Appendix

Algorithm 1 Multi-size Greedy Cutout

Input: Image label pair $(\mathbf{x}, y)$, vanilla classifier $\mathbb{F}$, mask sets $\mathbf{M}_{3\times3}, \mathbf{M}_{6\times6}$
Output: Augmented data $(\mathbf{x}'_1, y), (\mathbf{x}'_2, y)$

procedure ROUND-1($\mathbf{x}, \mathbb{F}, \mathbf{M}_{3\times3}, \mathbf{M}_{6\times6}$)
  loss $\leftarrow -1$.
  for each $\mathbf{m} \in \mathbf{M}_{3\times3}$ **do**
    if $\ell(\mathbb{F}(\mathbf{x} \odot \mathbf{m}, y) >$ loss **then**
      $\mathbf{M}_1 \leftarrow \mathbf{m}$ ▷ Mask with the largest loss
    end if
  end for
  $\widetilde{\mathbf{M}} \leftarrow$ {4 mask from $\mathbf{M}_{6\times6}$ that covers $\mathbf{M}_1$}.
  loss $\leftarrow -1$
  for each $\mathbf{m} \in \widetilde{\mathbf{M}}$ **do**
    if $\ell(\mathbb{F}(\mathbf{x} \odot \mathbf{m}, y) >$ loss **then**
      $\mathbf{M}_{11} \leftarrow \mathbf{m}$ ▷ Mask with the largest loss
    end if
  end for
  return $\mathbf{M}_1, \mathbf{M}_{11}$
end procedure

procedure ROUND-2($\mathbf{x}, \mathbb{F}, \mathbf{M}_{3\times3}, \mathbf{M}_{6\times6}$)
  $\mathbf{M}_1, \mathbf{M}_{11} \leftarrow$ ROUND-1($\mathbf{x}, \mathbb{F}, \mathbf{M}_{3\times3}, \mathbf{M}_{6\times6}$)
  loss $\leftarrow -1$
  for $\mathbf{m} \in \mathbf{M}_{3\times3}$ **do**
    if $\ell(\mathbb{F} \odot \mathbf{M}_{11} \odot \mathbf{m}) >$ loss **then**
      $\mathbf{m}^*_3 \leftarrow \mathbf{m}$ ▷ Mask with the largest loss
    end if
  end for
  $\mathbf{M}_2 \leftarrow \mathbf{M}_1 \odot \mathbf{m}^*_3$
  $\widehat{\mathbf{M}} \leftarrow$ {4 mask from $\mathbf{M}_{6\times6}$ that covers $\mathbf{m}^*_3$}.
  loss $\leftarrow -1$
  for each $\mathbf{m} \in \widehat{\mathbf{M}}$ **do**
    if $\ell(\mathbb{F} \odot \mathbf{M}_{11} \odot \mathbf{m}) >$ loss **then**
      $\mathbf{m}^*_6 \leftarrow \mathbf{m}$ ▷ Mask with the largest loss
    end if
  end for
  $\mathbf{M}_{22} \leftarrow \mathbf{M}_{11} \odot \mathbf{m}^*_6$
  $\mathbf{x}'_1 \leftarrow \mathbf{x} \odot \mathbf{M}_2$ ▷ Applying mask to augment data
  $\mathbf{x}'_2 \leftarrow \mathbf{x} \odot \mathbf{M}_{22}$ ▷ Applying mask to augment data
  return $(\mathbf{x}'_1, y), (\mathbf{x}'_2, y)$
end procedure

A Datasets

We use five popular image classfication datasets ranging from low-resolution to high resolution images.
(1) **ImageNet:** ImageNet is an image classification dataset Deng et al. (2009) with 1000 classes. It has 1.3 million training images and 50k validation images.
(2) **ImageNette:** ImageNette Howard et al. (2020) is a 10-class subset of ImageNet with 9469 training images and 3925 validation images.
(3) **CIFAR-10:** CIFAR-10 Krizhevsky et al. (2009) is a benchmark dataset for low-resolution image classification. The CIFAR-10 dataset consists of 60k 32x32 colour images in 10 classes, with 6k images per class.

There are 50k training images and, 10k test images.
(4) **CIFAR-100:** This dataset Krizhevsky et al. (2009) is just like the CIFAR-10, except it has 100 classes containing 600 images each. There are 500 training images and 100 testing images per class. The 100 classes in the CIFAR-100 are grouped into 20 superclasses.
(5) **SVHN:** Street View House Numbers (SVHN) Netzer et al. (2011) is a digit classification benchmark dataset that contains 600,000 32×32 RGB images of printed digits (from 0 to 9) cropped from pictures of house number plates.

Models: We use image classifiers from three different architecture families.
(1) **ResNet:** ResNets He et al. (2016) are deep neural networks which use skip connections. This approach makes it possible to train the network on thousands of layers without affecting performance. We use the ResNetV2-50x1 model from the timm Wightman et al. (2021) library.
(2) **Vision Transformers (ViT):** Convolutional Nets are designed based on inductive biases like translation invariance and a locally restricted receptive field. Unlike them, transformers are based on a self-attention mechanism that learns the relationships between elements of a sequence. We use ViT-B16-224 model.
(3) **ConvNeXt:** ConvNeXt Liu et al. (2022) is a pure convolutional model (ConvNet), inspired by the design of Vision Transformers. The design starts from a standard ResNet (e.g. ResNet50) and gradually "modernizes" the architecture to the construction of a hierarchical vision Transformer (e.g. Swin-T Liu et al. (2021)). We use the ConvNeXt_tiny_in22ft1k model from timm. It is trained on ImageNet-22k and fine-tuned on ImageNet-1k.

Table A1: Comparing certified robust accuracy of different masking strategies at two different mask set configurations $\mathbf{M}_{3\times3}$ and $\mathbf{M}_{6\times6}$ across different datasets on ViT. Certification pixels used 3% for ImageNette and ImageNet, and 2.4% for CIFAR-10, CIFAR-100 and SVHN. [1]

	Method	#passes	Mask set $\mathbf{M}_{3\times3}$					Mask set $\mathbf{M}_{6\times6}$				
			ImageNette	ImageNet	CIFAR-10	CIFAR-100	SVHN	ImageNette	ImageNet	CIFAR-10	CIFAR-100	SVHN
ViT	$\text{rand}_{3\times3}$	0	94.3	52.7	83.0	59.8	53.0	96.5	56.7	88.3	67.9	67.1
	$\text{rand}_{6\times6}$	0	93.6	50.6	79.5	54.9	42.2	96.0	56.7	86.0	64.4	61.3
	rand	0	94.7	52.7	83.1	60.1	52.0	96.4	58.3	88.4	68.5	67.7
	$\text{grid}_{3\times3}$	45	95.3	57.9	88.0	67.3	62.7	97.3	60.5	92.2	74.6	74.0
	$\text{grid}_{6\times6}$	666	95.3	-	-	-	61.1	97.2	-	-	-	78.1
	$\text{greedy}_{3\times3}$	17	95.1	58.3	87.9	66.6	62.5	97.2	61.2	92.2	74.2	73.8
	$\text{greedy}_{6\times6}$	71	95.1	56.6	86.2	64.4	60.6	97.5	63.8	91.8	74.3	77.9
	greedy (Ours)	25	95.5	57.7	87.5	66.0	63.3	97.3	62.3	92.0	74.5	76.8

Table A2: Table listing the number of forward passes needed in each batch training for grid search and our *Greedy Cutout* approach.

Method	# forward passes/batch training
$\text{grid search}_{3\times3}$	45 unique among $9\times9=81$
$\text{grid search}_{6\times6}$	666 unique among $36\times36=1296$
$\text{greedy}_{3\times3}$	9 (round 1) + 8 (round 2) = 17
$\text{greedy}_{6\times6}$	36 (round 1) + 35 (round 2) = 71
greedy (Ours)	13 (round 1) + 13 (round 2) = 26

B Masks selected

Figure A1 depicts the masks selected from exhaustive cutout and our Multi-size Greedy Cutout from both $\mathbf{M}_{3\times3}$ and $\mathbf{M}_{6\times6}$ on ImageNet training samples. It can be observed in the figure that masks could potentially cover the entire object (e.g. last row) and training on such instances would limit model learning capabilities. On the contrary, we observe that this training scheme encourages for higher robust accuracy as shown in Table 3, where $grid_{3\times3}$ and our Multi-size Greedy Cutout (greedy) achieve comparable results for $\mathbf{M}_{3\times3}$ and $grid_{6\times6}$ yield higher numbers than ours on $\mathbf{M}_{6\times6}$ with significantly higher training complexity. We

[1]Note that training with the masking strategy $greedy_{6\times6}$ on ImageNet is computationally costly and was not possible with available GPU resources. Same for Table 3.

Figure A1: Masks selected via cutout guided by exhaustive search and our Multi-size Greedy Cutout from both the mask sets $\mathbf{M}_{3\times3}$ and $\mathbf{M}_{6\times6}$ on ImageNet training samples.

hypothesize that images with partially covered objects are dominating in number and hence providing a strong training signal to the model, which making the noisy training signal from fully covered objects negligible.