# Defending Against Universal Perturbations With Shared Adversarial Training

## A. Supplementary material

### A.1. Threat Model

Here, we specify the capabilities of the adversary since the proposed defense mechanism aims at providing security under a specific threat model. We assume a *white-box* setting, where the adversary has full information about the model, i.e., it knows network architecture and weights, and can provide arbitrary inputs to the model and observe their corresponding outputs (and loss gradients). Moreover, we assume that the attacker can arbitrarily modify every pixel of the input but aims at keeping the $l_\infty$ norm of this perturbation minimal. In the case of a universal perturbation, we assume that the attacker can choose an arbitrary perturbation (while aiming to keep the $l_\infty$ norm minimal), but crucially does not know the inputs to which this perturbation will be applied. The adversary, however, has access to data points that have been sampled from the same data distribution as the future inputs.

### A.2. Relationship of different sharedness

Provided that the heap adversary finds a perturbation that is sufficiently close to the optimal perturbation of the heap and that heaps are composed hierarchically[1], we have the following relationship for $s = 2^i$ (we omit the dependence on $\sigma$, $\mathcal{S}$ and $f_{adv}/f_{heap}/f_{uni}$ for brevity):

$$\tilde{\rho}_{adv} = \tilde{\rho}^{(1)}_{heap} \geq \tilde{\rho}^{(2)}_{heap} \geq \tilde{\rho}^{(4)}_{heap} \geq \cdots \geq \tilde{\rho}^{(d)}_{heap} \geq \tilde{\rho}_{uni}(\sigma, \mathcal{S})$$

To see $\tilde{\rho}^{(s)}_{heap} \geq \tilde{\rho}^{(2s)}_{heap}$, let $\xi_1, \ldots, \xi_{d/(2s)}$ be the shared perturbations on the $d/(2s)$ heaps of $\tilde{\rho}^{(2s)}_{heap}$. Let $\xi_j$ be the shared perturbation for the $j$-th heap. Then, because of the hierarchical construction of the heaps, this heap is composed of two heaps used in $\tilde{\rho}^{(s)}_{heap}$. Let $j_1$ and $j_2$ be the corresponding indices of these heaps in $\tilde{\rho}^{(s)}_{heap}$. By setting $\xi_{j_1} = \xi_{j_2} = \xi_j$, we obtain $\tilde{\rho}^{(s)}_{heap} = \tilde{\rho}^{(2s)}_{heap}$.

### A.3. Configuration of Baselines for CIFAR10

For the defense proposed by Moosavi-Dezfooli et al. [29], we generated 10 different universal perturbations using the DeepFool-based method for generating universal-perturbations on 10,000 randomly sampled training images, ran 5 epochs of adversarial training with $\sigma = 0.5$, and chose the applied perturbation uniform randomly from the precomputed perturbations. After these 5 epochs, the robustness was evaluated. This procedure was iterated five times, which resulted in 5 accuracy-robustness points in Figure 1.

We run the defense proposed by Perolat et al. [37] for 45 epochs (sufficiently long for achieving convergence as evidenced by Figure 4 of Perolat et al. [37]). At the beginning of each episode, we generated one universal perturbation using the DeepFool-based method for generating universal-perturbations on the entire training set. We used $\sigma = 0.25$, and chose the applied perturbation uniform randomly from all universal perturbations computed so far. We report the accuracy and robustness at the end of these 45 epochs. We note that even though we did not replicate the exact setup of Perolat et al. [37], we achieve a similar accuracy-robustness trade-off in Figure 3 (right) as the one given in Figure 4 of Perolat et al. [37].

### A.4. Illustration of Universal Perturbations on CIFAR10

Figure A1 illustrates the minimal universal perturbation found for sharedness $s = 1$ and $s = 64$ for $\sigma = 0.3$ and $\varepsilon \in \{2, 8, 14, 18, 26\}$. Universal perturbations of the undefended model resemble high-frequency noise and are quasi-imperceptible when added to an image. Shared adversarial training increases robustness and the resulting perturbations are more perceptible (for small $\varepsilon$) or even dominate the image: for larger $\varepsilon$, the cat in the figure is completely hidden and the perturbed image could also not be classified correctly by a human. Moreover, the perturbation becomes more structured and even object-like for larger $\varepsilon$. Note that the perturbations shown for $s = 1$ also achieve high robustness but for smaller accuracy on clean data than those of shared adversarial training with $s = 64$.

[1] Heaps are composed hierarchically when a heap of sharedness $2s$ is always the union of two disjoint heaps of sharedness $s$.

## A.5. Selection of subset of ImageNet

Since the generation of the Pareto fronts on the entire ImageNet dataset would be computationally very expensive, we restrict the experiment to a subset of ImageNet. We use classes defined in TinyImageNet to filter out the samples from ImageNet dataset. We conducted our experiments on the samples of 200 classes from ImageNet, which results in 258,601 train and 10,000 validation images. Note that we take only the list of classes defined from TinyImageNet and use the data of those classes from ImageNet dataset with original resolution.

## A.6. Illustration of Universal Perturbations on ImageNet

We depict the universal perturbations with minimum magnitude on different models that are obtained from settings $\sigma = 1.0$, sharedness $s \in \{1, 32\}$ and different values of $\varepsilon$ on the subset of ImageNet in Figure A2. It can be clearly seen that both the standard ($s = 1$) and shared adversarial training ($s = 32$) increase robustness when compared against the undefended model but the latter handles the trade-off between performance on unperturbed data and robustness more gracefully. The universal perturbations become clearly visible on a model hardened with shared adversarial training with only a marginal loss of $5\%$ in top-1 accuracy and perturbations become much smoother for larger $\varepsilon$.

## A.7. Predicted Class after Untargeted Universal Perturbations

Figure A3 shows which class is predicted on ImageNet validation data after an untargeted universal perturbation (for the respective model) is added. While the undefended model predicts nearly always the same (wrong) class, the models defended with standard and shared adversarial training have a substantially higher entropy in their predictions. Prior work [18] has also observed that undefended models typically misclassify images perturbed with universal perturbation to the same class even though the attack is untargeted. Based on this observation, they hypothesized that directions in which a classifier is vulnerable to universal perturbations coincide with directions important for correct prediction on unperturbed data. We believe it would be important to re-examine these results for a defended model.

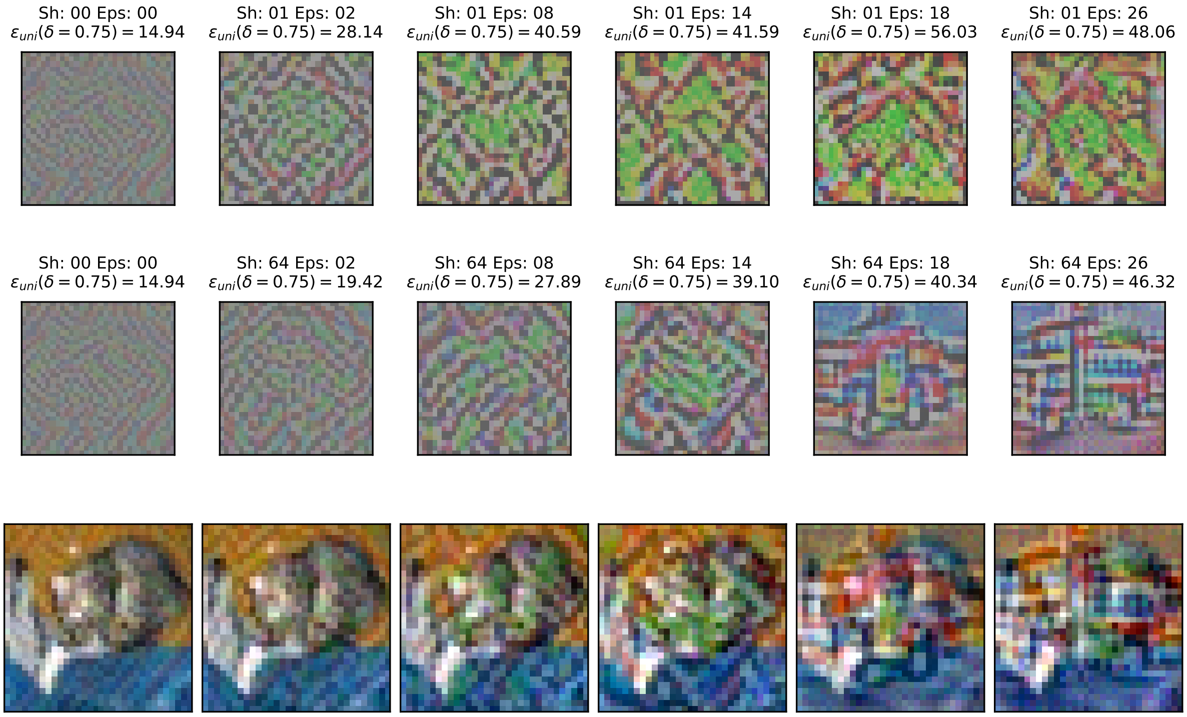
Sh: 00 Eps: 00
$\varepsilon_{uni}(\delta = 0.75) = 14.94$
Sh: 01 Eps: 02
$\varepsilon_{uni}(\delta = 0.75) = 28.14$
Sh: 01 Eps: 08
$\varepsilon_{uni}(\delta = 0.75) = 40.59$
Sh: 01 Eps: 14
$\varepsilon_{uni}(\delta = 0.75) = 41.59$
Sh: 01 Eps: 18
$\varepsilon_{uni}(\delta = 0.75) = 56.03$
Sh: 01 Eps: 26
$\varepsilon_{uni}(\delta = 0.75) = 48.06$
Sh: 00 Eps: 00
$\varepsilon_{uni}(\delta = 0.75) = 14.94$
Sh: 64 Eps: 02
$\varepsilon_{uni}(\delta = 0.75) = 19.42$
Sh: 64 Eps: 08
$\varepsilon_{uni}(\delta = 0.75) = 27.89$
Sh: 64 Eps: 14
$\varepsilon_{uni}(\delta = 0.75) = 39.10$
Sh: 64 Eps: 18
$\varepsilon_{uni}(\delta = 0.75) = 40.34$
Sh: 64 Eps: 26
$\varepsilon_{uni}(\delta = 0.75) = 46.32$

Figure A1. Illustration of universal perturbations on CIFAR10 for sharedness $s = 1$ (top row) and $s = 64$ (middle row) for different values of $\varepsilon$. The bottom row shows a test image of a cat with the respective perturbation of the middle row being added.

## A.8. Attacks on Semantic Image Segmentation

We illustrate universal perturbations for targeted and untargeted attacks on different models in this section. We illustrate the effect of the perturbations on one image; however, the perturbations are not specific for this image. For the model trained with empirical risk minimization, Figure A4 shows a targeted attack and Figure A5 an untargeted attack. For the model trained with adversarial training, Figure A6 shows a targeted attack and Figure A8 an untargeted attack. For the model trained with shared adversarial training, Figure A7 shows a targeted attack and Figure A9 an untargeted attack. We also illustrate the universal perturbations found for different models on targeted attacks in Figure A10.

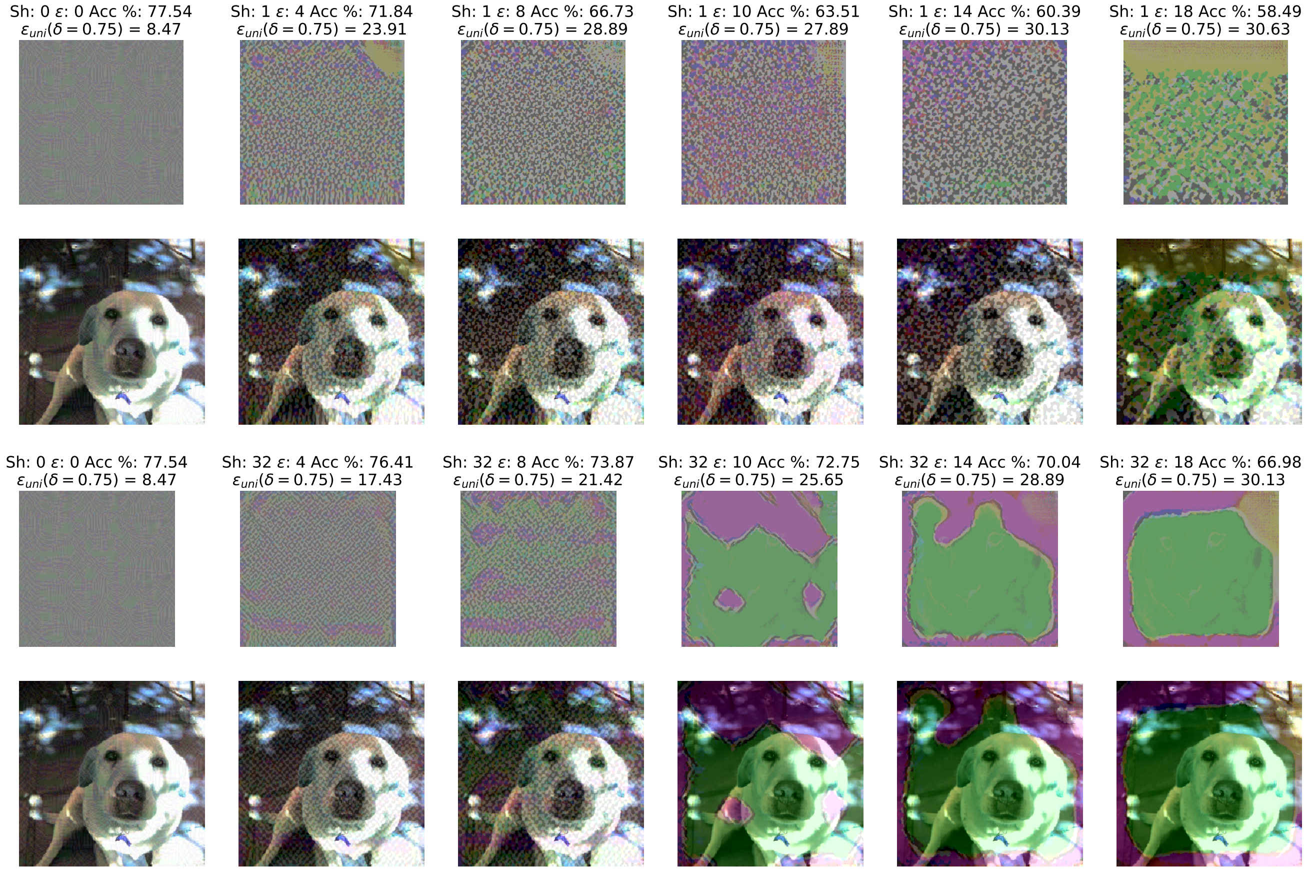
Sh: 0 $\varepsilon$: 0 Acc %: 77.54
$\varepsilon_{uni}(\delta = 0.75) = 8.47$
Sh: 1 $\varepsilon$: 4 Acc %: 71.84
$\varepsilon_{uni}(\delta = 0.75) = 23.91$
Sh: 1 $\varepsilon$: 8 Acc %: 66.73
$\varepsilon_{uni}(\delta = 0.75) = 28.89$
Sh: 1 $\varepsilon$: 10 Acc %: 63.51
$\varepsilon_{uni}(\delta = 0.75) = 27.89$
Sh: 1 $\varepsilon$: 14 Acc %: 60.39
$\varepsilon_{uni}(\delta = 0.75) = 30.13$
Sh: 1 $\varepsilon$: 18 Acc %: 58.49
$\varepsilon_{uni}(\delta = 0.75) = 30.63$
Sh: 0 $\varepsilon$: 0 Acc %: 77.54
$\varepsilon_{uni}(\delta = 0.75) = 8.47$
Sh: 32 $\varepsilon$: 4 Acc %: 76.41
$\varepsilon_{uni}(\delta = 0.75) = 17.43$
Sh: 32 $\varepsilon$: 8 Acc %: 73.87
$\varepsilon_{uni}(\delta = 0.75) = 21.42$
Sh: 32 $\varepsilon$: 10 Acc %: 72.75
$\varepsilon_{uni}(\delta = 0.75) = 25.65$
Sh: 32 $\varepsilon$: 14 Acc %: 70.04
$\varepsilon_{uni}(\delta = 0.75) = 28.89$
Sh: 32 $\varepsilon$: 18 Acc %: 66.98
$\varepsilon_{uni}(\delta = 0.75) = 30.13$

Figure A2. Illustration of universal perturbations (not amplified) on ImageNet that are generated from different models with the settings: sharedness $s = 1$ (top row) and $s = 32$ (third row), $\sigma = 1.0$ and different values of $\varepsilon$. The top-1 accuracy of the corresponding models and their smallest perturbation magnitude $\varepsilon$ that results in a misclassification rate of atleast $75\%$ are also shown. The second and bottom rows show a test image of a dog added with the respective universal perturbations from the first and third row. The models hardened with both standard and shared adversarial training demonstrate higher robustness when compared against the undefended model and universal perturbations become clearly visible. However, the shared adversarial training outperforms its counterpart in terms of robustness against perturbations and performance on unperturbed inputs. The perturbation of models from standard adversarial training resemble high frequency noise whereas the perturbations of the latter becomes much smoother for larger $\varepsilon$.

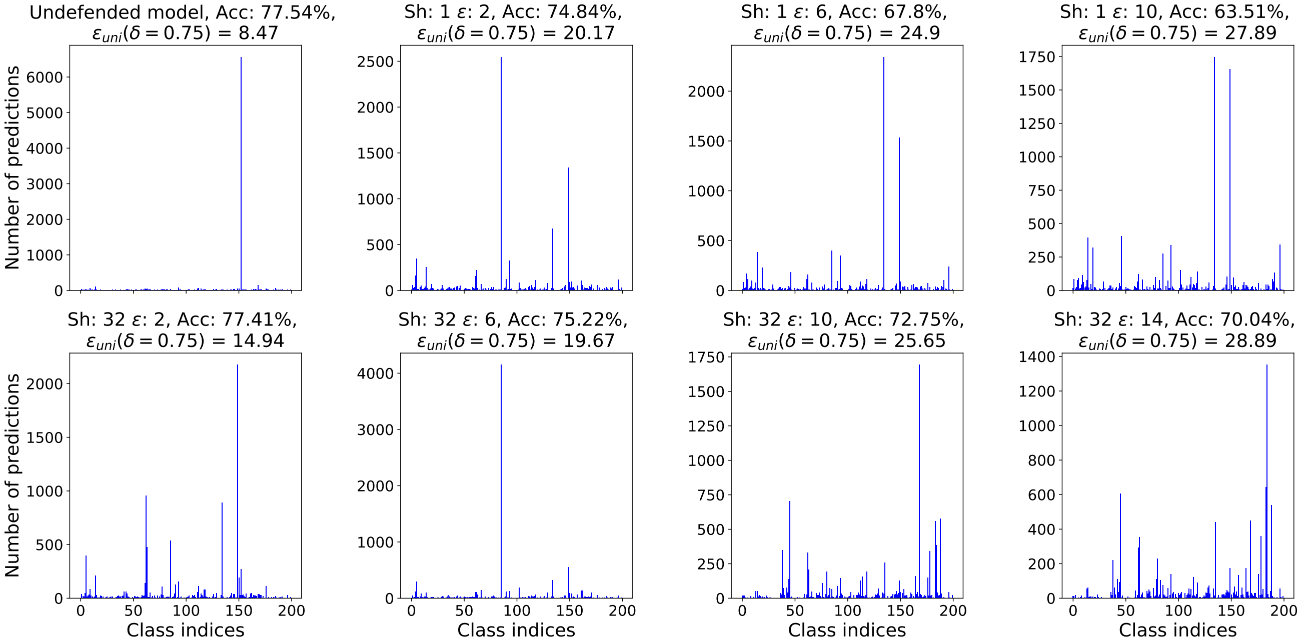

Figure A3. The figure shows histogram of the predicted classes over validation data of different models when an untargeted universal perturbation is added. The histogram is based on 200 classes ImageNet validation data. In other words, a bar in each histogram represents the number of times a class (represented by class index) is predicted over the validation samples. It is interesting to note that the undefended model almost always misclassified the adversarial samples (samples added with universal perturbations) under the same class even though attack is untargeted. In contrast, the defended models from standard and shared adversarial training have higher entropy in their predictions.

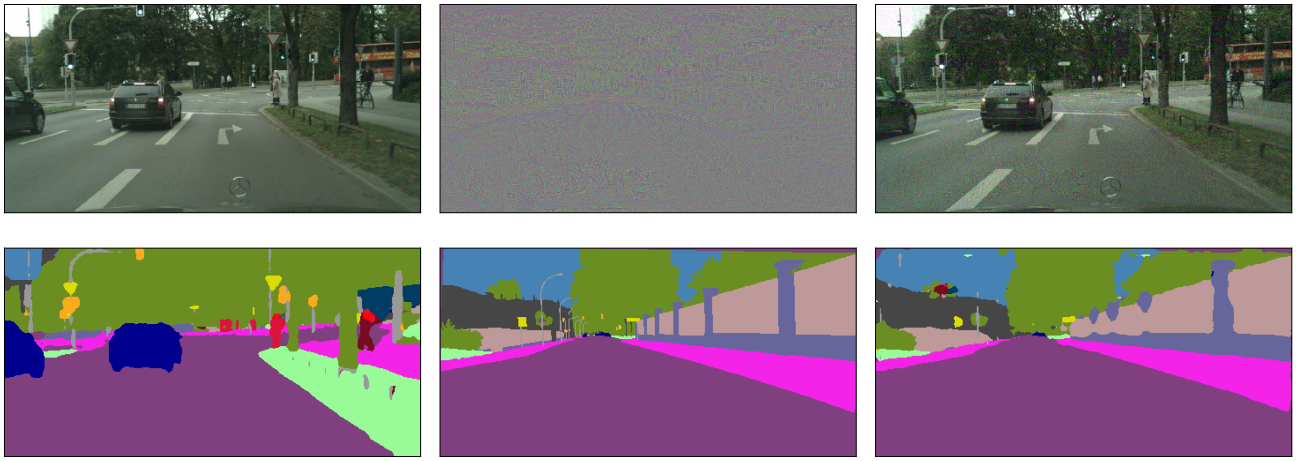

Figure A4. Targeted universal perturbations on Cityscapes for a model pretrained with empirical risk minimization. The shown perturbation upper bounds the robustness of the model to $\varepsilon_{uni}(\delta = 0.95) \leq 19.89$. Top row shows original image, universal perturbation, and perturbed image. Bottom row shows prediction on original image, target segmentation, and prediction on perturbed image.

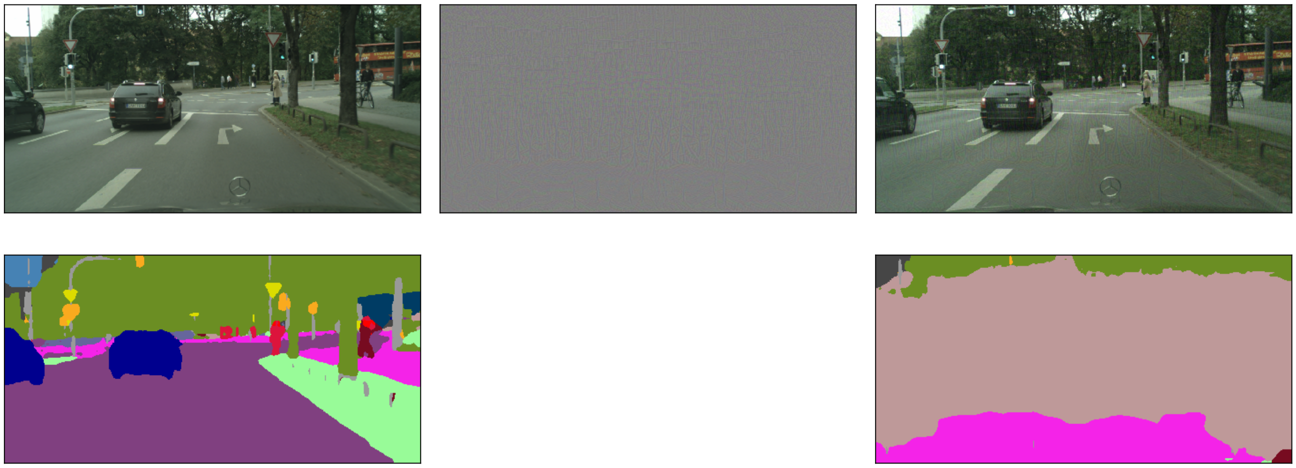

Figure A5. Untargeted universal perturbations on Cityscapes for a model pretrained with empirical risk minimization. The shown perturbation upper bounds the robustness of the model to $\varepsilon_{uni}(\delta = 0.95) \leq 8.5$. Top row shows original image, universal perturbation, and perturbed image. Bottom row shows prediction on original image and prediction on perturbed image.

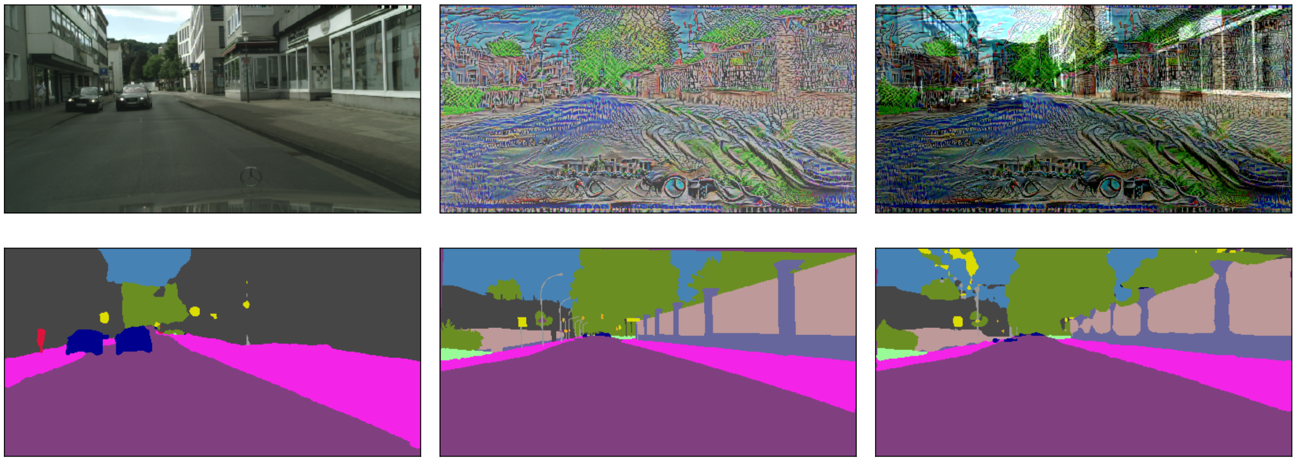

Figure A6. Targeted universal perturbations on Cityscapes for a model trained with adversarial training. The shown perturbation upper bounds the robustness of the model to $\varepsilon_{uni}(\delta = 0.95) \leq 62.5$. Top row shows original image, universal perturbation, and perturbed image. Bottom row shows prediction on original image, target segmentation, and prediction on perturbed image.

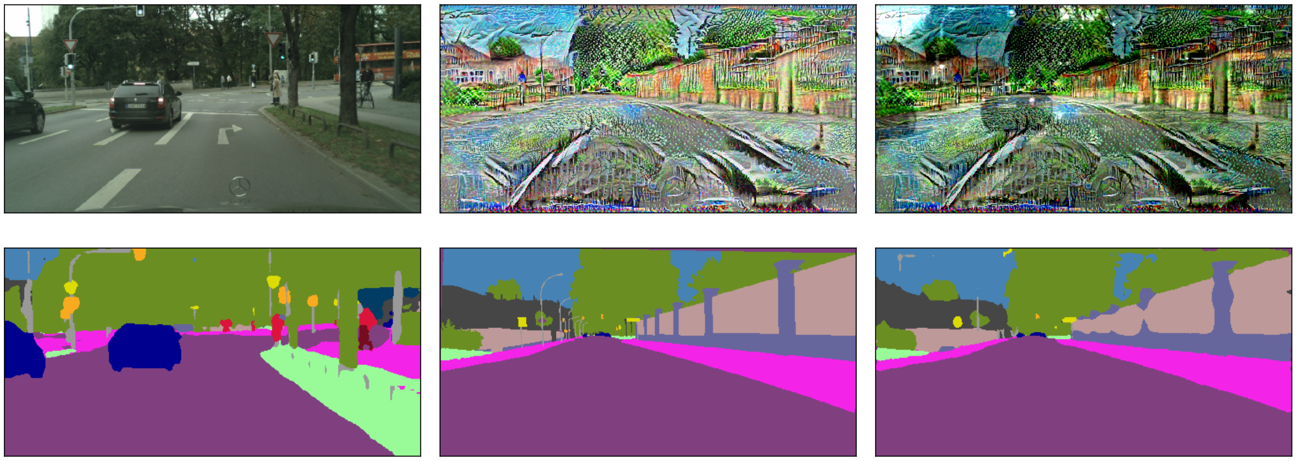

Figure A7. Universal perturbations on Cityscapes for a model trained with shared adversarial training. The shown perturbation upper bounds the robustness of the model to $\varepsilon_{uni}(\delta = 0.95) \leq 111.7$. Top row shows original image, universal perturbation, and perturbed image. Bottom row shows prediction on original image, target segmentation, and prediction on perturbed image.

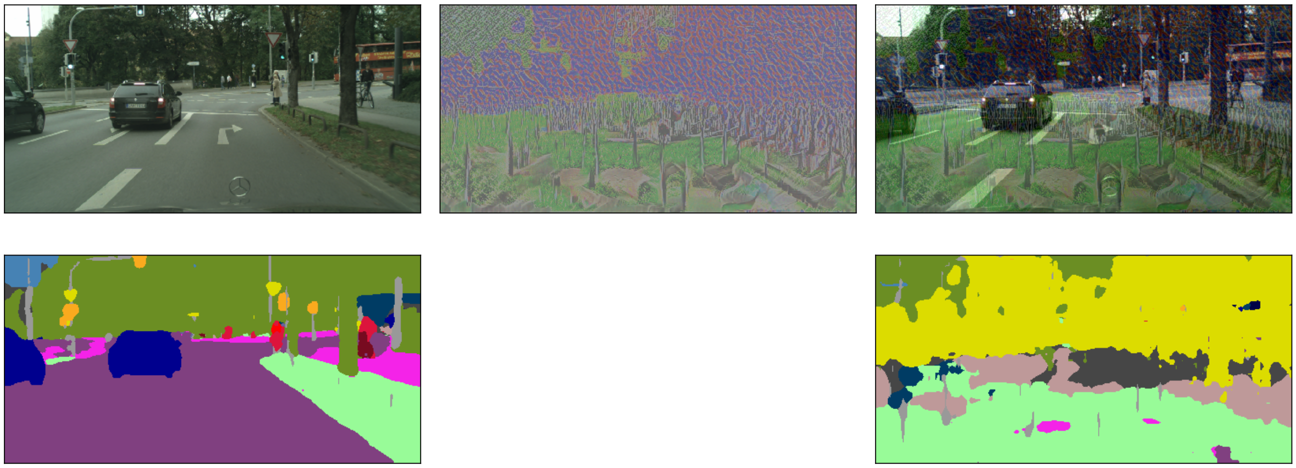

Figure A8. Untargeted universal perturbations on Cityscapes for a model trained with adversarial training. The shown perturbation upper bounds the robustness of the model to $\varepsilon_{uni}(\delta = 0.95) \leq 25$. Top row shows original image, universal perturbation, and perturbed image. Bottom row shows prediction on original image and prediction on perturbed image.

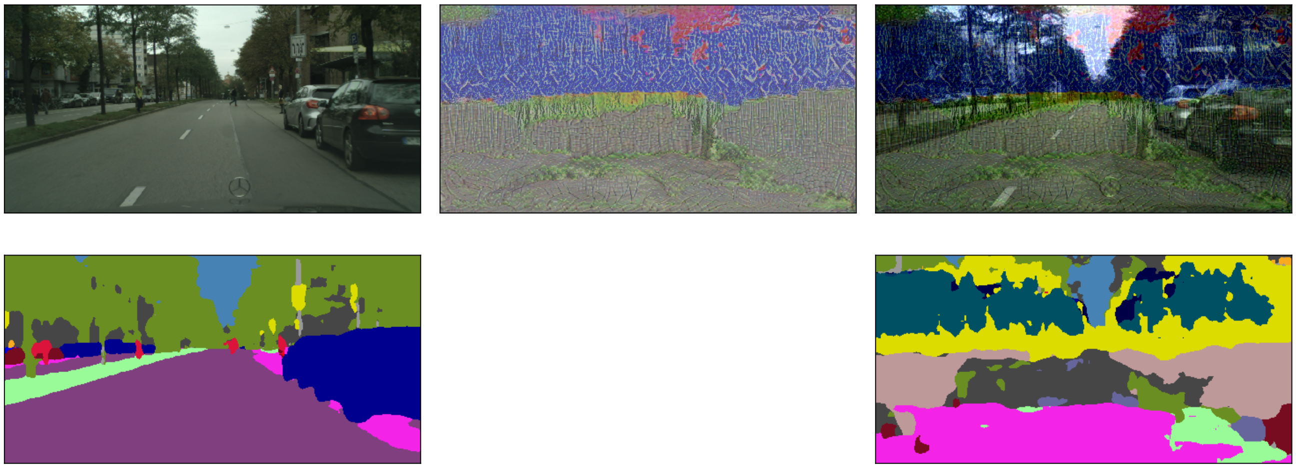

Figure A9. Untargeted universal perturbations on Cityscapes for a model trained with shared adversarial training. The shown perturbation upper bounds the robustness of the model to $\varepsilon_{uni}(\delta = 0.95) \leq 47.8$. Top row shows original image, universal perturbation, and perturbed image. Bottom row shows prediction on original image, target segmentation, and prediction on perturbed image.

Table A1. Universal robustness and classification accuracy of ResNet20 trained with standard adversarial training ($s = 1$) and *shared adversarial training* $s \in \{8, 64\}$ against S-PGD universal perturbations on CIFAR10 under different range of attack parameters $\varepsilon$ and $\sigma$. The pictorial representation of these entries are depicted in Figure 3 (left). The bold entries represent the model trained with *shared adversarial training* $s = 64$ that yields threefold increase in robustness when compared to undefended model with a drop of less than $3.5\%$ in accuracy.

| $\sigma$ | $s$ | | $\varepsilon = 2$ | $\varepsilon = 4$ | $\varepsilon = 6$ | $\varepsilon = 8$ | $\varepsilon = 10$ | $\varepsilon = 14$ | $\varepsilon = 18$ | $\varepsilon = 22$ | $\varepsilon = 26$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.3 | 1 | Acc.(%) | 92.16 | 90.42 | 88.76 | 84.02 | 78.54 | 73.61 | 70.32 | 67.63 | 61.73 |
| | | $\varepsilon_{uni}(\delta = 0.75)$ | 28.14 | 33.62 | 34.37 | 40.59 | 38.85 | 41.59 | 56.03 | 48.06 | 48.06 |
| | 8 | Acc.(%) | 93.04 | 92.05 | 91.34 | 90.75 | 89.09 | 85.87 | 82.68 | 79.69 | 76.35 |
| | | $\varepsilon_{uni}(\delta = 0.75)$ | 23.41 | 26.89 | 32.62 | 34.12 | 32.87 | 40.59 | 45.82 | 47.31 | 54.54 |
| | 64 | Acc.(%) | 93.72 | 93.36 | 92.84 | 92.47 | 91.90 | 90.89 | 88.93 | 86.13 | 83.89 |
| | | $\varepsilon_{uni}(\delta = 0.75)$ | 19.42 | 22.66 | 25.90 | 27.89 | 32.37 | 39.10 | 40.34 | 42.58 | 46.32 |
| 0.5 | 1 | Acc.(%) | 91.77 | 89.70 | 87.60 | 85.44 | 82.64 | 71.81 | 66.61 | 64.46 | 62.16 |
| | | $\varepsilon_{uni}(\delta = 0.75)$ | 30.63 | 38.10 | 39.35 | 42.33 | 47.31 | 61.26 | 55.78 | 93.27 | 63.75 |
| | 8 | Acc.(%) | 92.62 | 91.69 | 90.86 | 90.11 | 89.04 | 86.16 | 82.92 | 80.54 | 75.99 |
| | | $\varepsilon_{uni}(\delta = 0.75)$ | 23.16 | 29.88 | 33.87 | 36.61 | 38.10 | 42.83 | 43.83 | 53.54 | 64.00 |
| | 64 | Acc.(%) | 93.55 | 93.09 | 92.50 | 91.99 | 91.62 | 90.58 | 88.52 | 86.65 | 84.13 |
| | | $\varepsilon_{uni}(\delta = 0.75)$ | 20.67 | 23.91 | 28.14 | 29.88 | 38.10 | 37.60 | 41.34 | 47.31 | 51.05 |
| 0.7 | 1 | Acc.(%) | 91.55 | 89.07 | 86.58 | 84.47 | 81.97 | 78.75 | 73.29 | 64.15 | 63.63 |
| | | $\varepsilon_{uni}(\delta = 0.75)$ | 31.63 | 42.08 | 39.84 | 45.82 | 46.82 | 54.04 | 63.75 | 95.63 | 89.65 |
| | 8 | Acc.(%) | 92.51 | 91.29 | 90.23 | 89.12 | 88.08 | 85.70 | 83.74 | 80.27 | 78.00 |
| | | $\varepsilon_{uni}(\delta = 0.75)$ | 26.40 | 30.38 | 35.11 | 39.84 | 44.08 | 45.07 | 45.82 | 50.55 | 54.54 |
| | 64 | Acc.(%) | 93.34 | 92.81 | 92.27 | 91.67 | 91.30 | **89.94** | 88.25 | 85.96 | 83.94 |
| | | $\varepsilon_{uni}(\delta = 0.75)$ | 20.67 | 23.91 | 27.89 | 30.88 | 36.11 | **44.08** | 43.58 | 51.05 | 50.80 |
| 0.9 | 1 | Acc.(%) | 91.45 | 88.54 | 85.51 | 82.97 | 80.38 | 76.35 | 72.94 | 70.14 | 68.20 |
| | | $\varepsilon_{uni}(\delta = 0.75)$ | 32.87 | 38.85 | 46.07 | 50.55 | 54.54 | 58.27 | 63.25 | 64.50 | 59.77 |
| | 8 | Acc.(%) | 92.27 | 90.97 | 89.89 | 88.63 | 86.89 | 84.25 | 81.63 | 79.13 | 76.56 |
| | | $\varepsilon_{uni}(\delta = 0.75)$ | 26.40 | 33.12 | 37.10 | 40.34 | 44.08 | 45.82 | 47.81 | 55.28 | 58.02 |
| | 64 | Acc.(%) | 93.18 | 92.61 | 92.12 | 91.42 | 90.85 | 89.32 | 87.41 | 85.26 | 83.10 |
| | | $\varepsilon_{uni}(\delta = 0.75)$ | 22.16 | 25.40 | 29.63 | 32.87 | 36.61 | 43.33 | 46.07 | 45.32 | 54.04 |

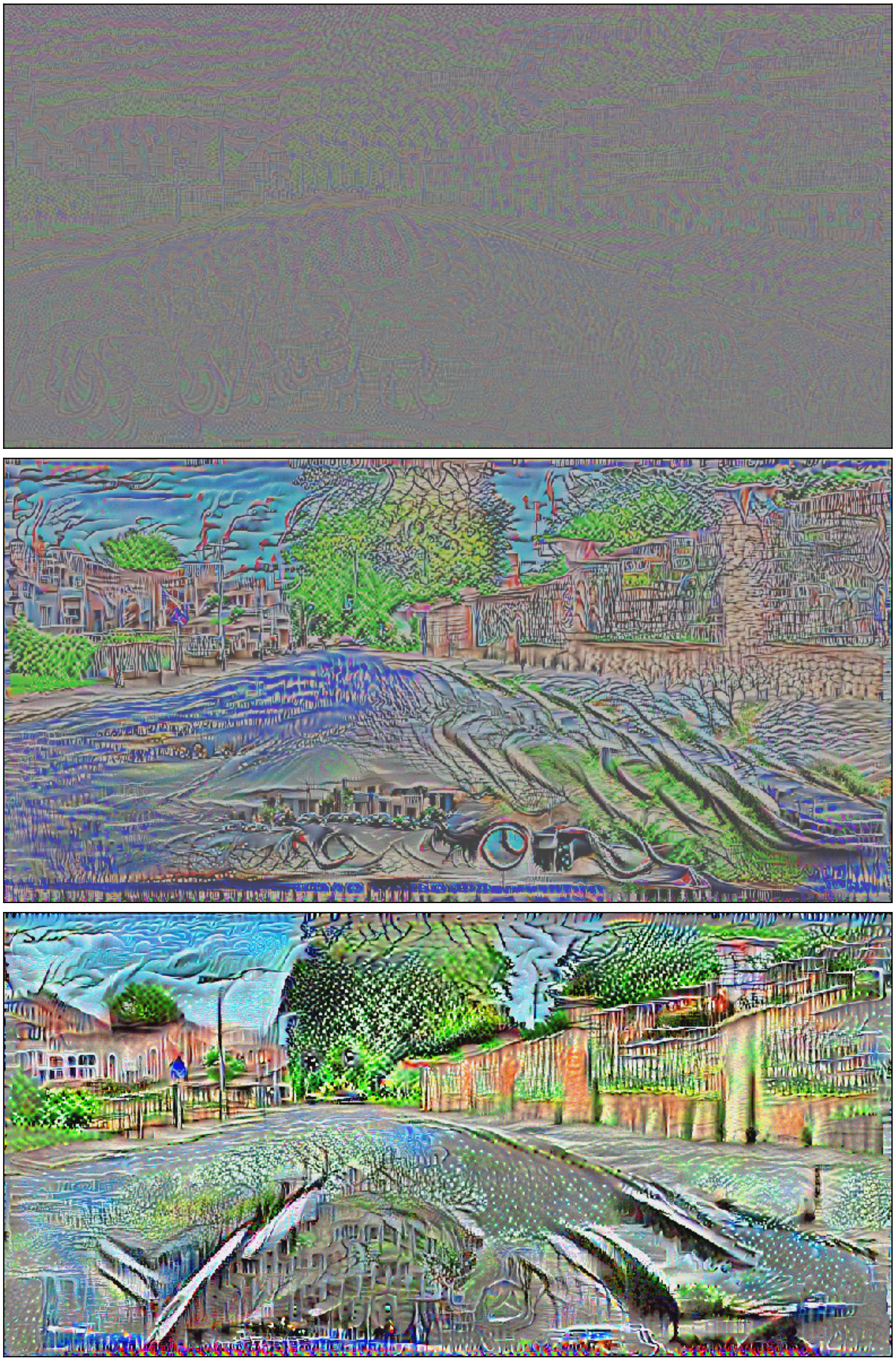

Figure A10. Illustration of targeted universal perturbation for empirical risk minimization (top), adversarial training (middle), and shared adversarial training (bottom).

Table A2. Universal robustness and classification accuracy of WRN-50-2-bottleneck trained with standard adversarial training ($s = 1$) and *shared adversarial training* ($s = 32$) against S-PGD universal perturbations on a subset of ImageNet (200 classes) under different range of attack parameters $\varepsilon$ and $\sigma$. The pictorial representation of these entries are depicted in Figure 4. The bold entries represent the model trained with *shared adversarial training* $s = 32$ that yields threefold increase in robustness when compared to undefended model with a drop of less than $5\%$ in accuracy.

| $\sigma$ | $s$ | | $\varepsilon = 2$ | $\varepsilon = 4$ | $\varepsilon = 6$ | $\varepsilon = 8$ | $\varepsilon = 10$ | $\varepsilon = 14$ | $\varepsilon = 18$ | $\varepsilon = 22$ | $\varepsilon = 26$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.5 | 1 | Acc.(%) | 75.81 | 74.49 | 71.14 | 71.48 | 68.78 | 66.25 | 64.05 | 59.04 | 57.54 |
| | | $\varepsilon_{uni}(\delta = 0.75)$ | 17.92 | 18.92 | 23.90 | 17.92 | 13.94 | 20.66 | 20.41 | 29.38 | 25.89 |
| | 32 | Acc.(%) | 77.50 | 77.23 | 75.46 | 72.77 | 68.43 | 64.21 | 56.48 | 54.58 | 48.30 |
| | | $\varepsilon_{uni}(\delta = 0.75)$ | 13.44 | 15.93 | 17.43 | 12.70 | 18.67 | 28.13 | 33.36 | 43.33 | 45.57 |
| 1.0 | 1 | Acc.(%) | 74.83 | 71.84 | 67.79 | 66.72 | 63.51 | 60.38 | 58.49 | 57.23 | 55.19 |
| | | $\varepsilon_{uni}(\delta = 0.75)$ | 20.17 | 23.90 | 24.90 | 28.88 | 27.89 | 30.13 | 30.62 | 32.12 | 32.37 |
| | 32 | Acc.(%) | 77.41 | 76.41 | 75.22 | 73.86 | **72.74** | 70.04 | 66.97 | 62.36 | 56.99 |
| | | $\varepsilon_{uni}(\delta = 0.75)$ | 14.94 | 17.43 | 19.67 | 21.41 | **25.64** | 28.88 | 30.13 | 29.38 | 33.36 |