Supplementary Material for:
Selection of Source Images Heavily Influences the Effectiveness of Adversarial Attacks

A L_p norms and perturbation visibility

Although we guarantee the discretization property, in order to maintain comparability with the literature, the perturbation amounts reported in the main text (both L_2 and L_∞) are calculated based on the assumption that pixel values lie in $[0,1]$. Based on this, we calculate the L_2 and L_∞ distance between two vectors with size $k = 3 \times 224 \times 224$ (channel $\times$ height $\times$ width) as follows:

$$L_2(\boldsymbol{x}, \hat{\boldsymbol{x}}) = ||\boldsymbol{x} - \hat{\boldsymbol{x}}||_2\,, \tag{1}$$
$$L_\infty(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \max(|\boldsymbol{x} - \hat{\boldsymbol{x}}|)\,, \tag{2}$$

where $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ represent an initial (source) image and its adversarial counterpart, respectively. In Figure I, we provide a number of qualitative examples that illustrate the measurement of perturbation visibility.

B Non-adversarial perturbations

In the main text, we compare the adversarial transferability of images modified through adversarial perturbation with that of images changed through non-adversarial noise. The different types of non-adversarial noise we employ are (1) uniform noise, (2) normal noise, and (3) change in contrast. For the aforementioned types of noise, we initialize a vector $\boldsymbol{p} = \boldsymbol{0} \in \mathbb{R}^k$ that has the same size as the input, filling its values as described below, with all non-adversarial perturbation generation methods respecting the L_∞ perturbation limit set for PGD, hereby using $\Pi_{\epsilon=38}$.

Uniform noise – Similar to the usage of PGD, we rely on an iterative approach for the application of uniform noise. As such, each of the elements of $\boldsymbol{p}$ is sampled from a uniform distribution $\mathcal{U}[-1,1]$. However, instead of using the values themselves, we use their signature, applying perturbation as follows:

$$[\hat{\boldsymbol{x}}]_{n+1} = \Pi_\epsilon([\hat{\boldsymbol{x}}]_n + [\boldsymbol{p}]_n)\,, [p_k]_n \sim \text{sign}(\mathcal{U}[-1,1])\,. \tag{3}$$

with $[\hat{\boldsymbol{x}}]_1 = \boldsymbol{x}$. Similar to the usage of PGD, if the "adversarial example" created this way does not achieve model-to-model transferability, we perform the same operation four more times.

Gaussian noise – Instead of an iterative approach, we follow a different methodology for the application of Gaussian noise. We sample only one noise vector, with every element of this vector originating from a Gaussian distribution with zero mean and standard deviation 10. We then apply this noise vector to the data point at hand as follows:

$$\hat{\boldsymbol{x}} = \Pi_\epsilon(\boldsymbol{x} + \boldsymbol{p})\,, p_k \sim \mathcal{N}(0, 10^2)\,. \tag{4}$$

If the resulting image does not achieve adversarial transferability, we perform the same operation up to ten times more, with newly sampled values from the same normal distribution.

Change in contrast – A change in contrast in the image domain means that all pixel values are modified with the same value. To that end, we evaluate all possible values within the allowed L_∞ limit, creating a set of adversarial examples originating from an input image as follows:

$$\widehat{\mathcal{X}} := \{\hat{\boldsymbol{x}}_b \,|\, \hat{\boldsymbol{x}}_b = \boldsymbol{x} + \mathbf{1} * b, b \in \{-38, \ldots, 38\}\}\,. \tag{5}$$

C Detailed transferability graphs

In Figure II and Figure III, we provide the model-to-model transferability plots presented in Figure 2 in the main text, but in a higher resolution and with more details. In addition to the untargeted transferability details provided in the aforementioned figures, in Figure IV, we provide the targeted adversarial transferability success of the produced adversarial examples.

In Figure V, Figure VI, and Figure VII, we provide detailed model-to-model transferability details for (left) fragile and (right) hard images, respectively, as identified with the help of non-adversarial perturbations.

In Figure 3(b) of the main text, we histogrammed $\overline{T}(\Theta, \widehat{\mathcal{X}}^{(A)}, \boldsymbol{y})$ for all adversarial examples, hereby displaying the transferability count of the source images. In Figure VIII, we provide the same information with $T(\Theta, \widehat{\mathcal{X}}^{(A)}, \boldsymbol{y})$, but specifically for adversarial examples created through the use of individual attacks.

D Correlation between transferability and perturbation

In Figure IX, we plot the adversarial transferability count for each source image, as obtained with $T(\Theta, \widehat{\mathcal{X}}^{(A)}, \boldsymbol{y})$, against the minimum required L_p perturbation to achieve adversarial transferability $D_{\{2,\infty\}}(\Theta, \widehat{\mathcal{X}}^{(\mathrm{A})})$, for all adversarial examples, as well as the subset of adversarial examples produced with individual attacks. Here, we observe a mild negative correlation between the added noise and the transferability count, where the adversarial examples originating from source images that achieve higher transferability counts are also the ones that require less perturbation.

E Required perturbation for adversarial transferability

In Figure 5 of the main text, we provided, for ViT-B, the $L_{\{2,\infty\}}$ norms of adversarial perturbations obtained through the usage of a number of source images, where this number is progressively reduced based on the transferability count of those images. From Figure X to Figure XVI, we provide the same results for the other models and for all adversarial attacks.

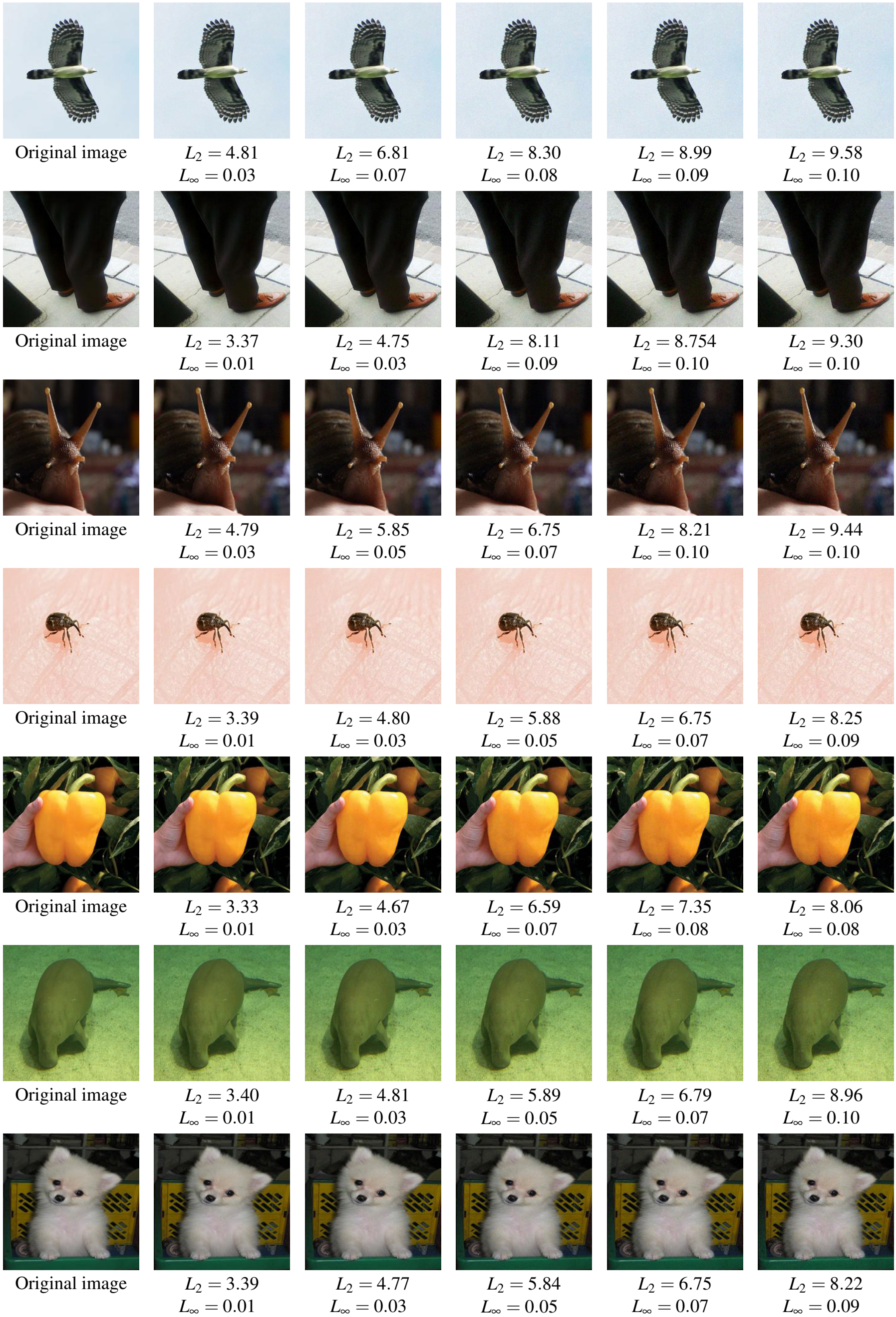


Figure I: Application of adversarial perturbations to images.

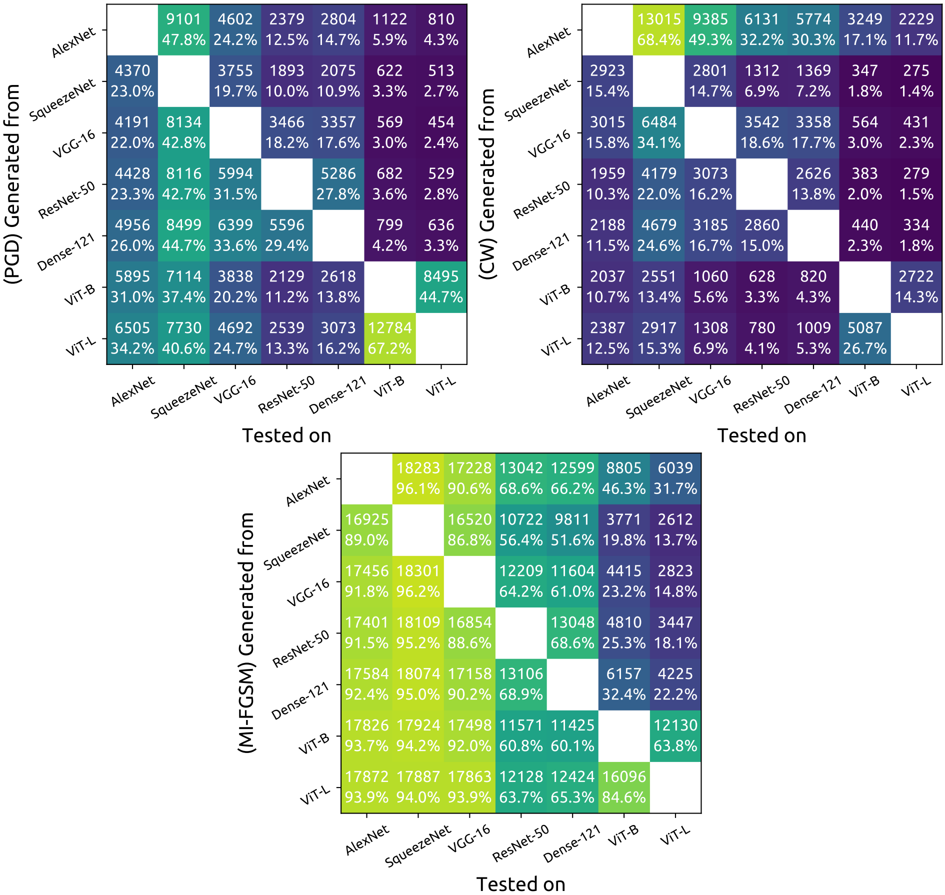


Figure II: Number (proportion) of source images that achieved (untargeted) adversarial transferability through the usage of (left) PGD, (right) CW, and (bottom) MI-FGSM. Adversarial examples are generated from the models listed on the y-axis and are tested on the models listed on the x-axis.

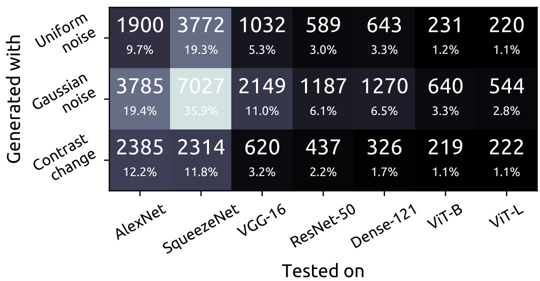


Figure III: Number (proportion) of source images that have their classification changed through the usage of non-adversarial perturbation.

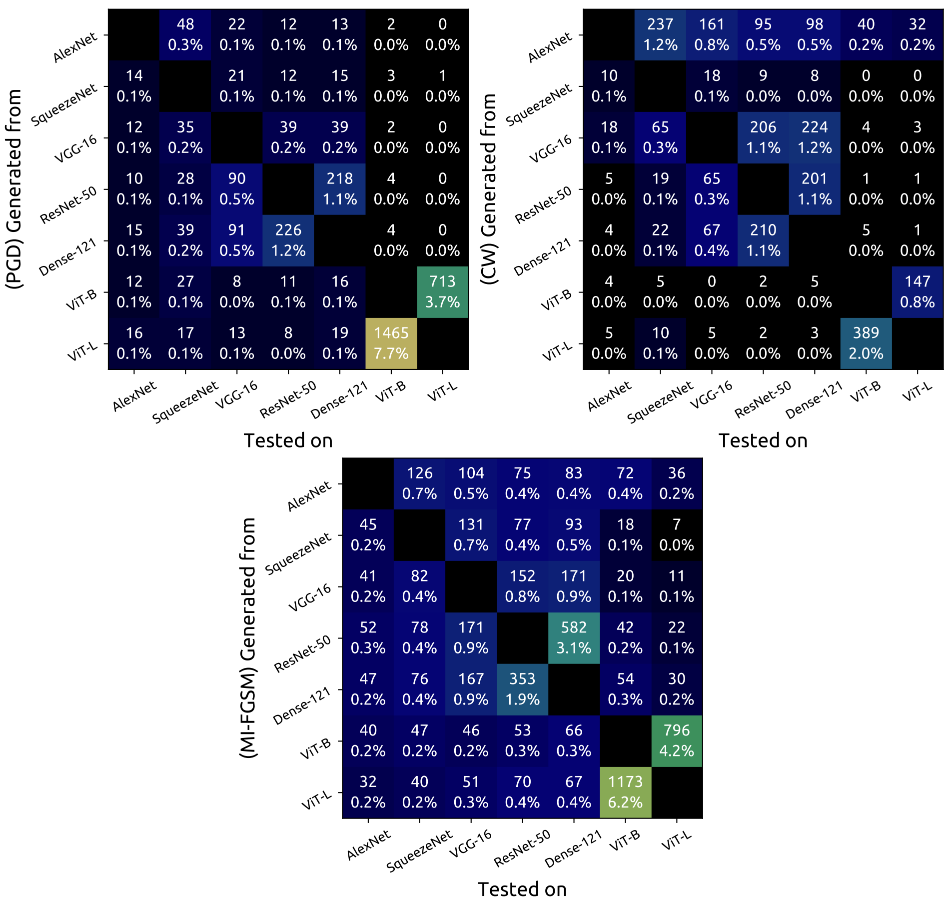
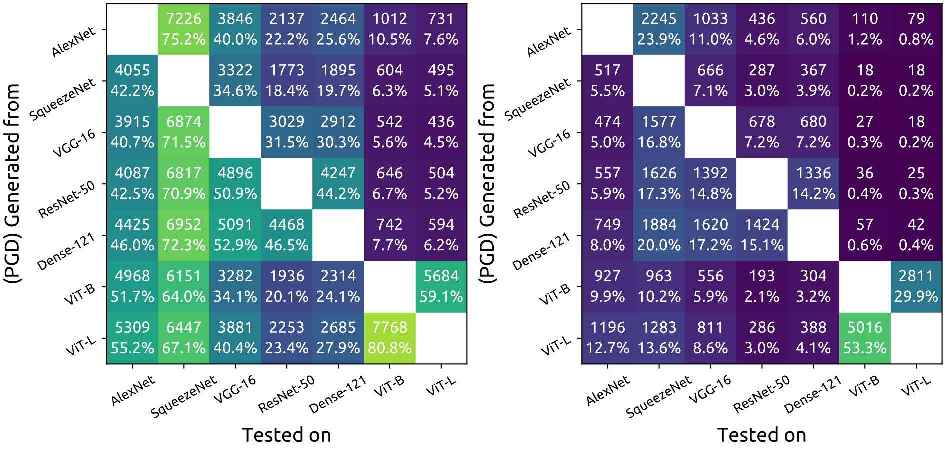
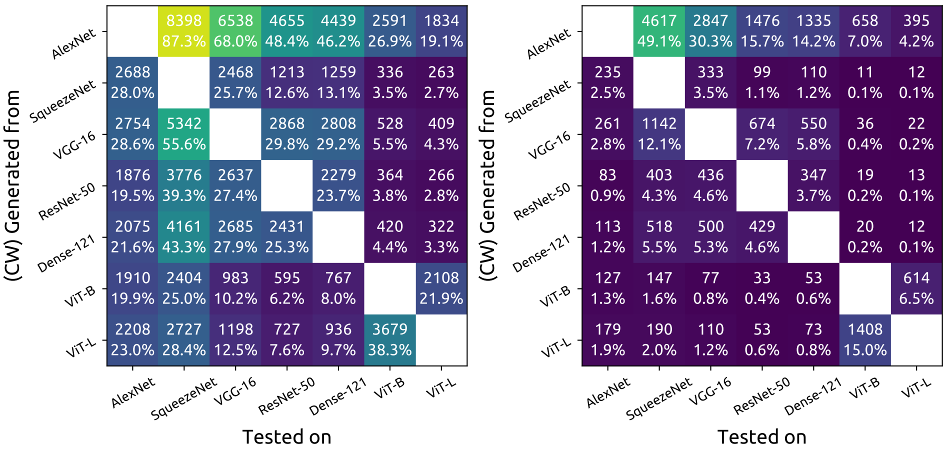


Figure IV: Number (proportion) of source images that achieved (targeted) adversarial transferability through the usage of (left) PGD, (right) CW, and (bottom) MI-FGSM. Adversarial examples are generated from the models listed on the *y*-axis and are tested on the models listed on the *x*-axis.



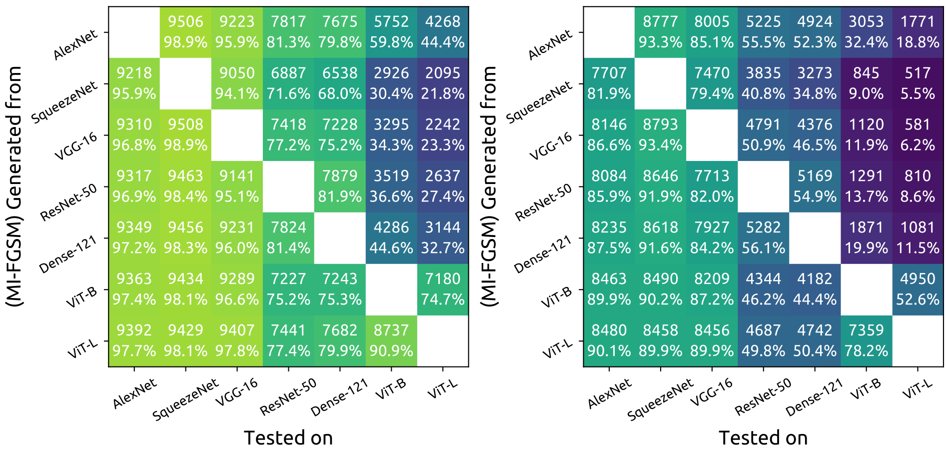
(a) Fragile images ($\mathbb{S}_f$) (b) Hard images ($\mathbb{S}_h$)

Figure V: Number (proportion) of source images that achieved (untargeted) adversarial transferability through the usage of **PGD** for source images taken from (left) $\underline{\mathbb{S}_f}$, and (right) $\underline{\mathbb{S}_h}$. Adversarial examples are generated from the models listed on the y-axis and are tested on the models listed on the x-axis.



(a) Fragile images ($\mathbb{S}_f$) (b) Hard images ($\mathbb{S}_h$)

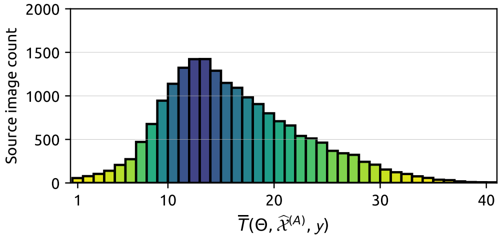
Figure VI: Number (proportion) of source images that achieved (untargeted) adversarial transferability through the usage of **CW** for source images taken from (left) $\underline{\mathbb{S}_f}$, and (right) $\underline{\mathbb{S}_h}$. Adversarial examples are generated from the models listed on the y-axis and are tested on the models listed on the x-axis.



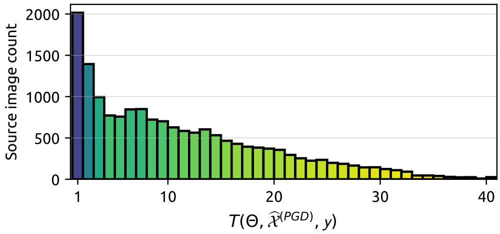
(a) Fragile images ($\mathbb{S}_f$)

(b) Hard images ($\mathbb{S}_h$)

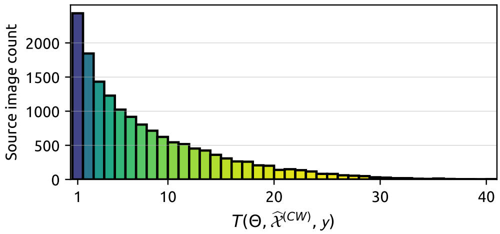
Figure VII: Number (proportion) of source images that achieved adversarial transferability through the usage of **MI-FGSM** for source images taken from (left) $\overline{\mathbb{S}_f}$, and (right) $\underline{\mathbb{S}_h}$. Adversarial examples are generated from the models listed on the y-axis and are tested on the models listed on the x-axis.



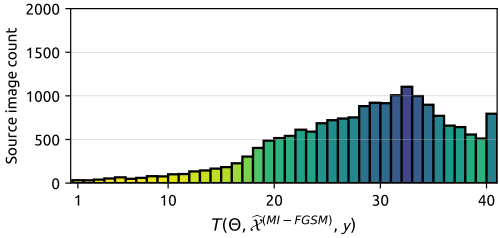
(a) All adversarial examples



(b) All adversarial examples produced with PGD

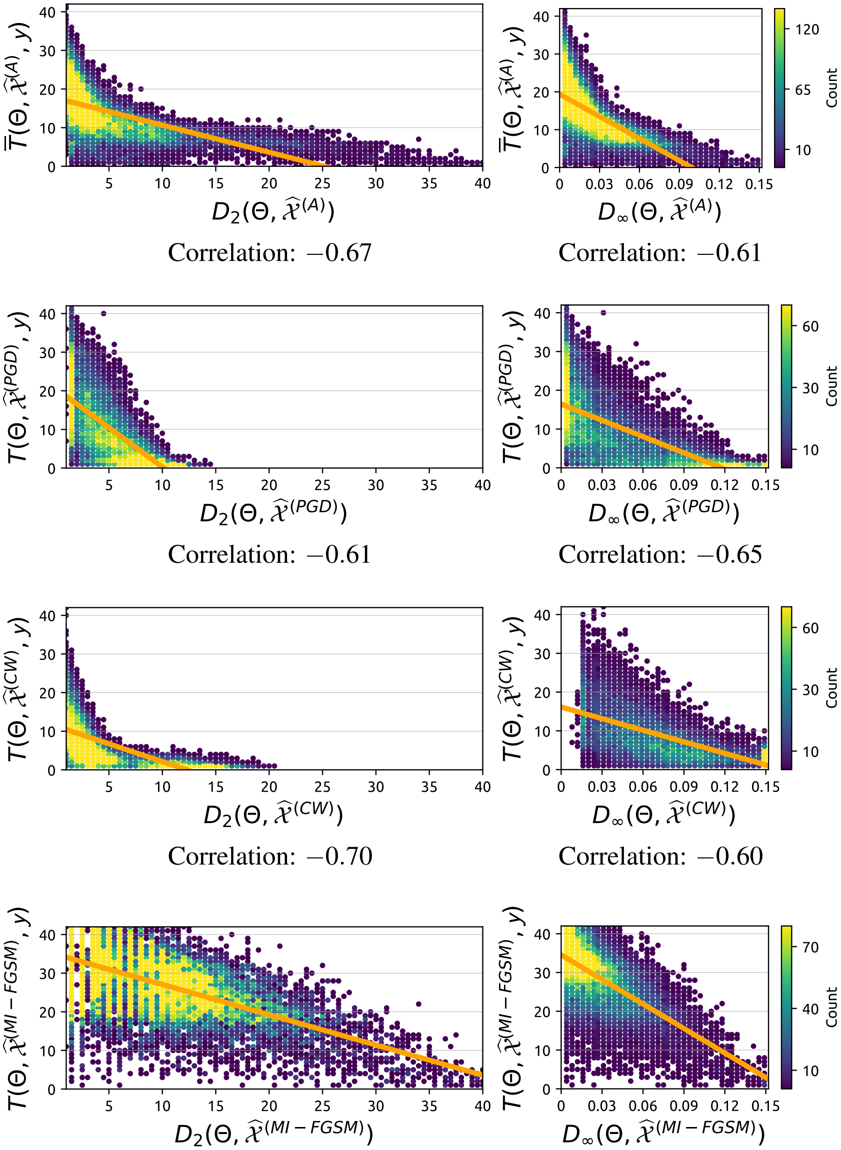


(c) All adversarial examples produced with CW



(d) All adversarial examples produced with MI-FGSM

Figure VIII: Histogram of source images and their transferability count according to $T(\Theta, \widehat{\mathcal{X}}^{(\mathrm{A})}, \boldsymbol{y})$, calculated with (top) all adversarial examples and (bottom three) individual attacks.



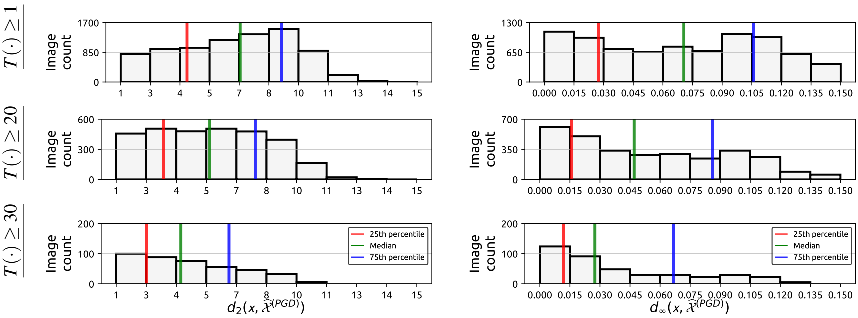
Correlation: -0.67 Correlation: -0.61

Correlation: -0.61 Correlation: -0.65

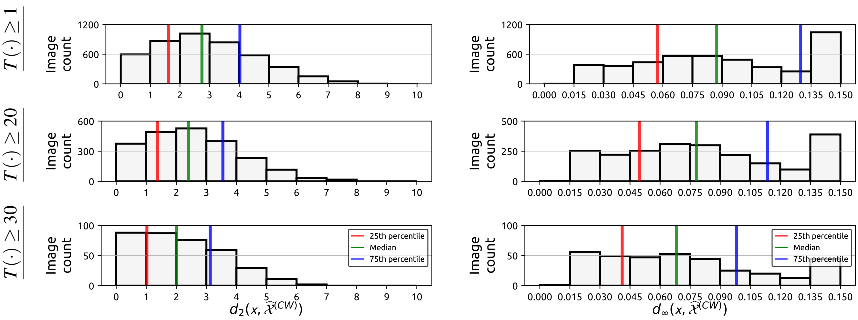
Correlation: -0.70 Correlation: -0.60

Correlation: -0.64 Correlation: -0.64

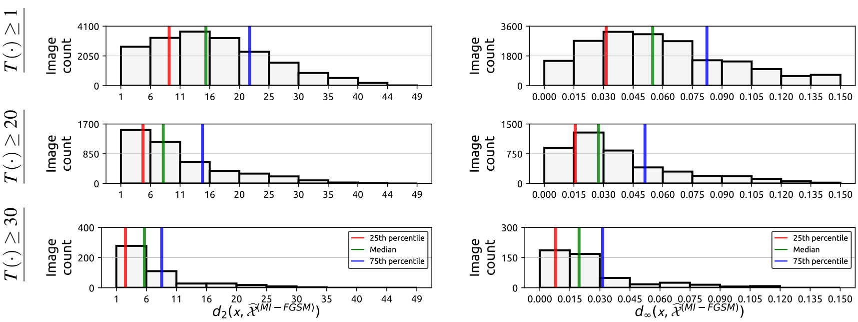
Figure IX: Scatter plot of $D_p(\Theta, \widehat{\mathcal{X}}^{(A)})$, the minimum amount of perturbation required for each source image, against adversarial transferability count $T(\Theta, \widehat{\mathcal{X}}^{(\mathrm{A})}, \boldsymbol{y})$, for $p = 2$ (left) and $p = \infty$ (right). The top graph shows the results for all adversarial examples, whereas the following ones present results for individual attacks. The regression line is shown in orange.



(a) Adversarial examples transferred to **AlexNet** with **PGD**.

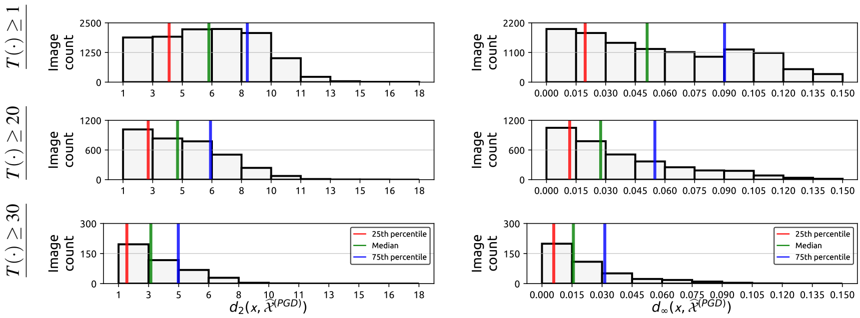


(b) Adversarial examples transferred to **AlexNet** with **CW**.

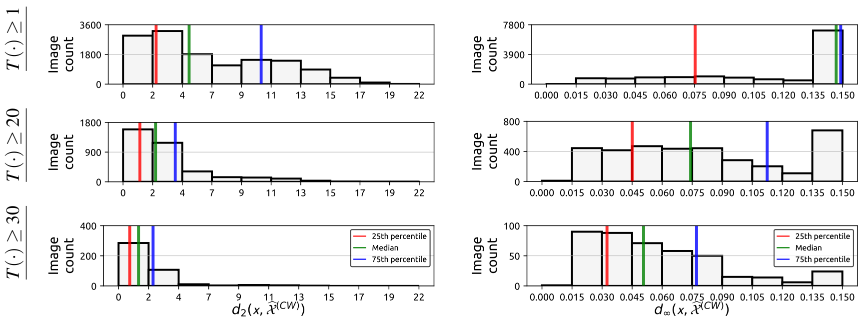


(c) Adversarial examples transferred to **AlexNet** with **MI-FGSM**.

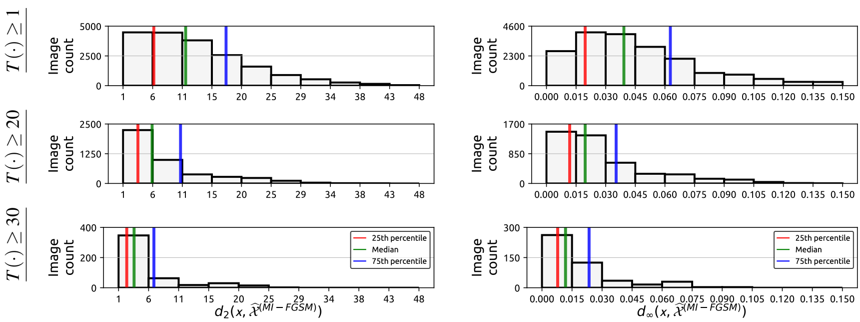
Figure X: Source images that achieved adversarial transferability to **AlexNet** are selected based on transferability count, with $T(\Theta, \widehat{\mathcal{X}}^{(\mathrm{A})}, \boldsymbol{y}) \geq \{1, 20, 30\}$. The minimum amount of perturbation required for creating adversarial examples from these source images is histogrammed, measuring the perturbation using $d_p(\boldsymbol{x}, \widehat{\mathcal{X}}^{(\mathrm{A})})$, with $p \in \{2, \infty\}$. The median perturbation, as well as the 25th and the 75th percentile, are provided in order to improve interpretability.



(a) Adversarial examples transferred to **SqueezeNet** with **PGD**.

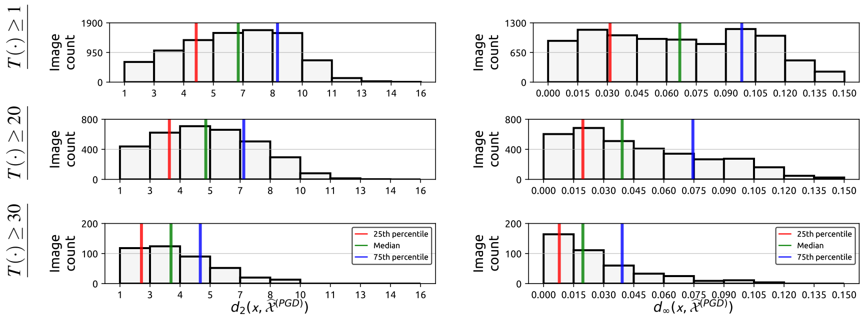


(b) Adversarial examples transferred to **SqueezeNet** with **CW**.

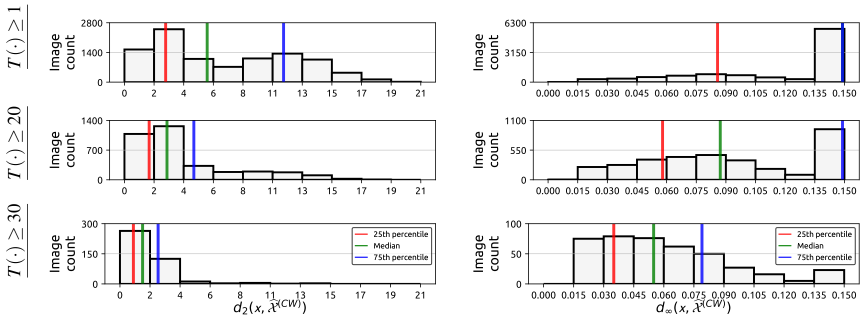


(c) Adversarial examples transferred to **SqueezeNet** with **MI-FGSM**.

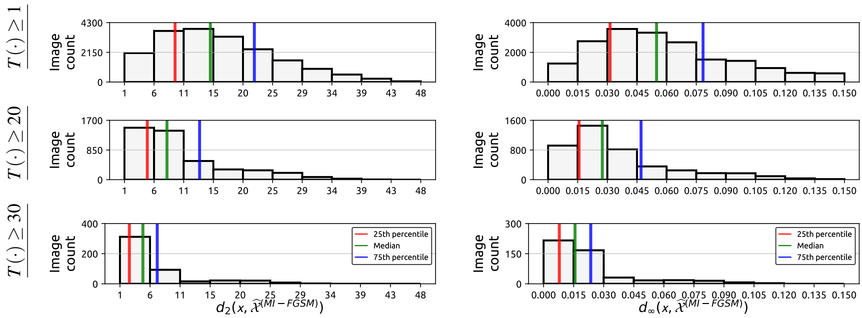
Figure XI: Source images that achieved adversarial transferability to **SqueezeNet** are selected based on transferability count, with $T(\Theta, \widehat{\mathcal{X}}^{(\mathrm{A})}, \boldsymbol{y}) \geq \{1, 20, 30\}$. The minimum amount of perturbation required for creating adversarial examples from these source images is histogrammed, measuring the perturbation using $d_p(\boldsymbol{x}, \widehat{\mathcal{X}}^{(\mathrm{A})})$, with $p \in \{2, \infty\}$. The median perturbation, as well as the 25th and the 75th percentile, are provided in order to improve interpretability.



(a) Adversarial examples transferred to **VGG-16** with **PGD**.

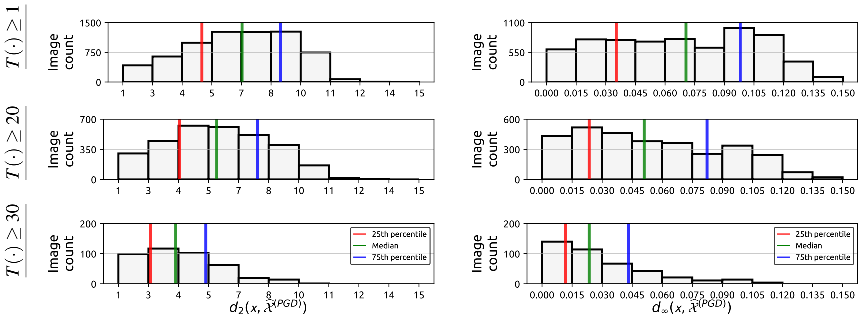


(b) Adversarial examples transferred to **VGG-16** with **CW**.

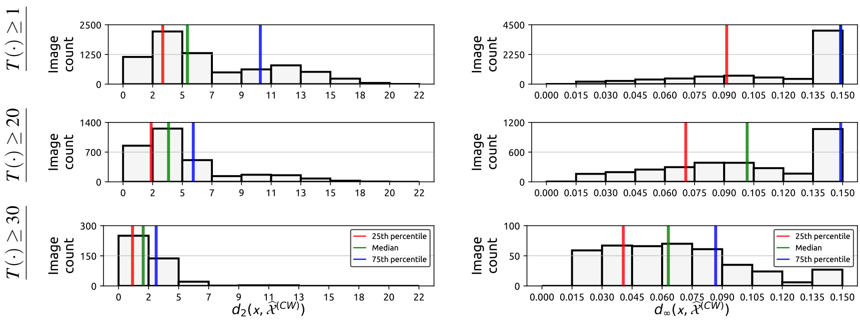


(c) Adversarial examples transferred to **VGG-16** with **MI-FGSM**.

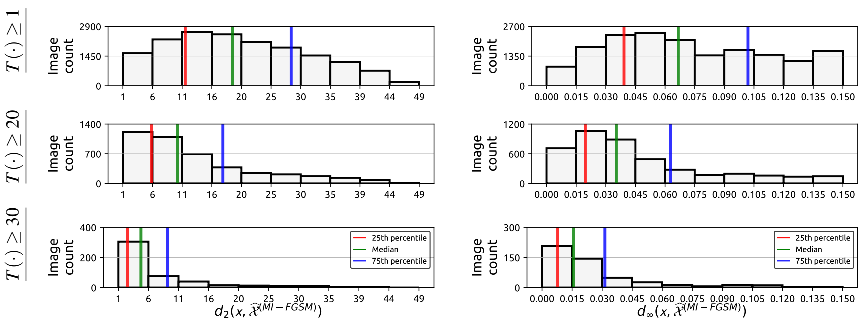
Figure XII: Source images that achieved adversarial transferability to **VGG-16** are selected based on transferability count, with $T(\Theta, \widehat{\mathcal{X}}^{(A)}, \boldsymbol{y}) \geq \{1, 20, 30\}$. The minimum amount of perturbation required for creating adversarial examples from these source images is histogrammed, measuring the perturbation using $d_p(\boldsymbol{x}, \widehat{\mathcal{X}}^{(A)})$, with $p \in \{2, \infty\}$. The median perturbation, as well as the 25th and the 75th percentile, are provided in order to improve interpretability.



(a) Adversarial examples transferred to **ResNet-50** with **PGD**.

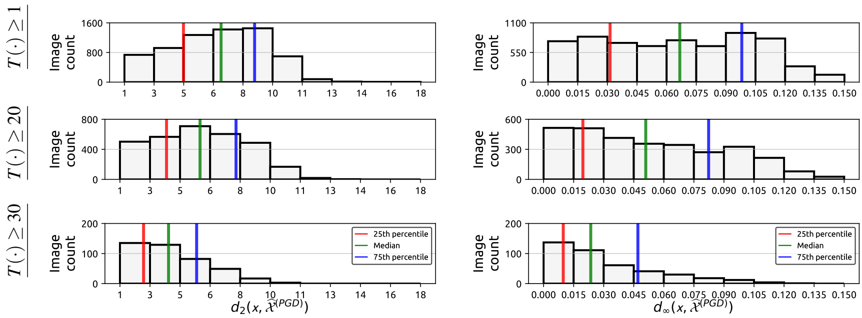


(b) Adversarial examples transferred to **ResNet-50** with **CW**.

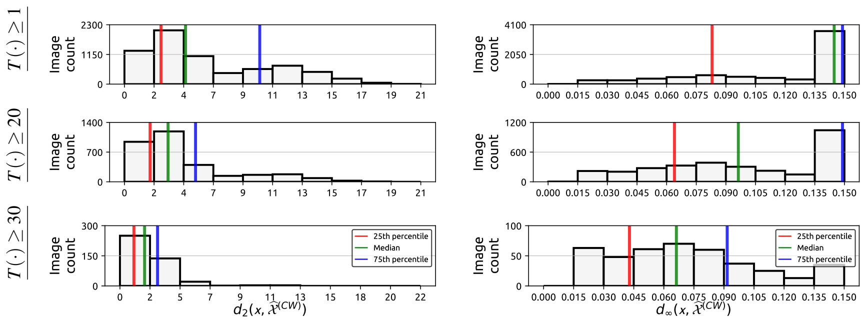


(c) Adversarial examples transferred to **ResNet-50** with **MI-FGSM**.

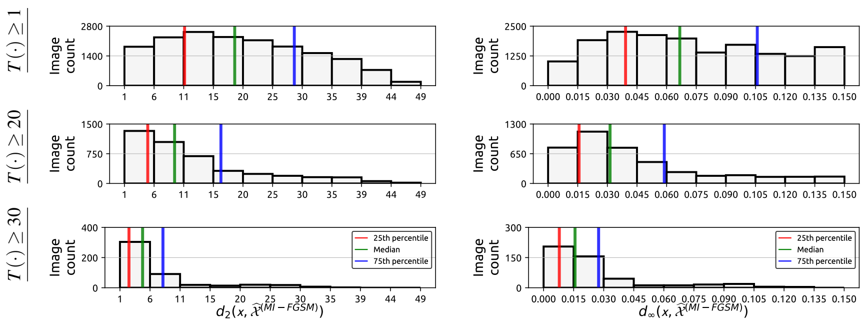
Figure XIII: Source images that achieved adversarial transferability to **ResNet-50** are selected based on transferability count, with $T(\Theta, \widehat{\mathcal{X}}^{(\mathrm{A})}, \boldsymbol{y}) \geq \{1, 20, 30\}$. The minimum amount of perturbation required for creating adversarial examples from these source images is histogrammed, measuring the perturbation using $d_p(\boldsymbol{x}, \widehat{\mathcal{X}}^{(\mathrm{A})})$, with $p \in \{2, \infty\}$. The median perturbation, as well as the 25th and the 75th percentile, are provided in order to improve interpretability.



(a) Adversarial examples transferred to **DenseNet-121** with **PGD**.

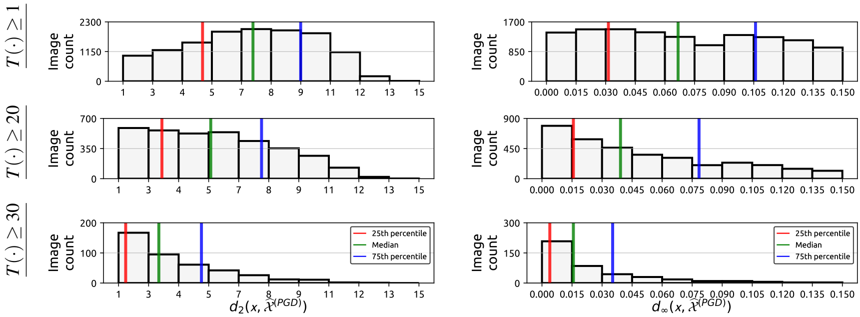


(b) Adversarial examples transferred to **DenseNet-121** with **CW**.

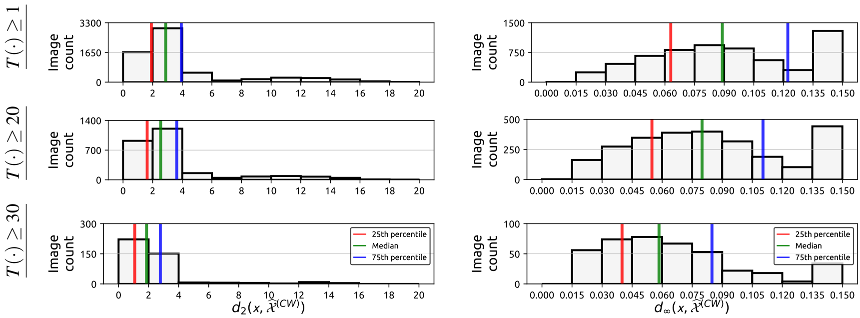


(c) Adversarial examples transferred to **DenseNet-121** with **MI-FGSM**.

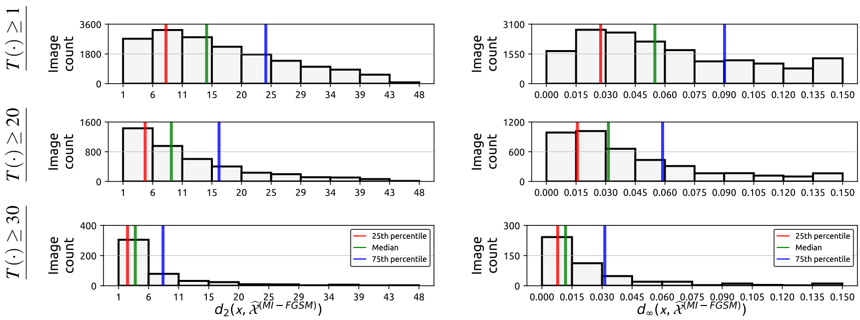
Figure XIV: Source images that achieved adversarial transferability to **DenseNet-121** are selected based on transferability count, with $T(\Theta, \widehat{\mathcal{X}}^{(A)}, \boldsymbol{y}) \geq \{1, 20, 30\}$. The minimum amount of perturbation required for creating adversarial examples from these source images is histogrammed, measuring the perturbation using $d_p(\boldsymbol{x}, \widehat{\mathcal{X}}^{(A)})$, with $p \in \{2, \infty\}$. The median perturbation, as well as the 25th and the 75th percentile, are provided in order to improve interpretability.



(a) Adversarial examples transferred to **ViT-B** with **PGD**.

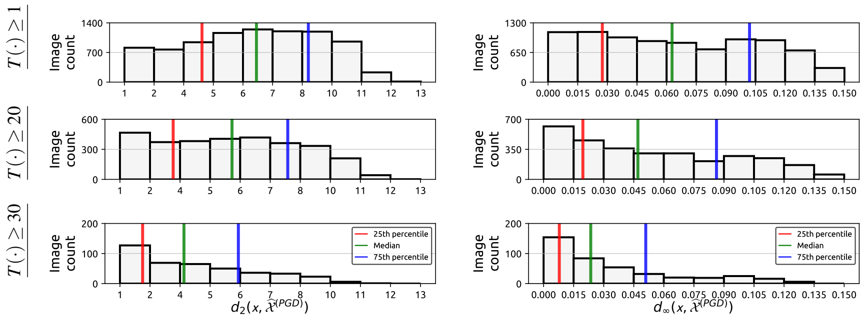


(b) Adversarial examples transferred to **ViT-B** with **CW**.

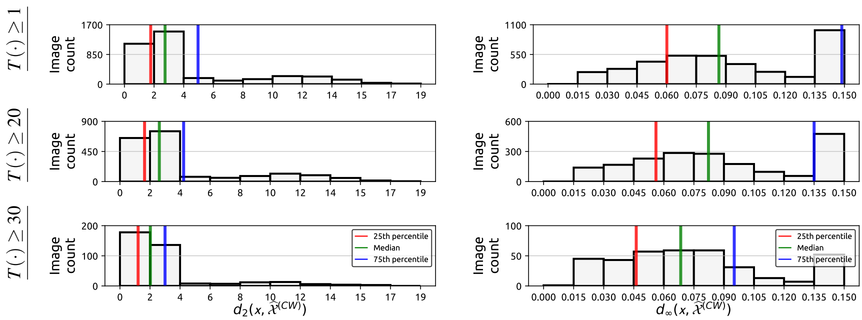


(c) Adversarial examples transferred to **ViT-B** with **MI-FGSM**.

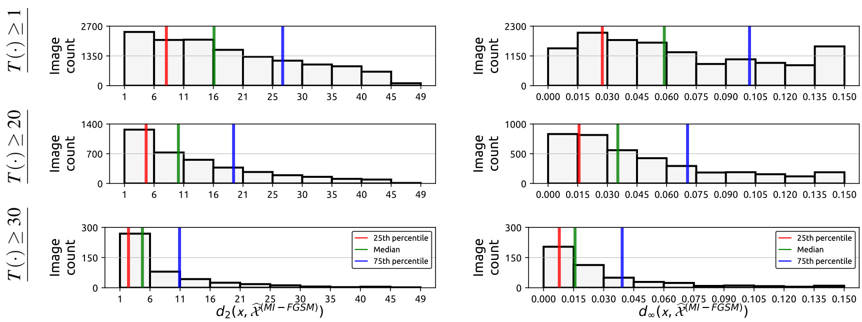
Figure XV: Source images that achieved adversarial transferability to **ViT-B** are selected based on transferability count, with $T(\Theta, \widehat{\mathcal{X}}^{(\mathrm{A})}, \boldsymbol{y}) \geq \{1, 20, 30\}$. The minimum amount of perturbation required for creating adversarial examples from these source images is histogrammed, measuring the perturbation using $d_p(\boldsymbol{x}, \widehat{\mathcal{X}}^{(\mathrm{A})})$, with $p \in \{2, \infty\}$. The median perturbation, as well as the 25th and the 75th percentile, are provided in order to improve interpretability.



(a) Adversarial examples transferred to **ViT-L** with **PGD**.



(b) Adversarial examples transferred to **ViT-L** with **CW**.



(c) Adversarial examples transferred to **ViT-L** with **MI-FGSM**.

Figure XVI: Source images that achieved adversarial transferability to **ViT-L** are selected based on transferability count, with $T(\Theta, \widehat{\mathcal{X}}^{(A)}, \boldsymbol{y}) \geq \{1, 20, 30\}$. The minimum amount of perturbation required for creating adversarial examples from these source images is histogrammed, measuring the perturbation using $d_p(\boldsymbol{x}, \widehat{\mathcal{X}}^{(A)})$, with $p \in \{2, \infty\}$. The median perturbation, as well as the 25th and the 75th percentile, are provided in order to improve interpretability.

F Error estimates

In the main text, we briefly mentioned the usage of a number of error estimates in order to measure mistakes made in the prediction of source images. We denote with $\boldsymbol{y}$ the true probabilistic categorical distribution associated with a data point $\boldsymbol{x}$ and assume that $c = \arg\max(\boldsymbol{y})$ is the true class and that $\hat{\boldsymbol{y}} = P(\boldsymbol{\theta}, \boldsymbol{x})$ is the prediction obtained with a model described by its parameters $\boldsymbol{\theta}$. The error estimates are then defined, in the context of ImageNet, as follows:

$$\mathrm{MSE}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{1}{1,000} \sum_{k=0}^{1,000} \left(y_k - \hat{y}_k\right)^2, \tag{6}$$

$$\mathrm{Q}(\hat{\boldsymbol{y}}) = \frac{\max_{k \neq c}(\hat{y}_k)}{\max_c(\hat{y}_c)}, \tag{7}$$

$$\mathrm{WD}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \inf_{\pi \in \mathcal{P}(\hat{\boldsymbol{y}}, \boldsymbol{y})} \int_{\mathbb{R} \times \mathbb{R}} |\hat{\boldsymbol{y}} - \boldsymbol{y}| \, d\pi(\hat{\boldsymbol{y}}, \boldsymbol{y}), \tag{8}$$

with $\mathcal{P}(u, v)$ representing the set of probability distributions on $\mathbb{R} \times \mathbb{R}$, where the first factor has marginal distribution u and the second one marginal distribution v. Note that the fourth estimate used in the main paper, $1 - \max(P(\boldsymbol{\theta}, \boldsymbol{x}))$, corresponds to $\frac{1}{2}\mathrm{MAE}(\hat{\boldsymbol{y}}, \boldsymbol{y})$, since all source images in this study are initially correctly classified by all models. For this reason, we omit the mean absolute distance from the set of measured estimates.

From Table I to Table VII, and based on source image filtering, we provide results regarding the transferability and required perturbation for all models considered in this study, when the adversarial examples are generated from the model that has the highest transferability to the model under inspection according to Figure II.

G Categorical information

We could observe that a large number of adversarial examples are misclassified into categories that are semantically close to the categories of their source images. This leads to the following question: does a misclassification made for ImageNet, where the prediction is a semantically similar class (i.e., a brown dog breed is misclassified as another brown dog breed), carry the same weight as a misclassification made by an automated system in a self-driving car scenario (i.e., a human or a vehicle not identified)?

In Figure XVII, we provide a number of qualitative examples where the adversarial examples on the left are misclassified into the categories on the right. Note that both categories are semantically very similar to each other. As such, we believe an important item for future work is the analysis of misclassification categories, taking into account the semantic similarity of classes.

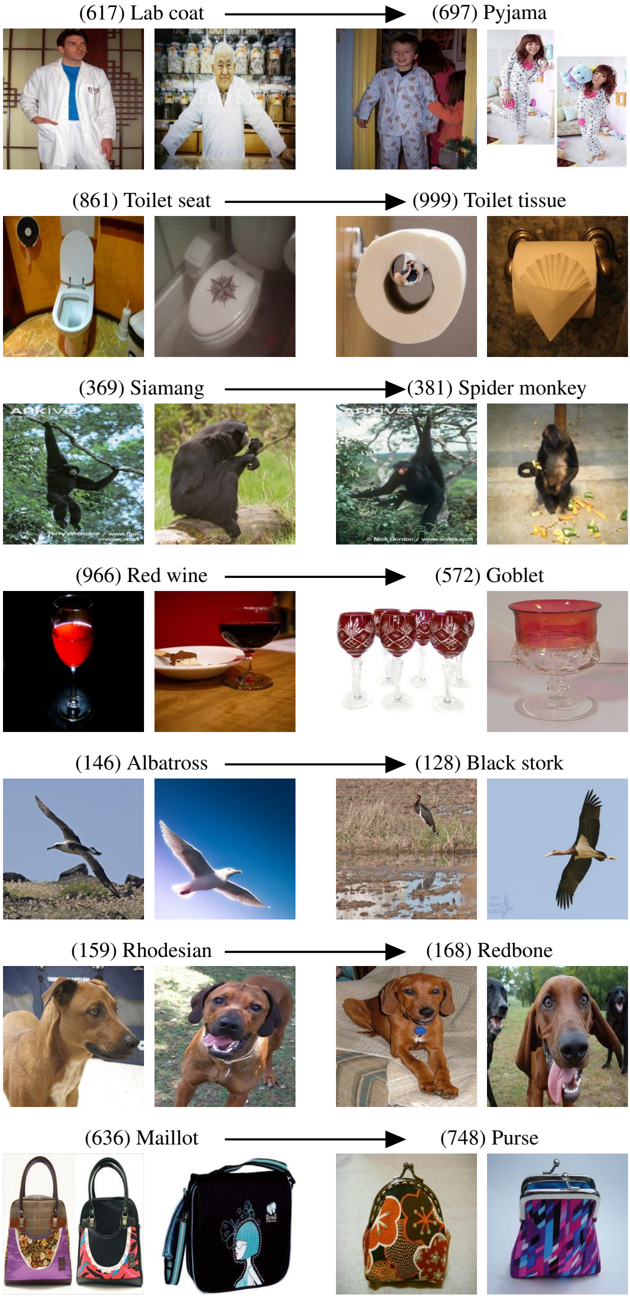


Figure XVII: Adversarial examples shown on the left are misclassified into similar categories shown on the right, by multiple models used in this study.

Table I: The lowest, the highest, and the average transferability, as well as the $L_{\{2,\infty\}}$ perturbations, are provided for adversarial examples created by randomly sampling $1,000$ source images $10,000$ times from the datasets provided in the second row. Statistics are provided using adversarial examples that are created from ViT-L and tested on **AlexNet**.

			All images	Hard images		Easy (fragile) images		Filtered images	
			$\mathbb{S}$	$\mathbb{S}_{Q<10}$	$\mathbb{S}_{Q<25}$	$\mathbb{S}_{Q>90}$	$\mathbb{S}_{Q>75}$	$\mathbb{S}\setminus(\mathbb{S}_{Q<10}\cup\mathbb{S}_{Q>90})$	$\mathbb{S}\setminus(\mathbb{S}_{Q<25}\cup\mathbb{S}_{Q>75})$
Source images in set:			19,025	1,904	4,758	1,904	4,758	15,219	9,511
Transferability	PGD	Low	28.1%	0.4%	2.2%	85.4%	71.0%	26.7%	25.8%
		Avg	34.2%	1.7%	4.7%	88.2%	75.1%	31.5%	30.1%
		High	40.4%	1.9%	6.2%	90.9%	79.4%	36.1%	33.4%
	CW	Low	6.1%	0.0%	0.0%	58.7%	36.3%	9.1%	2.3%
		Avg	12.5%	0.0%	0.2%	62.4%	41.5%	10.1%	4.0%
		High	18.2%	0.0%	0.6%	66.2%	48.0%	12.3%	5.1%
	MI-FGSM	Low	89.5%	76.2%	80.3%	97.2%	96.2%	92.6%	93.4%
		Avg	94.2%	80.6%	83.2%	98.9%	97.5%	94.1%	94.5%
		High	98.4%	84.1%	85.5%	99.5%	99.0%	96.3%	96.1%
Perturbation (L_2 / L_∞)	PGD	Low	7.15 / 0.07	7.81 / 0.09	9.08 / 0.10	5.43 / 0.04	6.40 / 0.06	8.04 / 0.09	8.71 / 0.10
		Avg	7.52 / 0.08	9.76 / 0.12	9.58 / 0.12	5.70 / 0.05	6.73 / 0.06	8.59 / 0.09	9.01 / 0.10
		High	8.50 / 0.09	11.3 / 0.14	10.75 / 0.13	5.95 / 0.05	7.01 / 0.07	9.07 / 0.10	9.52 / 0.11
	CW	Low	2.35 / 0.07	– / –	2.58 / 2.58	2.12 / 0.06	2.31 / 0.07	2.78 / 0.08	2.93 / 0.09
		Avg	2.69 / 0.08	– / –	2.95 / 0.13	2.23 / 0.07	2.54 / 0.07	3.15 / 0.08	3.41 / 0.09
		High	3.11 / 0.09	– / –	4.11 / 0.14	2.37 / 0.07	2.75 / 0.08	3.41 / 0.09	3.78 / 0.10
	MI-FGSM	Low	18.1 / 0.07	26.7 / 0.10	26.1 / 0.10	10.1 / 0.03	12.1 / 0.04	19.3 / 0.07	19.3 / 0.07
		Avg	19.8 / 0.07	27.1 / 0.10	26.5 / 0.10	10.7 / 0.03	13.3 / 0.05	19.8 / 0.07	19.9 / 0.07
		High	20.3 / 0.07	27.7 / 0.11	27.1 / 0.10	11.5 / 0.04	14.5 / 0.05	20.6 / 0.07	20.8 / 0.07

Table II: The lowest, the highest, and the average transferability, as well as the $L_{\{2,\infty\}}$ perturbations, are provided for adversarial examples created by randomly sampling $1,000$ source images $10,000$ times from the datasets provided in the second row. Statistics are provided using adversarial examples that are created from AlexNet and tested on **SqueezeNet**.

			All images	Hard images		Easy (fragile) images		Filtered images	
			$\mathbb{S}$	$\mathbb{S}_{Q<10}$	$\mathbb{S}_{Q<25}$	$\mathbb{S}_{Q>90}$	$\mathbb{S}_{Q>75}$	$\mathbb{S}\setminus(\mathbb{S}_{Q<10}\cup\mathbb{S}_{Q>90})$	$\mathbb{S}\setminus(\mathbb{S}_{Q<25}\cup\mathbb{S}_{Q>75})$
Source images in set:			19,025	1,904	4,758	1,904	4,758	15,219	9,511
Transferability	PGD	Low	41.2%	3.9%	9.1%	92.5%	82.8%	41.9%	39.6%
		Avg	47.8%	5.6%	13.0%	94.0%	86.6%	47.3%	45.8%
		High	54.8%	7.2%	16.9%	96.4%	90.2%	51.2%	50.4%
	CW	Low	61.3%	22.4%	34.3%	95.3%	90.8%	63.5%	65.5%
		Avg	68.4%	26.3%	38.7%	97.0%	93.2%	70.0%	69.0%
		High	74.2%	30.7%	44.4%	98.2%	96.0%	73.5%	73.1%
	MI-FGSM	Low	94.1%	88.4%	90.1%	98.9%	97.5%	95.6%	95.8%
		Avg	96.2%	90.3%	92.2%	99.5%	98.5%	96.4%	96.5%
		High	97.5%	92.6%	93.8%	99.9%	99.3%	97.2%	97.2%
Perturbation (L_2 / L_∞)	PGD	Low	6.72 / 0.05	8.97 / 0.09	8.73 / 0.09	4.48 / 0.03	5.63 / 0.04	7.55 / 0.07	8.03 / 0.07
		Avg	7.34 / 0.06	9.61 / 0.10	9.28 / 0.10	4.73 / 0.03	5.95 / 0.04	7.93 / 0.07	8.38 / 0.07
		High	7.91 / 0.07	10.2 / 0.11	9.84 / 0.11	4.98 / 0.03	6.29 / 0.05	8.30 / 0.08	8.60 / 0.08
	CW	Low	7.61 / 0.11	10.4 / 0.14	10.26 / 0.14	4.35 / 0.09	5.99 / 0.11	8.34 / 0.12	8.97 / 0.12
		Avg	8.37 / 0.13	11.3 / 0.14	10.82 / 0.14	4.65 / 0.09	6.04 / 0.11	8.86 / 0.13	9.23 / 0.13
		High	9.05 / 0.13	11.8 / 0.14	11.35 / 0.14	4.94 / 0.10	6.48 / 0.12	9.35 / 0.13	9.61 / 0.13
	MI-FGSM	Low	15.9 / 0.07	24.7 / 0.08	22.1 / 0.07	7.1 / 0.02	10.3 / 0.03	16.7 / 0.06	16.7 / 0.06
		Avg	16.8 / 0.07	25.3 / 0.09	22.6 / 0.08	8.1 / 0.02	10.8 / 0.03	17.1 / 0.06	17.2 / 0.06
		High	17.5 / 0.08	25.8 / 0.09	23.5 / 0.08	8.6 / 0.02	11.5 / 0.04	17.4 / 0.07	17.5 / 0.07

Table III: The lowest, the highest, and the average transferability, as well as the $L_{\{2,\infty\}}$ perturbations, are provided for adversarial examples created by randomly sampling $1,000$ source images $10,000$ times from the datasets provided in the second row. Statistics are provided using adversarial examples that are created from DenseNet-121 and tested on **VGG-16**.

			All images	Hard images		Easy (fragile) images		Filtered images	
			$\mathbb{S}$	$\mathbb{S}_{Q<10}$	$\mathbb{S}_{Q<25}$	$\mathbb{S}_{Q>90}$	$\mathbb{S}_{Q>75}$	$\mathbb{S}\setminus(\mathbb{S}_{Q<10}\cup\mathbb{S}_{Q>90})$	$\mathbb{S}\setminus(\mathbb{S}_{Q<25}\cup\mathbb{S}_{Q>75})$
Source images in set:			19,025	1,904	4,758	1,904	4,758	15,219	9,511
Transferability	PGD	Low	27.2%	3.2%	7.3%	72.0%	56.4%	27.4%	28.5%
		Avg	33.6%	5.4%	10.2%	75.4%	61.4%	31.9%	31.4%
		High	39.8%	7.5%	14.4%	78.9%	66.3%	36.1%	36.0%
	CW	Low	12.2%	0.1%	1.4%	51.5%	33.5%	9.4%	9.6%
		Avg	16.7%	0.8%	2.8%	55.8%	38.6%	13.8%	12.7%
		High	21.6%	1.5%	4.7%	60.1%	43.8%	18.3%	16.5%
	MI-FGSM	Low	87.4%	77.7%	80.3%	94.5%	91.4%	89.4%	89.6%
		Avg	90.5%	80.0%	84.1%	95.6%	94.3%	90.2%	90.6%
		High	92.4%	82.8%	88.4%	97.2%	96.5%	92.3%	92.2%
Perturbation (L_2 / L_∞)	PGD	Low	6.33 / 0.05	7.95 / 0.09	7.87 / 0.08	4.80 / 0.04	5.61 / 0.05	7.06 / 0.06	7.23 / 0.06
		Avg	6.93 / 0.06	8.56 / 0.09	8.53 / 0.09	5.06 / 0.04	5.98 / 0.05	7.44 / 0.07	7.62 / 0.07
		High	7.41 / 0.08	9.16 / 0.10	8.86 / 0.10	5.30 / 0.04	6.32 / 0.06	7.84 / 0.08	7.98 / 0.07
	CW	Low	2.66 / 0.07	3.93 / 0.08	3.06 / 0.08	2.96 / 0.06	2.55 / 0.07	3.08 / 0.08	3.22 / 0.08
		Avg	3.10 / 0.08	4.75 / 0.10	3.74 / 0.10	2.46 / 0.07	2.77 / 0.07	3.41 / 0.08	3.52 / 0.08
		High	3.50 / 0.09	5.31 / 0.14	4.35 / 0.11	2.61 / 0.08	3.00 / 0.08	3.74 / 0.09	3.82 / 0.09
	MI-FGSM	Low	19.7 / 0.06	25.6 / 0.09	23.4 / 0.08	13.1 / 0.04	15.4 / 0.05	19.9 / 0.06	19.9 / 0.06
		Avg	20.4 / 0.07	26.1 / 0.09	24.7 / 0.08	13.5 / 0.04	16.0 / 0.05	20.3 / 0.06	20.4 / 0.06
		High	21.1 / 0.07	27.0 / 0.09	25.8 / 0.09	14.2 / 0.05	16.9 / 0.06	21.0 / 0.07	21.0 / 0.07

Table IV: The lowest, the highest, and the average transferability, as well as the $L_{\{2,\infty\}}$ perturbations, are provided for adversarial examples created by randomly sampling $1,000$ source images $10,000$ times from the datasets provided in the second row. Statistics are provided using adversarial examples that are created from DenseNet-121 and tested on **ResNet-50**.

			All images	Hard images		Easy (fragile) images		Filtered images	
			$\mathbb{S}$	$\mathbb{S}_{Q<10}$	$\mathbb{S}_{Q<25}$	$\mathbb{S}_{Q>90}$	$\mathbb{S}_{Q>75}$	$\mathbb{S}\setminus(\mathbb{S}_{Q<10}\cup\mathbb{S}_{Q>90})$	$\mathbb{S}\setminus(\mathbb{S}_{Q<25}\cup\mathbb{S}_{Q>75})$
Source images in set:			19,025	1,904	4,758	1,904	4,758	15,219	9,511
Transferability	PGD	Low	23.9%	5.2%	6.9%	65.8%	50.1%	22.3%	21.5%
		Avg	29.4%	7.4%	9.8%	69.2%	55.8%	27.1%	25.9%
		High	35.2%	9.8%	13.1%	72.8%	61.2%	32.6%	30.6%
	CW	Low	10.3%	0.8%	1.6%	43.8%	29.0%	8.7%	8.4%
		Avg	15.0%	1.7%	3.2%	48.6%	33.7%	12.4%	11.5%
		High	19.8%	2.8%	5.2%	52.5%	39.2%	16.1%	15.2%
	MI-FGSM	Low	63.1%	50.1%	53.3%	79.5%	75.7%	64.5%	65.6%
		Avg	68.2%	53.2%	57.8%	81.7%	79.5%	69.7%	69.8%
		High	72.5%	56.3%	62.7%	84.1%	82.1%	74.2%	72.9%
Perturbation (L_2 / L_∞)	PGD	Low	6.41 / 0.06	7.50 / 0.08	7.47 / 0.08	5.28 / 0.04	5.86 / 0.05	6.97 / 0.07	7.09 / 0.07
		Avg	6.97 / 0.07	8.01 / 0.09	8.10 / 0.09	5.54 / 0.05	6.25 / 0.06	7.39 / 0.08	7.52 / 0.08
		High	7.50 / 0.08	8.53 / 0.10	8.65 / 0.10	5.78 / 0.06	6.49 / 0.06	7.09 / 0.08	7.93 / 0.08
	CW	Low	2.77 / 0.07	2.95 / 0.08	2.97 / 0.8	2.42 / 0.05	2.68 / 0.07	3.15 / 0.09	3.22 / 0.09
		Avg	3.21 / 0.08	3.41 / 0.09	3.58 / 0.9	2.59 / 0.06	2.91 / 0.07	3.50 / 0.09	3.58 / 0.09
		High	3.66 / 0.10	3.89 / 0.10	4.36 / 0.11	2.75 / 0.07	3.18 / 0.08	3.83 / 0.10	3.90 / 0.10
	MI-FGSM	Low	20.7 / 0.07	26.7 / 0.09	25.1 / 0.09	14.7 / 0.05	16.9 / 0.06	21.1 / 0.07	21.2 / 0.07
		Avg	22.2 / 0.07	27.7 / 0.09	26.5 / 0.09	15.5 / 0.05	17.9 / 0.06	22.5 / 0.07	22.6 / 0.07
		High	23.6 / 0.08	28.6 / 0.10	27.7 / 0.10	16.4 / 0.05	19.2 / 0.06	23.6 / 0.08	23.5 / 0.08

Table V: The lowest, the highest, and the average transferability, as well as the $L_{\{2,\infty\}}$ perturbations, are provided for adversarial examples created by randomly sampling $1,000$ source images $10,000$ times from the datasets provided in the second row. Statistics are provided using adversarial examples that are created from ResNet-50 and tested on **DenseNet-121**.

			All images	Hard images		Easy (fragile) images		Filtered images	
			$\mathbb{S}$	$\mathbb{S}_{Q<10}$	$\mathbb{S}_{Q<25}$	$\mathbb{S}_{Q>90}$	$\mathbb{S}_{Q>75}$	$\mathbb{S}\setminus(\mathbb{S}_{Q<10}\cup\mathbb{S}_{Q>90})$	$\mathbb{S}\setminus(\mathbb{S}_{Q<25}\cup\mathbb{S}_{Q>75})$
Source images in set:			$19,025$	$1,904$	$4,758$	$1,904$	$4,758$	$15,219$	$9,511$
Transferability	PGD	Low	21.3%	3.2%	4.7%	69.7%	50.8%	19.9%	18.3%
		Avg	27.7%	5.4%	7.8%	73.2%	57.4%	24.8%	22.9%
		High	34.0%	7.5%	10.7%	77.7%	63.0%	29.7%	27.0%
	CW	Low	9.1%	0.3%	0.7%	47.9%	47.9%	7.2%	6.7%
		Avg	13.6%	1.2%	1.9%	52.7%	52.7%	10.5%	8.7%
		High	19.1%	2.3%	3.4%	56.6%	56.6%	14.0%	12.0%
	MI-FGSM	Low	64.2%	48.2%	52.2%	80.9%	76.9%	65.0%	64.1%
		Avg	68.6%	51.7%	56.8%	83.5%	79.3%	68.8%	69.4%
		High	72.3%	54.5%	61.5%	86.2%	83.5%	73.0%	74.6%
Perturbation (L_2 / L_∞)	PGD	Low	6.09 / 0.06	7.08 / 0.07	7.15 / 0.07	4.83 / 0.04	5.60 / 0.05	6.86 / 0.07	7.10 / 0.07
		Avg	6.74 / 0.07	7.88 / 0.08	7.91 / 0.08	5.11 / 0.04	5.96 / 0.05	7.31 / 0.07	7.51 / 0.07
		High	7.35 / 0.08	8.62 / 0.09	8.58 / 0.09	5.37 / 0.05	6.30 / 0.06	7.75 / 0.08	7.93 / 0.08
	CW	Low	2.44 / 0.07	2.02 / 0.06	2.66 / 0.07	2.18 / 0.06	2.18 / 0.06	2.83 / 0.08	2.88 / 0.08
		Avg	2.85 / 0.08	3.03 / 0.08	3.39 / 0.09	2.32 / 0.06	2.32 / 0.06	3.18 / 0.09	3.21 / 0.09
		High	3.27 / 0.09	4.11 / 0.10	3.99 / 0.11	2.46 / 0.07	2.46 / 0.07	3.53 / 0.09	3.58 / 0.09
	MI-FGSM	Low	20.6 / 0.07	27.3 / 0.09	25.7 / 0.09	13.2 / 0.04	15.2 / 0.05	21.2 / 0.07	21.3 / 0.07
		Avg	22.0 / 0.07	28.5 / 0.09	27.4 / 0.09	14.1 / 0.04	17.7 / 0.06	22.1 / 0.08	22.2 / 0.08
		High	23.1 / 0.08	29.6 / 0.10	28.4 / 0.10	15.1 / 0.05	18.5 / 0.06	23.0 / 0.08	23.0 / 0.08

Table VI: The lowest, the highest, and the average transferability, as well as the $L_{\{2,\infty\}}$ perturbations, are provided for adversarial examples created by randomly sampling $1,000$ source images $10,000$ times from the datasets provided in the second row. Statistics are provided using adversarial examples that are created from ViT-L and tested on **ViT-B**.

			All images	Hard images		Easy (fragile) images		Filtered images	
			$\mathbb{S}$	$\mathbb{S}_{Q<10}$	$\mathbb{S}_{Q<25}$	$\mathbb{S}_{Q>90}$	$\mathbb{S}_{Q>75}$	$\mathbb{S}\setminus(\mathbb{S}_{Q<10}\cup\mathbb{S}_{Q>90})$	$\mathbb{S}\setminus(\mathbb{S}_{Q<25}\cup\mathbb{S}_{Q>75})$
Source images in set:			$19,025$	$1,904$	$4,758$	$1,904$	$4,758$	$15,219$	$9,511$
Transferability	PGD	Low	61.7%	48.1%	49.9%	83.2%	76.5%	60.7%	61.0%
		Avg	67.2%	52.6%	54.4%	86.0%	80.8%	66.6%	66.5%
		High	74.0%	57.1%	60.3%	89.0%	84.7%	71.4%	71.1%
	CW	Low	20.6%	9.5%	9.4%	52.8%	40.3%	19.9%	19.5%
		Avg	26.7%	12.3%	13.4%	56.9%	45.5%	24.7%	23.9%
		High	33.4%	15.2%	17.5%	61.4%	50.4%	29.5%	28.3%
	MI-FGSM	Low	80.1%	75.9%	76.9%	89.9%	86.2%	81.9%	82.0%
		Avg	84.6%	78.4%	80.2%	91.2%	89.5%	85.2%	85.3%
		High	89.2%	80.5%	83.5%	93.5%	92.1%	88.4%	87.7%
Perturbation (L_2 / L_∞)	PGD	Low	6.49 / 0.06	7.40 / 0.07	7.38 / 0.07	4.94 / 0.04	5.57 / 0.05	6.81 / 0.06	6.87 / 0.06
		Avg	6.93 / 0.07	7.71 / 0.07	7.70 / 0.08	5.21 / 0.04	5.98 / 0.05	7.14 / 0.07	7.20 / 0.07
		High	7.35 / 0.07	8.03 / 0.08	8.04 / 0.08	5.54 / 0.05	6.34 / 0.06	7.47 / 0.07	7.54 / 0.07
	CW	Low	2.39 / 0.07	2.64 / 0.08	2.58 / 0.07	1.98 / 0.06	2.20 / 0.06	2.54 / 0.08	2.63 / 0.08
		Avg	2.64 / 0.08	2.87 / 0.08	2.88 / 0.08	2.11 / 0.06	2.37 / 0.07	2.77 / 0.09	2.82 / 0.08
		High	2.91 / 0.09	3.12 / 0.09	3.15 / 0.09	2.31 / 0.07	2.55 / 0.08	2.99 / 0.09	3.05 / 0.09
	MI-FGSM	Low	15.0 / 0.05	18.8 / 0.06	17.0 / 0.06	11.0 / 0.04	12.9 / 0.04	15.2 / 0.05	15.0 / 0.05
		Avg	16.9 / 0.05	19.7 / 0.07	18.2 / 0.06	11.7 / 0.04	13.8 / 0.05	16.4 / 0.06	16.2 / 0.06
		High	17.5 / 0.06	19.5 / 0.07	19.5 / 0.06	12.3 / 0.04	14.5 / 0.05	17.6 / 0.06	17.5 / 0.06

Table VII: The lowest, the highest, and the average transferability, as well as the $L_{\{2,\infty\}}$ perturbations, are provided for adversarial examples created by randomly sampling $1,000$ source images $10,000$ times from the datasets provided in the second row. Statistics are provided using adversarial examples that are created from ViT-B and tested on **ViT-L**.

			All images	Hard images		Easy (fragile) images		Filtered images	
			$\mathbb{S}$	$\mathbb{S}_{Q<10}$	$\mathbb{S}_{Q<25}$	$\mathbb{S}_{Q>90}$	$\mathbb{S}_{Q>75}$	$\mathbb{S}\setminus(\mathbb{S}_{Q<10}\cup\mathbb{S}_{Q>90})$	$\mathbb{S}\setminus(\mathbb{S}_{Q<25}\cup\mathbb{S}_{Q>75})$
Source images in set:			$19,025$	$1,904$	$4,758$	$1,904$	$4,758$	$15,219$	$9,511$
Transferability	PGD	Low	38.7%	23.2%	27.7%	69.2%	57.9%	38.8%	37.8%
		Avg	44.7%	27.5%	32.2%	72.8%	63.0%	43.5%	42.0%
		High	51.2%	30.8%	37.2%	77.4%	69.7%	47.3%	45.4%
	CW	Low	9.4%	2.0%	2.9%	40.1%	25.8%	10.1%	8.7%
		Avg	14.6%	3.8%	5.3%	44.2%	30.8%	13.5%	11.0%
		High	19.2%	5.4%	8.0%	49.7%	35.7%	17.7%	14.2%
	MI-FGSM	Low	59.1%	48.1%	52.9%	72.8%	67.1%	59.0%	58.9%
		Avg	63.6%	50.3%	56.2%	75.7%	75.5%	63.5%	63.1%
		High	68.2%	53.7%	59.5%	78.1%	76.6%	68.1%	67.4%
Perturbation (L_2 / L_∞)	PGD	Low	6.00 / 0.05	6.79 / 0.07	6.67 / 0.06	4.68 / 0.03	5.31 / 0.04	6.27 / 0.06	6.41 / 0.06
		Avg	6.49 / 0.06	7.14 / 0.07	7.10 / 0.07	4.98 / 0.04	5.67 / 0.05	6.76 / 0.06	6.88 / 0.06
		High	7.01 / 0.07	7.54 / 0.08	7.49 / 0.08	5.26 / 0.04	6.01 / 0.05	6.98 / 0.07	7.14 / 0.07
	CW	Low	1.88 / 0.06	2.09 / 0.08	2.13 / 0.07	1.72 / 0.05	1.85 / 0.06	2.08 / 0.06	2.02 / 0.06
		Avg	2.25 / 0.08	2.56 / 0.09	2.53 / 0.08	1.85 / 0.05	2.05 / 0.06	2.42 / 0.07	2.40 / 0.07
		High	2.71 / 0.09	2.91 / 0.10	2.84 / 0.09	1.94 / 0.06	2.87 / 0.07	2.74 / 0.08	2.63 / 0.07
	MI-FGSM	Low	15.5 / 0.05	19.2 / 0.07	18.8 / 0.06	11.2 / 0.04	13.3 / 0.04	16.8 / 0.05	16.4 / 0.05
		Avg	17.6 / 0.06	21.7 / 0.07	19.2 / 0.06	12.6 / 0.04	14.6 / 0.05	17.5 / 0.05	17.1 / 0.05
		High	18.2 / 0.06	22.5 / 0.08	20.5 / 0.07	13.9 / 0.04	15.7 / 0.05	18.1 / 0.06	18.0 / 0.06