Supplementary Materials

A. Experiment setup

A.1. Training details

The models were trained in a multi-task style. During training, the model was required to predict labels of race & gender & gender of a training sample, while during inference we only use the gender label predicted by the model. For training hyperparameters, we used a batch-size of 128, an initial learning rate of $1e^{-3}$ which decays by 10 times at epoch 13 and epoch 17. We trained all the models with a total number of 21 epochs. We used an Adam optimizer.

A.2. Computing resources

All the experiments were run with NVIDIA Tesla A100 GPUs. Training a model per run took 1 GPU hour. Applying FGSM attack to test dataset of FairFace took $\sim 1,000$ seconds, applying CW attack to test dataset of FairFace took $\sim 6,000$ seconds.

B. Model performance

We report our networks' accuracies for different race groups on non-OOD test images in Table 1, to demonstrate that they all achieve reasonably high accuracies on both datasets. The performances do not significantly vary with FLKS because the training and testing images are all from the same distribution.

Table 1. **Model performances on *unperturbed* (non-OOD) images from the Fairface & UTKFace datasets.** We report the trained models' performances on test sets split by race group. Each number is an average over 3 different trained models. The first column indicates the dataset name and model's First Layer Kernel Size (FLKS). Fairface has 7 annotated race groups and UTKFace has 4. The performances are relatively constant with a variation to kernel size because the test and training images belong to the same distribution.

Dataset (FLKS)	Overall	White	Black	East Asian	Indian	Southeast Asian	Latino	Mid. Eastern
Fairface (3)	0.947	0.950	0.894	0.942	0.945	0.894	0.957	0.977
Fairface (5)	0.949	0.947	0.896	0.939	0.957	0.896	0.957	0.980
Fairface (7)	0.946	0.943	0.895	0.947	0.956	0.895	0.963	0.978
Fairface (9)	0.947	0.946	0.895	0.937	0.951	0.895	0.960	0.979
Fairface (11)	0.946	0.949	0.892	0.937	0.949	0.885	0.967	0.979
UTKFace (3)	0.929	0.949	0.905	0.931	0.942	/	/	/
UTKFace (5)	0.935	0.951	0.901	0.940	0.951	/	/	/
UTKFace (7)	0.934	0.955	0.905	0.939	0.953	/	/	/
UTKFace (9)	0.937	0.955	0.910	0.941	0.955	/	/	/
UTKFace (11)	0.936	0.950	0.901	0.943	0.955	/	/	/

C. Adversarial attack example

We show an example of a CW and FGSM attack for the same input image in the Supplementary, which further shows that CW perturbation is an order of magnitude smaller due to the effect of its regularization.

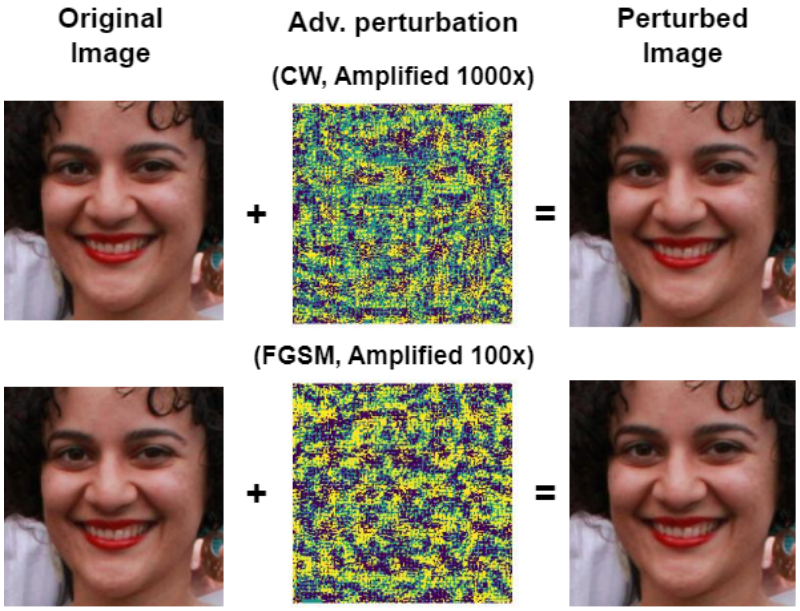

Figure 1. **Example of adversarial attack perturbations**. By adding tiny noise-like perturbations (center image, amplified 100/1000 times for visualization) to a test image (left), a target neural network will output a wrong prediction. However, the perturbed image (right) has no perceptible differences with the original image to the human eye. We use the CW and FGSM attacks in our experiments.

D. Results on DenseNet121

To further test the robustness and universality of our framework and conclusion, we also tested on DenseNet121 – another popular face analysis model. We also vary the first convolutional kernel size from $\{3, 5, 7, , 9, 11\}$. We report the averaged spectra in Figure 2, and its corresponding perturbation distance & $f_{0.5}$ scores in Figure 3.

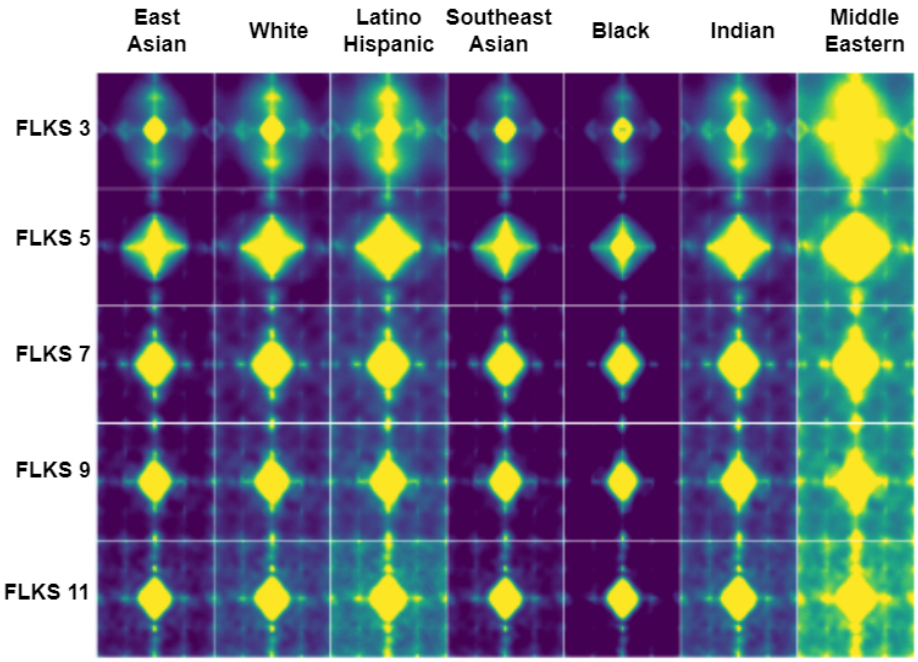

Figure 2. **Perturbation Spectrum Visualization on Fairface using DenseNet121.** Similar to results in Figure 3 in main text, each row represents a model with a different First Layer Kernel Size(FLKS), and each column corresponds to protected attribute groups. We observed a similar trending as discovered in the other 2 results above: generally, the perturbation shifts it's attention to low-frequency information as FLKS increases, and the perturbations for Black always have lower high-frequency focus compared to other race group.

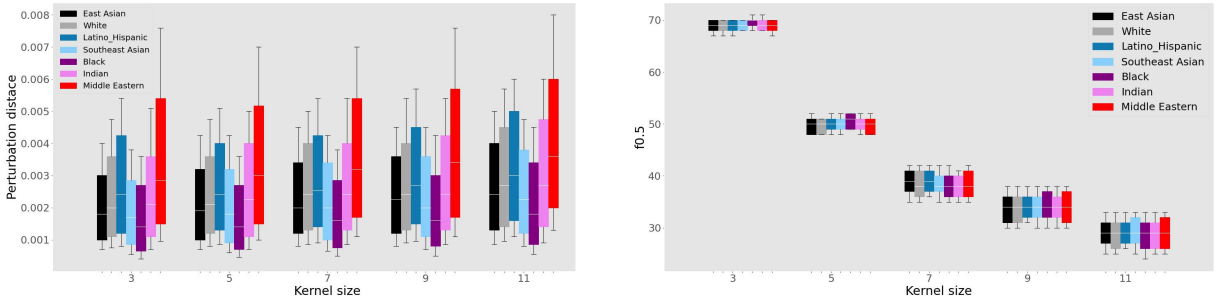

Figure 3. **Perturbation spectra $f_{0.5}$ & Perturbation distance for DenseNet121.** We visualize the perturbation and $f_{0.5}$ scores in the same way discussed in Sec. 4.1 and Figure 4. We observed a similar trend: there is a significant trend that the $f_{0.5}$ drops as the FLKS increases for all demographic groups and as the FLKS increases, the perturbation distances generally increase too for all the demographic groups.

E. Results on Vgg16

To further test the robustness and universality of our framework and conclusion, we also tested on Vgg16– another popular face analysis model. We also vary the first convolutional kernel size from $\{3, 5, 7, , 9, 11\}$. We report the averaged spectra in Figure 4, and its corresponding perturbation distance & $f_{0.5}$ scores in Figure 5.

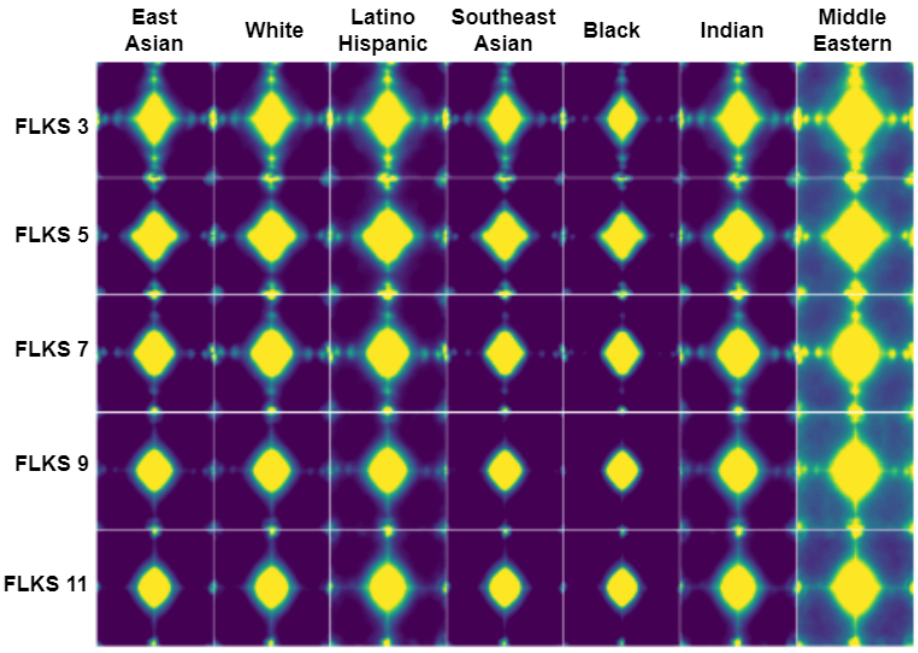

Figure 4. **Perturbation Spectrum Visualization on Fairface using Vgg16.** Similar to results in Figure 3 in main text, each row represents a model with a different First Layer Kernel Size(FLKS), and each column corresponds to protected attribute groups. We observed a similar trending as discovered in the other 2 results above: generally, the perturbation shifts it's attention to low-frequency information as FLKS increases, and the perturbations for Black always have lower high-frequency focus compared to other race group.

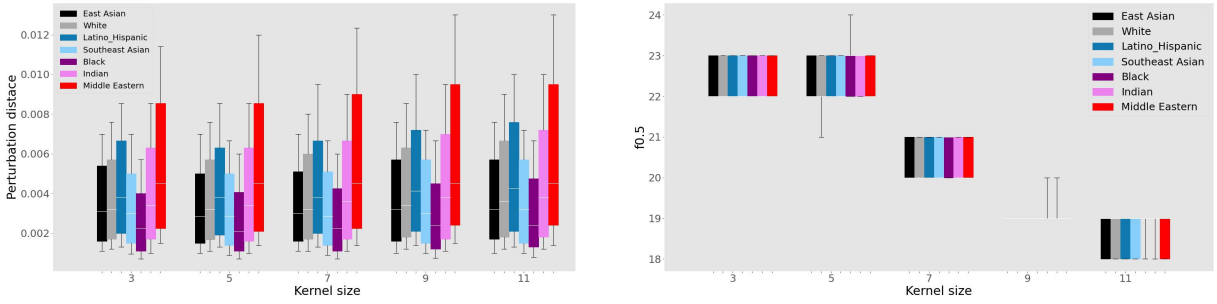

Figure 5. **Perturbation spectra $f_{0.5}$ & Perturbation distance for Vgg16.** We visualize the perturbation and $f_{0.5}$ scores in the same way discussed in Sec. 4.1 and Figure 4 We observed a similar trend: there is a significant trend that the $f_{0.5}$ drops as the FLKS increases for all demographic groups and as the FLKS increases, the perturbation distances generally increase too for all the demographic groups.

F. Regression results

We report the coefficients $\boldsymbol{\beta}$ (left) and $\boldsymbol{\gamma}$ (right) for the regression described in Section 4.1.3. We also report the standard deviations of the coefficient values and calculate their t values according to $t = \frac{coefvalue}{stderr}$, as well as $P > |t|$. A $P \leq 0.05$ indicates the value is significant.

Table 2. **Regression results of perturbation distance**. The β coefficient names use subscripts corresponding to race (E: East Asian, W: White, B: Black, I: Indian, L: Latino, S: Southeast Asian), and gender (M: Male, F: Female). Some large disparities between $\boldsymbol{\beta}$ values across groups are obvious, e.g., the values for Black Asian Female is $\sim 100\%$ higher than that of East Asian Female. Refer to Figure 6 in Main paper for plots on $\boldsymbol{\beta}$.

coef name	coef value	std err	t	$P > \|t\|$	coef name	coef value	std err	t	$P > \|t\|$
$\boldsymbol{\beta}_{EM}$	0.0227	0.003	6.930	0.000	$\boldsymbol{\gamma}_{EM}$	0.4251	0.025	17.308	0.000
$\boldsymbol{\beta}_{EF}$	0.0274	0.003	8.323	0.000	$\boldsymbol{\gamma}_{EF}$	0.4232	0.025	17.065	0.000
$\boldsymbol{\beta}_{WM}$	0.0306	0.003	11.392	0.000	$\boldsymbol{\gamma}_{WM}$	0.4453	0.020	21.969	0.000
$\boldsymbol{\beta}_{WF}$	0.0254	0.003	8.614	0.000	$\boldsymbol{\gamma}_{WF}$	0.4302	0.022	19.361	0.000
$\boldsymbol{\beta}_{LM}$	0.0351	0.003	11.143	0.000	$\boldsymbol{\gamma}_{LM}$	0.2088	0.012	17.566	0.000
$\boldsymbol{\beta}_{LF}$	0.0269	0.003	8.521	0.000	$\boldsymbol{\gamma}_{LF}$	0.2425	0.012	20.304	0.000
$\boldsymbol{\beta}_{SM}$	0.0251	0.003	7.552	0.000	$\boldsymbol{\gamma}_{BM}$	0.1856	0.013	14.783	0.000
$\boldsymbol{\beta}_{SF}$	0.0189	0.004	5.982	0.000	$\boldsymbol{\gamma}_{BF}$	0.2163	0.014	15.281	0.000
$\boldsymbol{\beta}_{BM}$	0.0589	0.001	48.770	0.000	$\boldsymbol{\gamma}_{BM}$	0.2088	0.012	17.505	0.000
$\boldsymbol{\beta}_{BF}$	0.0659	0.001	48.588	0.000	$\boldsymbol{\gamma}_{BF}$	0.2425	0.012	20.234	0.000
$\boldsymbol{\beta}_{IM}$	0.0748	0.001	61.160	0.000	$\boldsymbol{\gamma}_{BM}$	0.1856	0.013	14.783	0.000
$\boldsymbol{\beta}_{IF}$	0.0846	0.001	65.294	0.000	$\boldsymbol{\gamma}_{BF}$	0.2163	0.014	15.821	0.000

G. Results of varying all convolutional kernel sizes

We also test our framework on the occasion where we modify all the convolutional layers' kernel sizes. The results are in Figure 6. Basically, we found that modifying all convolutional layers' kernel sizes doesn't make a significant difference comparing to *only* modify the first convolutional kernel size. Refer to the caption for more details.

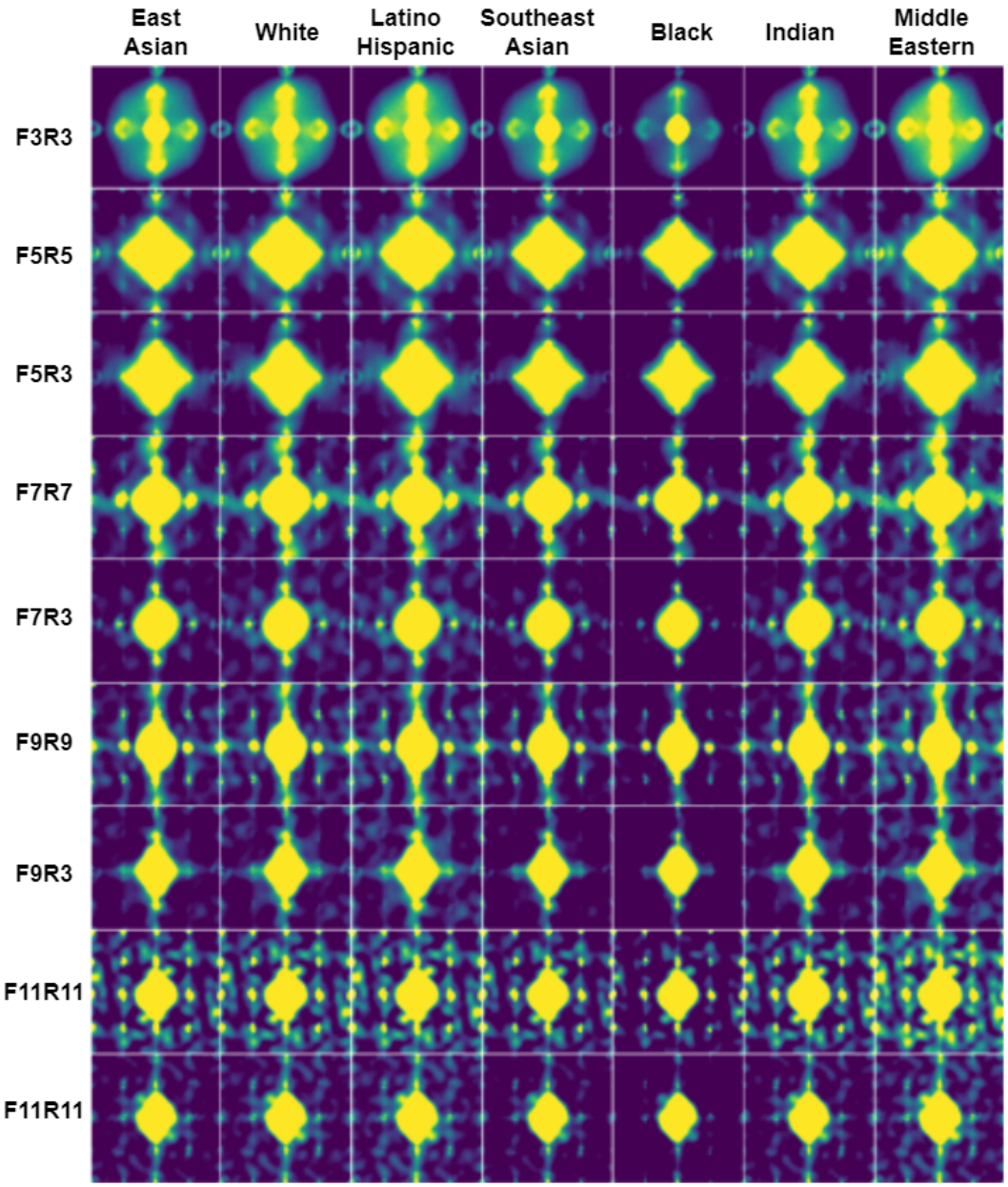

Figure 6. **Perturbation Spectrum Visualization on Fairface for modifying all convolutional layer kernel size.** Each row represents a model with a different architectual choice, for example, "F3R3" represents model with first layer kernel size of 3 and the rest layers kernel sizes of 3, and each column corresponds to protected attribute groups. We found that modifying all convolutional layers' kernel sizes doesn't make a significant difference comparing to *only* modify the first convolutional kernel size.

H. Results of applying CW attack to model trained on UTKFace

We also conduct experiments on UTKFace, another popular face image dataset and report the results in Figure 7. Refer to Figure 3 in main text for results on Fairface and analysis.

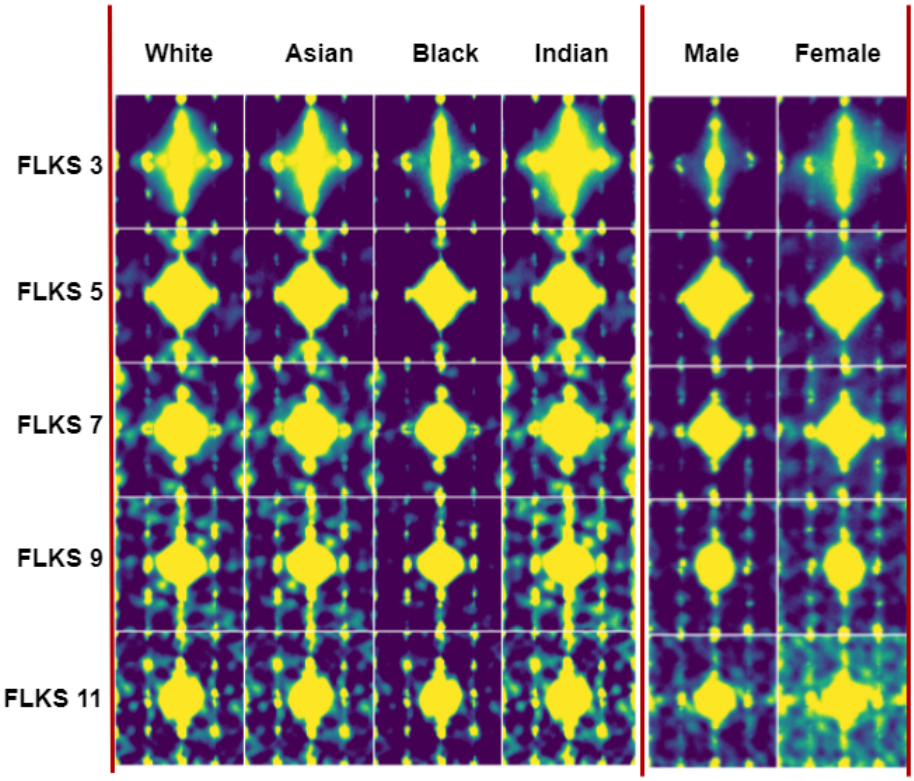

Figure 7. **Average spectra of adversarial perturbation images split by race and gender for UTKFace**. We see similar trends in these spectra to the ones shown for Fairface in Fig 3 in Main paper.

I. Results of applying FGSM to model

We also test our framework on the occasion where we apply FGSM attack to all the models trained on Fairface. The results are in Figure 9. It has basically the same trend with all the previous spectra visualization.

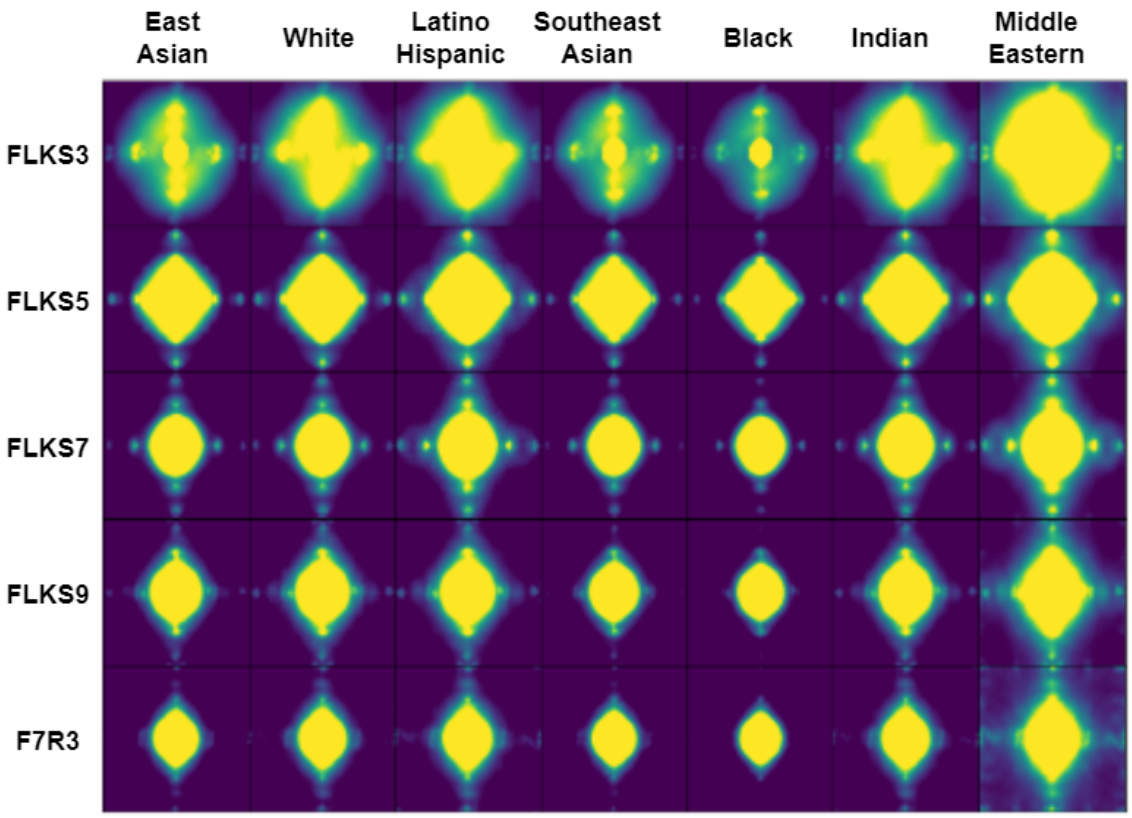

Figure 8. **Perturbation Spectrum Visualization on Fairface using FGSM.** Similar to results in Figure 4, each row represents a model with a different First Layer Kernel Size(FLKS), and each column corresponds to protected attribute groups. We observed a similar trending: generally, the perturbation shifts it's attention to low-frequency information as FLKS increases, and the perturbations for Black always have lower high-frequency focus compared to other race group.

J. Frequency energy injection result

Same to Figure 5 in main paper, we show models' performances with different FLKS for all race groups separately.

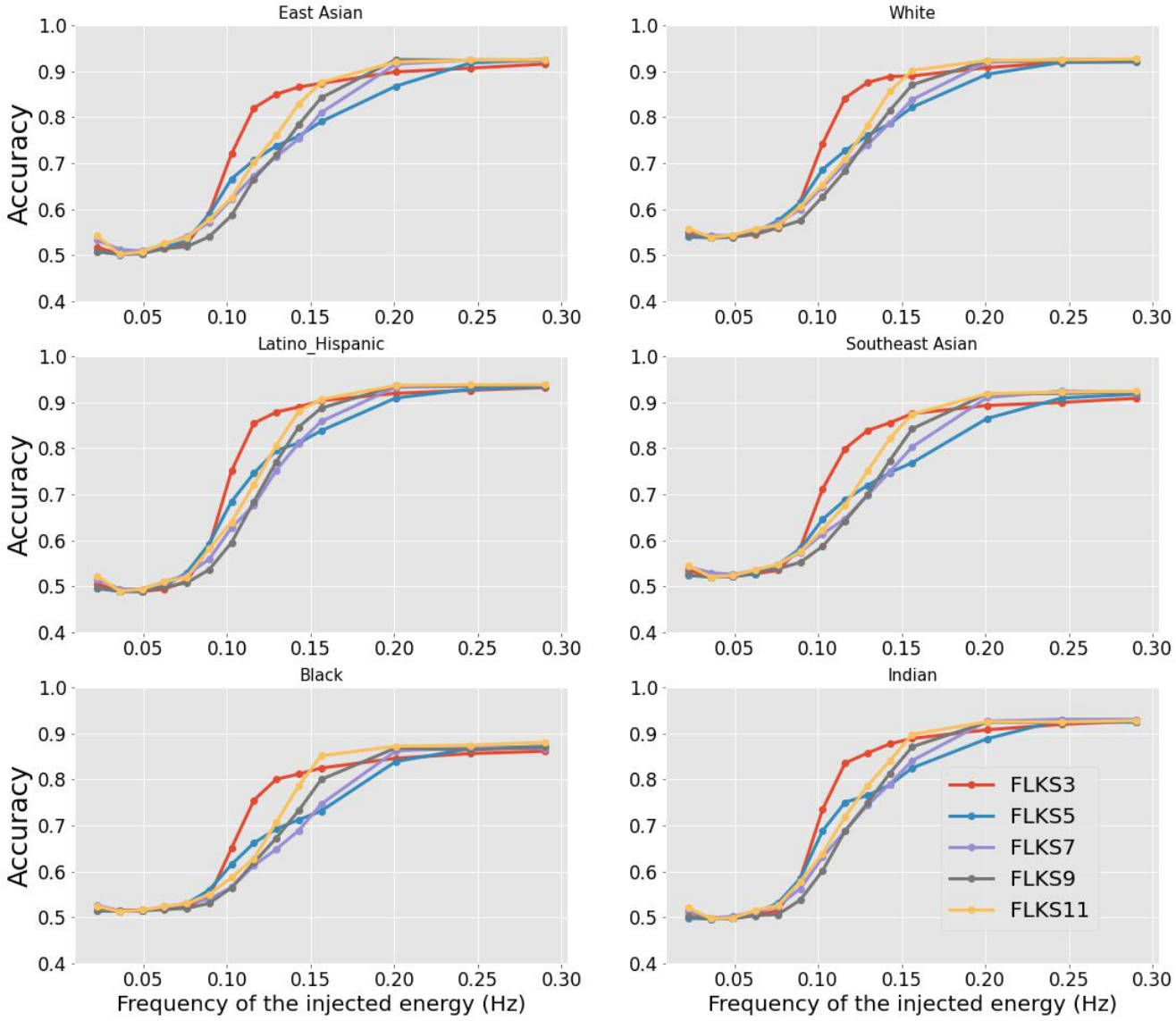

Figure 9. **Frequency energy injection result.** In each individual figure, the x-axis is the frequency we are injecting energy at and the y-axis is the accuracy of different models. It is obvious that all the models suffer from low to mid frequency's energy injections, and become robust to mid to high frequency noises. It is hard to directly tell which group is getting influenced more than the others, which furthers asks for a quantitatively analysis.