Supplementary Material for: Geometric robustness of deep networks: analysis and improvement

1 Sampling Random Transformations

This section provides a detailed examination on how the sampling of random transformations is performed in Section 5 of the main paper. Let $\boldsymbol{x}$ be the image we want to transform, and $v \in T_{\boldsymbol{x}}\,\mathcal{M}(\boldsymbol{x})$ be a vector of the space tangent to the manifold $\mathcal{M}(\boldsymbol{x})$ at point $\boldsymbol{x}$. In Section 4.2, we have used a retraction that first maps a vector to $T_e\,\mathcal{T}$ and then use matrix exponential to map it onto the set $\mathcal{T}$, which is done as

$$\tau = \exp\left(\sum_i v_i G_i\right), \tag{1}$$

where G_i is the basis of $T_e\,\mathcal{T}$ and τ is the matrix representation of the transformation. Although matrix exponential is defined on the whole $T_e\,\mathcal{T}$, it is not injective. Thus, the range of the matrix logarithm, which is used for mapping an element of $\mathcal{T}$ to $T_e\,\mathcal{T}$, is limited to only a subset of $T_e\,\mathcal{T}$. We say that a vector $v \in T_{\boldsymbol{x}}\,\mathcal{M}(\boldsymbol{x})$ is *valid* if its counterpart on $T_e\,\mathcal{T}$ is in the range of the matrix logarithm, i.e., if

$$\log\left(\exp\left(\sum_i v_i G_i\right)\right) = \sum_i v_i G_i. \tag{2}$$

Now, using this definition of validity, we start explaining the sampling of random transformations which are used for calculating the invariance of a classifier against geometric transformations in a random regime. Again, let $\boldsymbol{x}$ be the image we want to transform, and $v \in T_{\boldsymbol{x}}\,\mathcal{M}(\boldsymbol{x})$ with $\|v\| = 1$ be a vector sampled from the unit sphere of the tangential space. Then, for $\alpha > 0$, the path from $\boldsymbol{x}$ to $R_{\boldsymbol{x}}(\alpha v)$ is defined as

$$\gamma(t) = R_{\boldsymbol{x}}(t\alpha v), \quad t \in [0, 1]. \tag{3}$$

Because we assume that the geodesic path is direct, as long as αv is valid, γ is the geodesic path and thus its length is the geodesic distance, i.e., $d(\boldsymbol{x}, R_{\boldsymbol{x}}(\alpha v)) = L(\gamma)$. As α increases, there are two possibilities for $d(\boldsymbol{x}, R_{\boldsymbol{x}}(\alpha v))$:

1. $d(\boldsymbol{x}, R_{\boldsymbol{x}}(\alpha v))$ increases with α as the path gets longer until it reaches a bound where the images gets too distorted; further increases do not change the image. For example, the image can be scaled up so much so that it becomes a single color image and thus further scaling does not change anything.

2. $d(\boldsymbol{x}, R_{\boldsymbol{x}}(\alpha v))$ increases with α until αv becomes invalid. For example, for the set of rotations, α becomes invalid after $\alpha = \pi$.

In both cases, $d(\boldsymbol{x}, R_{\boldsymbol{x}}(\alpha v))$ is a non-decreasing function which reaches its upper bound for a certain α. For the vector v, let this bound on the geodesic distance be called $r^{(v)}_{\max}$. Then, since $d(\boldsymbol{x}, R_{\boldsymbol{x}}(0)) = d(\boldsymbol{x}, \boldsymbol{x}) = 0$, for a given geodesic distance $r \leq r^{(v)}_{\max}$, there exists a value of α such that $d(\boldsymbol{x}, R_{\boldsymbol{x}}(\alpha v)) = r$. Using this, we can sample the set of transformed images with given geodesic score r, $\mathcal{M}_r(\boldsymbol{x}) = \{\boldsymbol{x}_\tau : \tau \in \mathcal{T}, d_{\boldsymbol{x}}(e, \tau) = r\}$, by using rejection sampling with following steps:

1. Sample v from $U_{\boldsymbol{x}} = \{v \in T_{\boldsymbol{x}}\,\mathcal{M} : \|v\| = 1\}$

2. Check if $r \leq r^{(v)}_{\max}$. If not, return to step 1.

3. Return $R_{\boldsymbol{x}}(\alpha v)$ that corresponds to $d(\boldsymbol{x}, R_{\boldsymbol{x}}(\alpha v)) = r$.

Practically, this sampling is done by increasing the magnitude of a vector that is sampled from the unit sphere of $T_{\boldsymbol{x}}\,\mathcal{M}(\boldsymbol{x})$. Let r be the requested geodesic score of the output transform, η a chosen step size and ϵ a tolerance term. First, a vector v is sampled randomly from the unit sphere. The increase of magnitude is done iteratively. For the iteration i, the distance between the original image $\boldsymbol{x}$ and the transformed version $R_{\boldsymbol{x}}(i\eta v)$ is calculated using the distance calculated in the previous iteration as

$$d(\boldsymbol{x}, R_{\boldsymbol{x}}(i\eta v)) = d(\boldsymbol{x}, R_{\boldsymbol{x}}((i-1)\eta v)) + \|R_{\boldsymbol{x}}(i\eta v) - R_{\boldsymbol{x}}((i-1)\eta v)\| \tag{4}$$

starting with $d(\boldsymbol{x}, R_{\boldsymbol{x}}(0)) = 0$. If $d(\boldsymbol{x}, R_{\boldsymbol{x}}(i\eta v)) = d(\boldsymbol{x}, R_{\boldsymbol{x}}((i-1)\eta v))$, i.e., if the distance has reached its upper bound, then the algorithm returns to the beginning, starting by sampling a new $v \in U_{\boldsymbol{x}}$. If $d(\boldsymbol{x}, R_{\boldsymbol{x}}(i\eta v)) > r$, the iterations stop. Then, an α such that $i\eta \geq \alpha \geq (i-1)\eta$ and $d(\boldsymbol{x}, R_{\boldsymbol{x}}(\alpha v)) = r \pm \epsilon$ is found by doing binary search between $i\eta$ and $(i-1)\eta$. Finally, the validity of the vector αv is checked. If it is not valid, then the algorithm again returns to the beginning and samples a new $v \in U_{\boldsymbol{x}}$. However, if it is valid, then our sampled transformed image is found as $R_{\boldsymbol{x}}(\alpha v))$. In the main paper, this sampling is done multiple times for many different images to get a set of transformed images with same(within the tolerance value) geodesic score. This set is then used for estimating the misclassification rate of a classifier at a given geodesic score r. This is repeated for different r to estimate the robustness of the classifier to random transformations in Section 5 of the main text.

2 Supplementary Examples of Transformed Images

In Figures 1, 2, and 3, we are giving more examples of ILSVRC2012 images transformed using outputs of ManiFool using ResNet18. In all cases, the odd rows are the original images and the even rows are the 'fooled' images. The labels of each image is found at the bottom of the image.

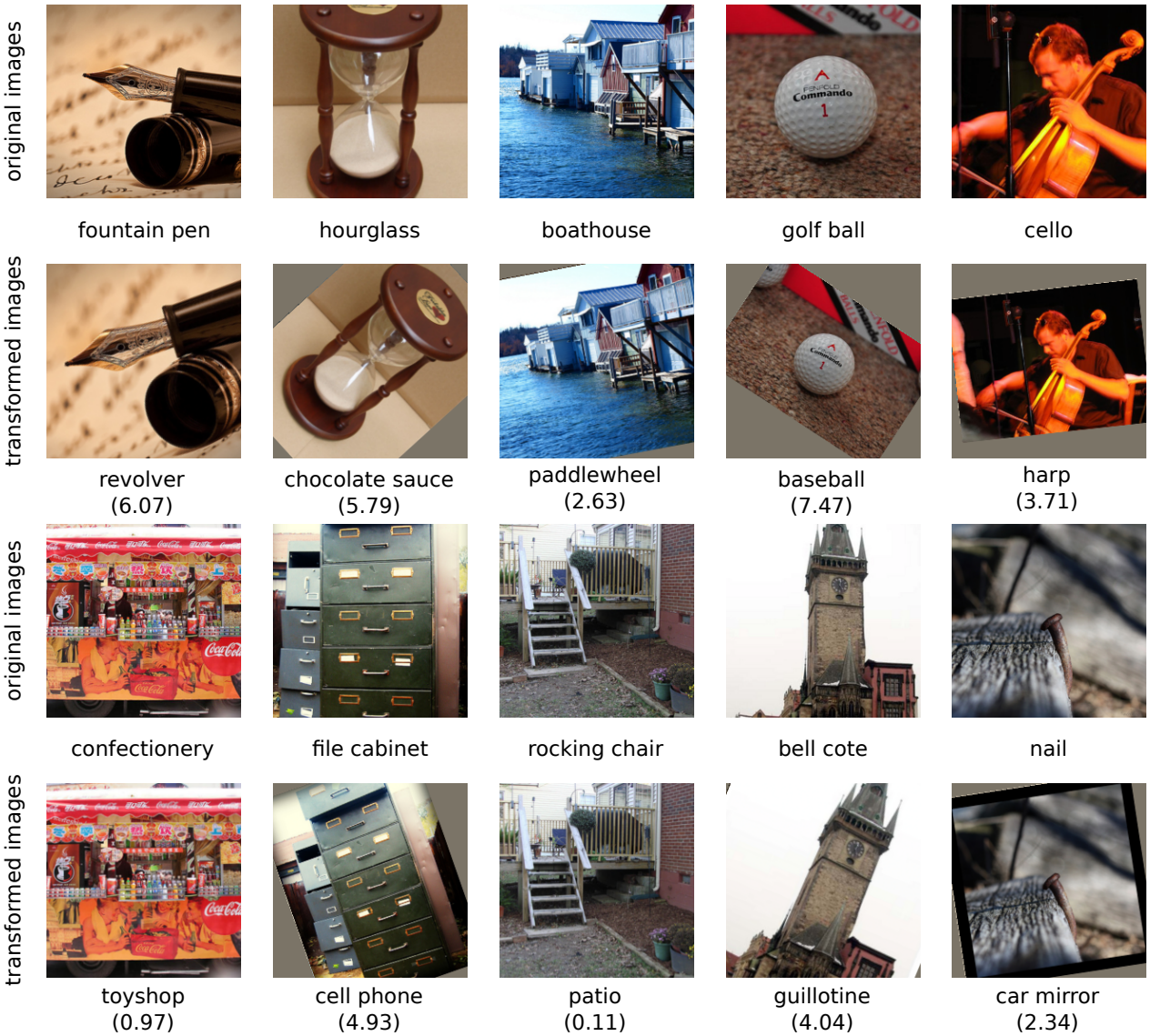

Figure 1: Examples that fool ResNet18 using similarity transforms that are generated using ManiFool. The geodesic score of these transformations are given in parentheses below the label of the transformed image. These scores are computed using the direct distance method that was explained in Section 4.3 of the main paper with $\eta = 0.05$.

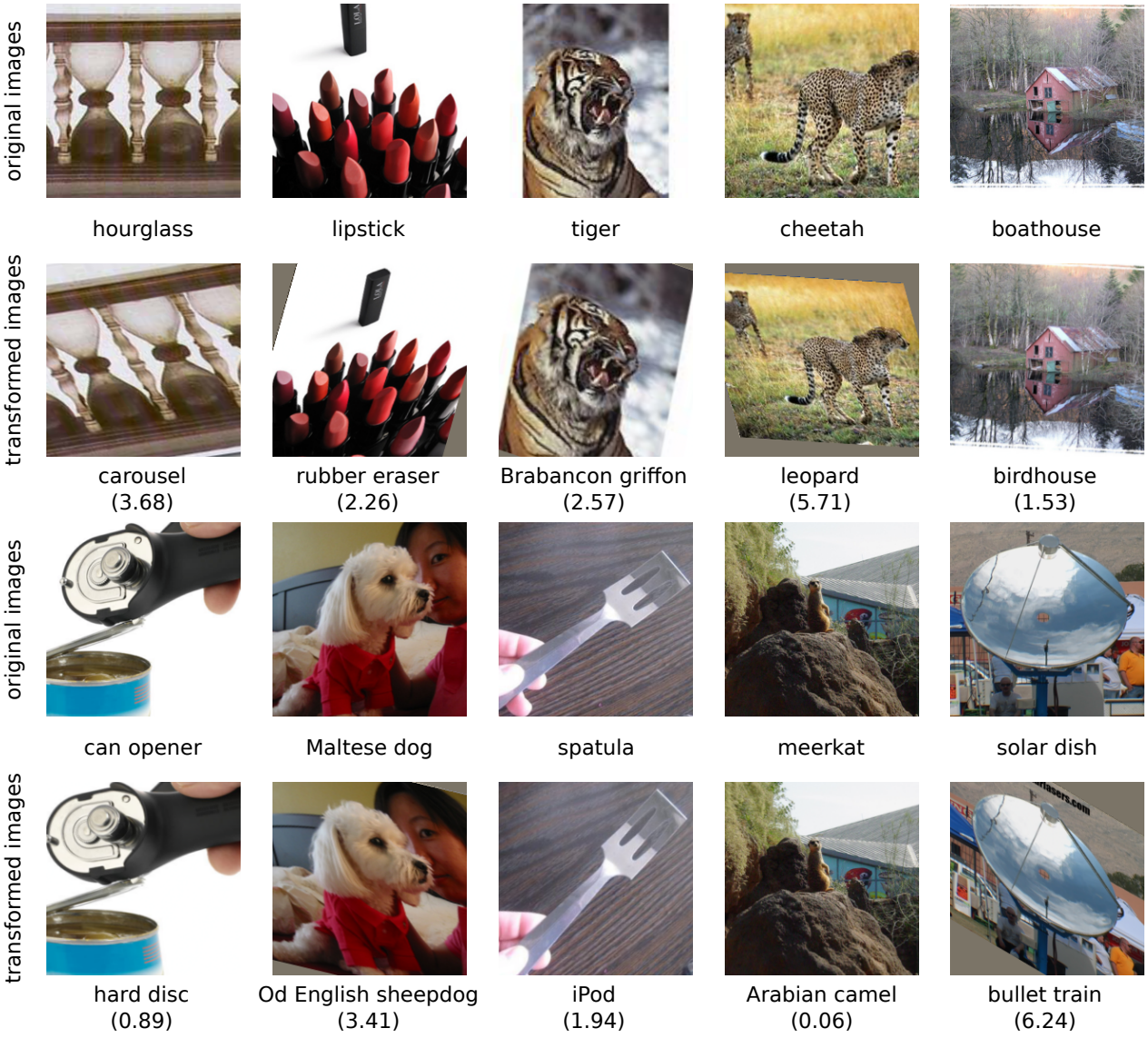

Figure 2: Examples that fool ResNet18 using affine transforms that are generated using ManiFool. The geodesic score of these transformations are given in parentheses below the label of the transformed image. These scores are computed using the direct distance method that was explained in Section 4.3 of the main paper with $\eta = 0.05$.

Figure 3: Examples that fool ResNet18 using projective transforms that are generated using ManiFool. The geodesic score of these transformations are given in parentheses below the label of the transformed image. These scores are computed using the direct distance method that was explained in Section 4.3 of the main paper with $\eta = 0.005$.