNet101/ResNet152/MobileNetV1. D, E and F show a measure of tapping rhythm - the logarithm of variation in the tapping frequency. The colored marks represent the different finger tapping conditions with blue denoting the 'Big/Fast' self-paced conditions, and yellow, green, red and purple the externally paced conditions at frequencies of 0.5Hz, 1Hz, 2Hz and 3Hz respectively. Circles are for conditions performed Near-To-Laptop (60cm) and triangles are for conditions performed Far-From-Laptop (80cm).

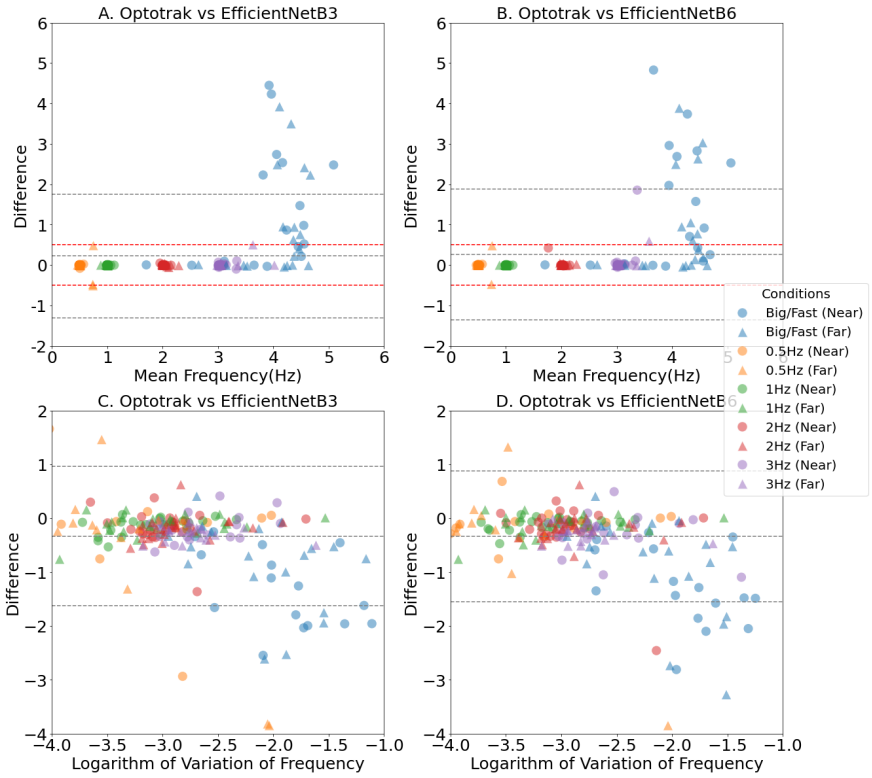

Figure 6: **Validity of computer measures of Tapping Frequency, demonstrated by the Bland Altman Plots, with the representation of the limits of agreement (red dashed lines), and from -1.96 standard deviation to +1.96 standard deviation (lower and upper grey dashed lines).** A and B show the mean tapping frequency comparison between the Optotrak system and EfficientNetB3/EfficientNetB6. C and D show a measure of tapping rhythm - the logarithm of variation in the tapping frequency. The colored marks represent the different finger tapping conditions with blue denoting the 'Big/Fast' self-paced conditions, and yellow, green, red and purple the externally paced conditions at frequencies of 0.5Hz, 1Hz, 2Hz and 3Hz respectively. Circles are for conditions performed Near-To-Laptop (60cm) and triangles are for conditions performed Far-From-Laptop (80cm).

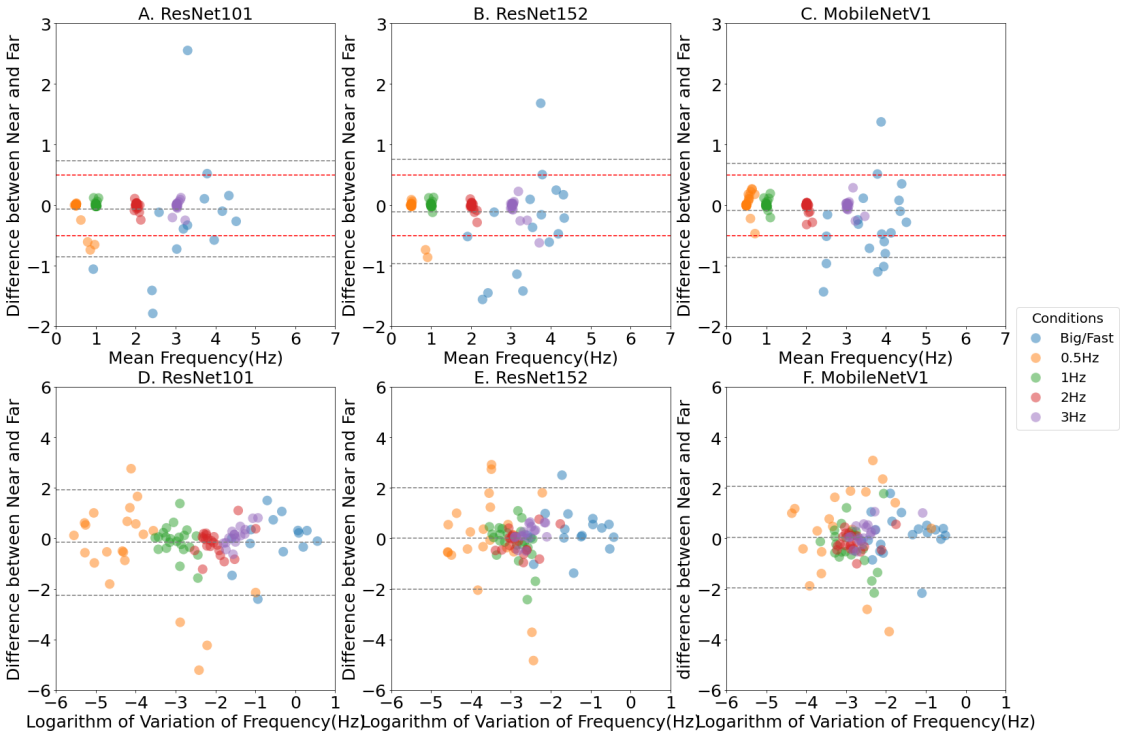

Figure 7: **Reliability of computer measures of Tapping Frequency at two distances from camera, demonstrated by the Bland Altman Plots, with the representation of the limits of agreement (red dashed lines), and from -1.96 standard deviation to +1.96 standard deviation (lower and upper grey dashed lines).** A, B and C show the mean tapping frequency scatter plots between Near-To-Laptop (60cm camera to hand) and Far-From-Laptop (80cm camera to hand) conditions for the ResNet101, ResNet152 and MobileNetV1 respectively. D, E and F show a measure of tapping rhythm - the logarithm of variation in the tapping frequency. The colored marks represent the different finger tapping conditions with blue denoting the 'Big/Fast' self-paced conditions, and yellow, green, red and purple are the externally paced conditions at frequencies of 0.5Hz, 1Hz, 2Hz and 3Hz respectively. The middle grey dashed line represents the mean difference between Near-To-Laptop and Far-From-Laptop conditions. The upper and lower grey dashed lines represent the upper and lower borders at 95% confidence level. The upper and lower red dashed lines represent the +/-0.5Hz agreement levels.

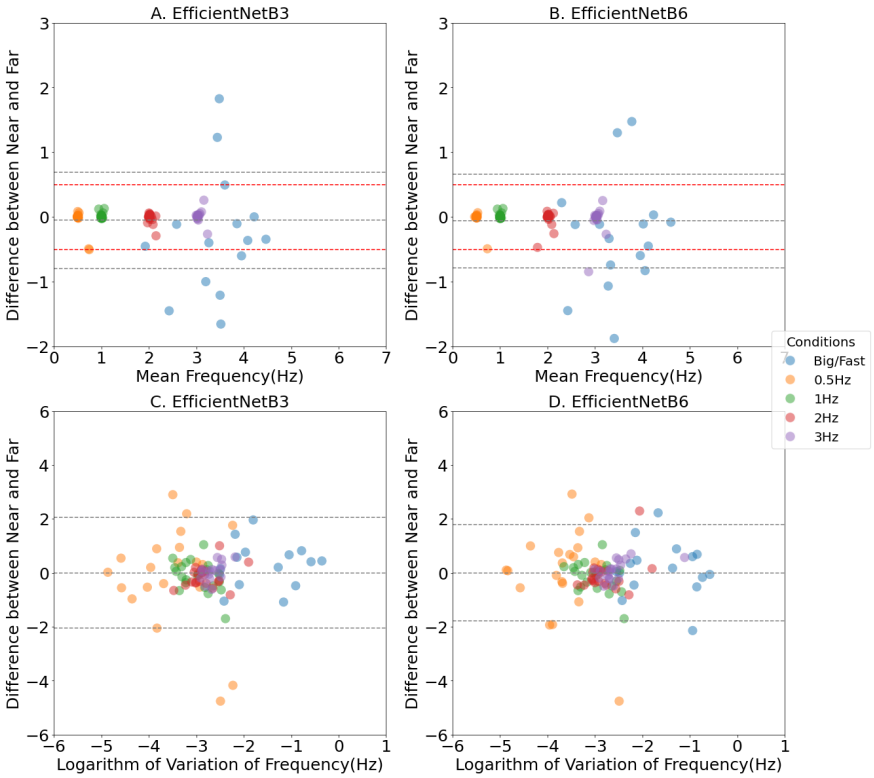

Figure 8: **Reliability of computer measures of Tapping Frequency at two distances from camera, demonstrated by the Bland Altman Plots, with the representation of the limits of agreement (red dashed lines), and from -1.96 standard deviation to +1.96 standard deviation (lower and upper grey dashed lines).** A and B show the mean tapping frequency scatter plots between Near-To-Laptop (60cm camera to hand) and Far-From-Laptop (80cm camera to hand) conditions for the EfficientNetB3 and EfficientNetB6 respectively. C and D show a measure of tapping rhythm - the logarithm of variation in the tapping frequency. The colored marks represent the different finger tapping conditions with blue denoting the 'Big/Fast' self-paced conditions, and yellow, green, red and purple are the externally paced conditions at frequencies of 0.5Hz, 1Hz, 2Hz and 3Hz respectively. The middle grey dashed line represents the mean difference between Near-To-Laptop and Far-From-Laptop conditions. The upper and lower grey dashed lines represent the upper and lower borders at 95% confidence level. The upper and lower red dashed lines represent the +/-0.5Hz agreement levels.