A. Training UAP with Images from a Single Class.

We report results of our Simple-UAP algorithm trained on images sampled from a single class in Table 1. The performance is close to the case of using images of all classes. Additionally, training on different single classes leads to different dominant labels, and the results are summarized in Table 2. When the single training class is fixed, the resulting dominant label with different runs is the same in most cases. However, when the single training class is changed, the corresponding dominant label also changes. This constitutes another empirical evidence that the dominant label does not necessarily occupy large regions in the image space as hypothesized by [3]. If the dominant label phenomenon is caused by the fact that the dominant label occupies large regions in the image space, the resulting dominant label is not supposed to change with the choice of a single training class.

Table 1. Simple-UAP algorithm comparison on the ImageNet validation dataset with the metric of fooling ratio (%). The single class algorithm trained only on samples from one class ("quilt").

<table>
<thead>
<tr>
<th>Method</th>
<th>AlexNet</th>
<th>GoogleNet</th>
<th>VGG16</th>
<th>VGG19</th>
<th>ResNet152</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Class</td>
<td>93.2</td>
<td>85.7</td>
<td>91.8</td>
<td>90.8</td>
<td>77.6</td>
</tr>
<tr>
<td>Normal Training</td>
<td>96.5</td>
<td>90.5</td>
<td>97.4</td>
<td>96.4</td>
<td>90.2</td>
</tr>
</tbody>
</table>

B. GD-UAP Results.

Here, we report the results for GD-UAP (see Figure 1). The results of GD-UAP resemble the trend for Cosine-UAP, in the sense of higher \( \cos(v, x + v) \) than \( \cos(x, x + v) \) for most latter layers and a positive relationship between cosine \( \cos(v, x + v) \) and fooling ratio. Overall, the Cosine-UAP results in higher \( \cos(v, x + v) \) compared with GD-UAP, especially in the very last few layers, which partially explains why the Cosine-UAP yields a higher fooling ratio.

![Figure 1. Layer-wise (left) model and step-wise (right) analysis of the GD-UAP on the model response in the untargeted setting.](image)

![Figure 2. Three examples of robust samples (first three columns) and vulnerable sample (last three columns) with their respective fourier transforms.](image)

*Equal contribution*
C. Visualization of Robust and Vulnerable Samples.

The visualization of robust and vulnerable samples is shown in Figure 2. Robust samples tend to have more high-frequency content.

D. Practical Data-free Black-Box Attack.

Experiment Setup. As widely reported in previous literature [3], random noise has very limited influence on accuracy. Our ablation study in Table 10 of the main manuscript also confirms this by showing the average accuracy under uniform noise perturbation is as high as 67.3%. Our analysis in Sec 4.3 of the main manuscript shows that content with repetitive patterns usually has a high influence on the joint DNN response when it is combined with another independent content as the DNN combined input. Inspired by the above finding, we design adversarial perturbation with repetitive patterns. The patterns that we investigate include horizontal, vertical, and checkerboard patterns, which are shown in Figure 4. Taking a horizontal pattern, for example, the pixel values are constant in the horizontal direction but receptively change in the vertical direction with values either $-\epsilon$ and $\epsilon$. Following [2], we set $\epsilon$ to 0.1. One hyper-parameter here is the width of the lines and we empirically find that the width of 2 pixels achieves satisfactory performance. For both horizontal and vertical patterns, we set it to 2 pixels. The checkerboard’s square size is set to $2 \times 2$ pixels. We optionally remove some HF content from the original images which can further enhance the attack success rate. For a fair comparison with [2], we finally clip the final resulting perturbation with the $l_\infty$ constraint $\epsilon$. We adopt two common ways for removing the high frequency with Fourier transform(FT) or SVD [1]. For adopting FT to remove HF content, we adopt the approach introduced in [4] and set the bandwidth to 36. For adopting SVD to remove HF content, we keep content corresponding to the top 12 singular values.

Additional Analysis. As demonstrated in the main manuscript, UAP has a dominant influence on the joint model response triggered by adversarial examples. Our designed adversarial perturbation with repetitive patterns is also universal since it can be added to any random image. For analyzing its influence on the joint model response of adversarial samples, i.e., images + our designed pattern, we first investigate the model prediction taking only the designed pattern.

References