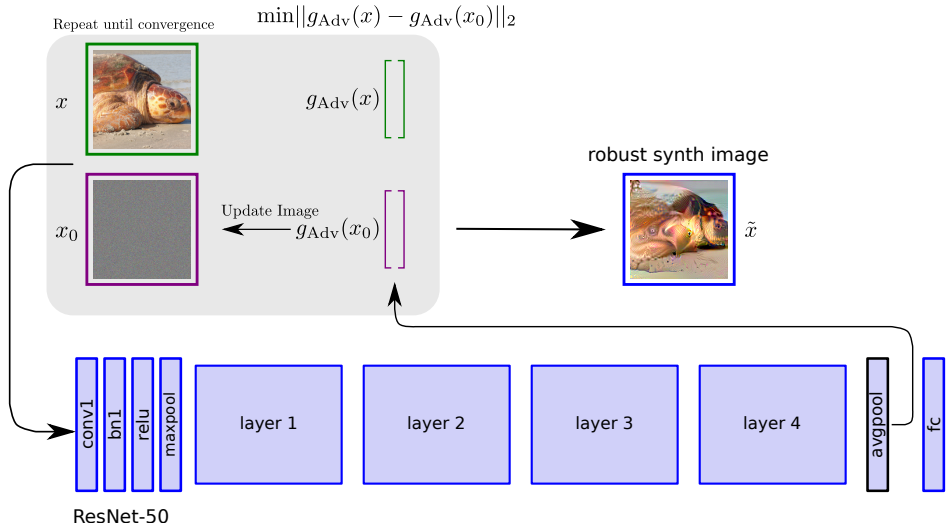

Figure 9: The Robust Image Synthesis pipeline: A noise image x_0 is passed through an adversarially trained ResNet-50 and the penultimate layer features $g_{\text{Adv}}(x_0)$ are matched wrt the original images' penultimate feature activation $g_{\text{Adv}}(x)$ via an L2 loss, and is repeated until convergence (Santurkar et al., 2019; Engstrom et al., 2019). Critically we use $g_{\text{Adv}}(\circ)$ as a summary statistic of peripheral processing in our experiments.

A Image Synthesis Details

	Classes								
RIN	Dog	Cat	Frog	Turtle	Bird	Primate	Fish	Crab	Insect
IN	151-268	281-285	30-32	33-37	68-100	365-382	389-397	118-121	300-319

Table 1: Classes of RestrictedImageNet (**RIN**) and the corresponding ImageNet (**IN**) class ranges.

A.1 Standard and Robust Stimuli

We used the publicly available code from Santurkar et al. (2019); Engstrom et al. (2019); Ilyas et al. (2019) found here to synthesize both standard and robust stimuli which where derived from a regularly and adversarially trained model respectively: `https://github.com/MadryLab/robust_representations`

A schematic that illustrates the robust stimuli rendering pipeline can be seen in Figure 9. Standard stimuli is generated with the same procedure, and number of iterations, but the network $g_{\text{Adv}}(\circ)$ is replaced with $g_{\text{Standard}}(\circ)$ instead.

A visualization of the convergence of the loss when performing the synthesis procedure can be seen in Figure 10.

A.2 Texform Stimuli

Texform stimuli were synthesized using the publicly available code of Deza et al. (2019a): `https://github.com/ArturoDeza/Fast-Texforms`

The following images (class:[image id's]) were removed as they did not converge:

- texform0: 0:[49],1:[9],2:[],3:[44],4:[],5:[],6:[10],7:[40],8:[].
- texform1: 0:[49],1:[9,44],2:[],3:[44],4:[],5:[],6:[10],7:[40],8:[]

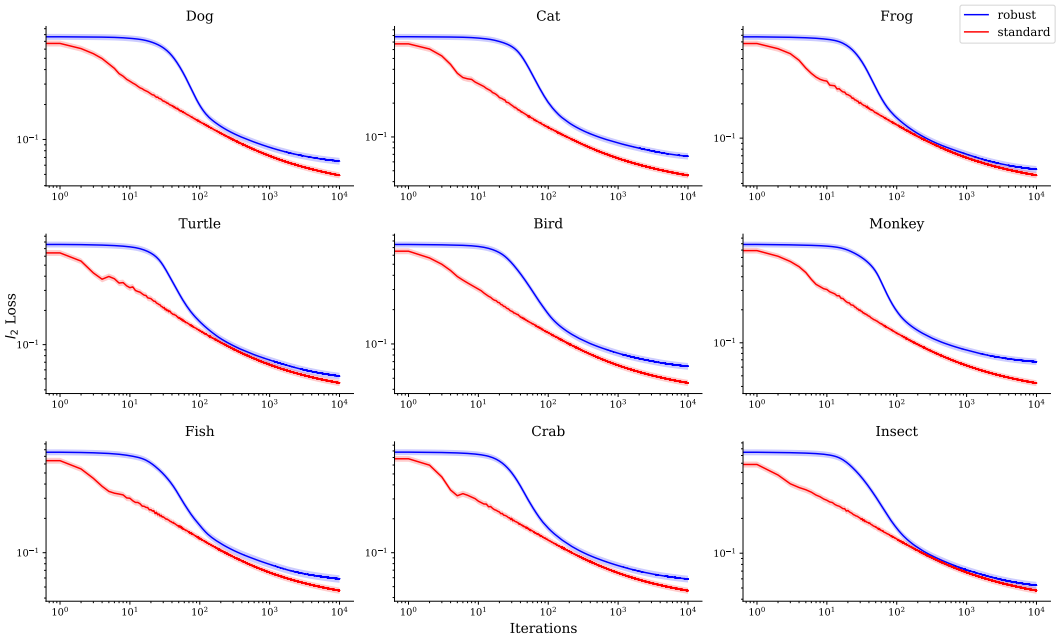

Figure 10: Per-Class Synthesis Loss visualizations for the Robust and Standard Stimuli across all samples. Errorbar represents 1 standard error.

In addition the following image id's were removed from our psychophysical analysis from the texform stimuli as they converged to the *exact* same image even when starting from different noise seeds. This was found while doing a post-hoc IQA analysis as the one shown in Figure 11. These stimuli only occurred for classes 0 (dog) and 1 (cat):

- texform: 0:[22,25,26,27,29,93,94,95,96,97,98,99],1:[20,21,22,23,73,74]

We found that Standard and Robust stimuli did not have this identical convergence problem over the 900 rendered pairs (1800 stimuli in total for Standard and 1800 in total for Robust).

Note 1a: A common mis-conception is that Freeman & Simoncelli (2011)-derived stimuli (such as texforms) *do not* contain structural priors and only performs localized texture synthesis over smoothly overlapping log-polar receptive fields. This has been investigated with great detail in Wallis et al. (2016; 2017); Liu et al. (2016) that showed that without spectral constraints it is impossible to generate metameric images from non-stationary textures for the human observer when showing such stimuli in the visual periphery. For texforms the metameric constraint is purposely broken because we'd like to test how a specific biologically-plausible family of transformations (embodied through the synthesis procedure) interacts with eccentricity when the eccentricity-dependent and scaling factors texform parameters are fixed. See (z_*, s_*) from Eq. 7.

Note 1b: The Freeman & Simoncelli (2011) synthesis model is not equivalent to the Portilla & Simoncelli (2000) synthesis model. The Freeman & Simoncelli (2011) is a super-ordinate synthesis model class that locally uses the Portilla & Simoncelli (2000) synthesis model over smoothly overlapping receptive fields in addition to adding a global structural prior. Texforms are rendered with the Freeman & Simoncelli (2011) model, by placing he simulated point of fixation *outside* the image (Long et al., 2018; Deza et al., 2019a).

Note 1c: Usual texform rendering time is about 1 day per image, though the rendering procedure has been accelerated to the order of minutes as shown in Deza et al. (2019a). We used their publicly available code in our experiments. Thus, it is worth noting that synthesizing texforms in the order of hundreds of thousands (or millions) for supervised learning experiments – has not been done before and is computationally expensive (may take months), which is why Figure 2 displays no information on texform-trained CNN's. This direction is current work.

Note 2: A first naive criticism to the selection of making texforms fixed and not varying as a function of eccentricity – given the model they were based on (Freeman & Simoncelli, 2011) – is that they will not create metameric stimuli. Our anticipated reply to this is three-fold, and partially aligned with the motivation of Long et al. (2018):

1. Our goal is *not* to make metameric stimuli out of texforms or robust stimuli, but to examine how perceptual discriminability rates of a *fixed stimuli* change as a function of retinal eccentricity. By checking if these perceptual decays are similar (which we show) we can connect both functions that give rise to these apparently un-related transformations (the stimuli). Recall Eq. 6.
2. Having a "metameric texform" that changes as a function of eccentricity would defeat the purpose of using it as a control in our experiments. Had this been the road taken, we would now have a control curve that will presumably be horizontal and at chance, providing no information about how the transformation that gives rise to the robust stimuli is linked to the texform transformation.
3. The goal of this paper is *not* to make a foveated metamer model that fools human observers similar to that of Freeman & Simoncelli (2011); Rosenholtz et al. (2012); Deza et al. (2019b); Wallis et al. (2019) that would be based on a foveated adversarially trained network. The previous idea however is highly interesting and is being explored in current work, and this work provides a proof of concept that it is tractable.

A.3 Synthesis vs Synthesis and Original vs Original

The goal of combining these experimental variations into a block (called *'stimulus roving'*) in our experiments was two-fold: 1) to add difficulty to the tasks thus reducing the likelihood of ceiling effects; 2) to gather two psychometric functions per family of stimuli, which portrays a better description of each stimulus's evoked perceptual signatures. Synthesis vs Synthesis experiments probe the diversity of samples in pixel space that can potentially yield visual metamerism, while the Original vs Synthesis condition yields a stronger condition for visual metamerism. Several works have explored these paradigms (Wallis et al., 2016; Deza et al., 2019b).

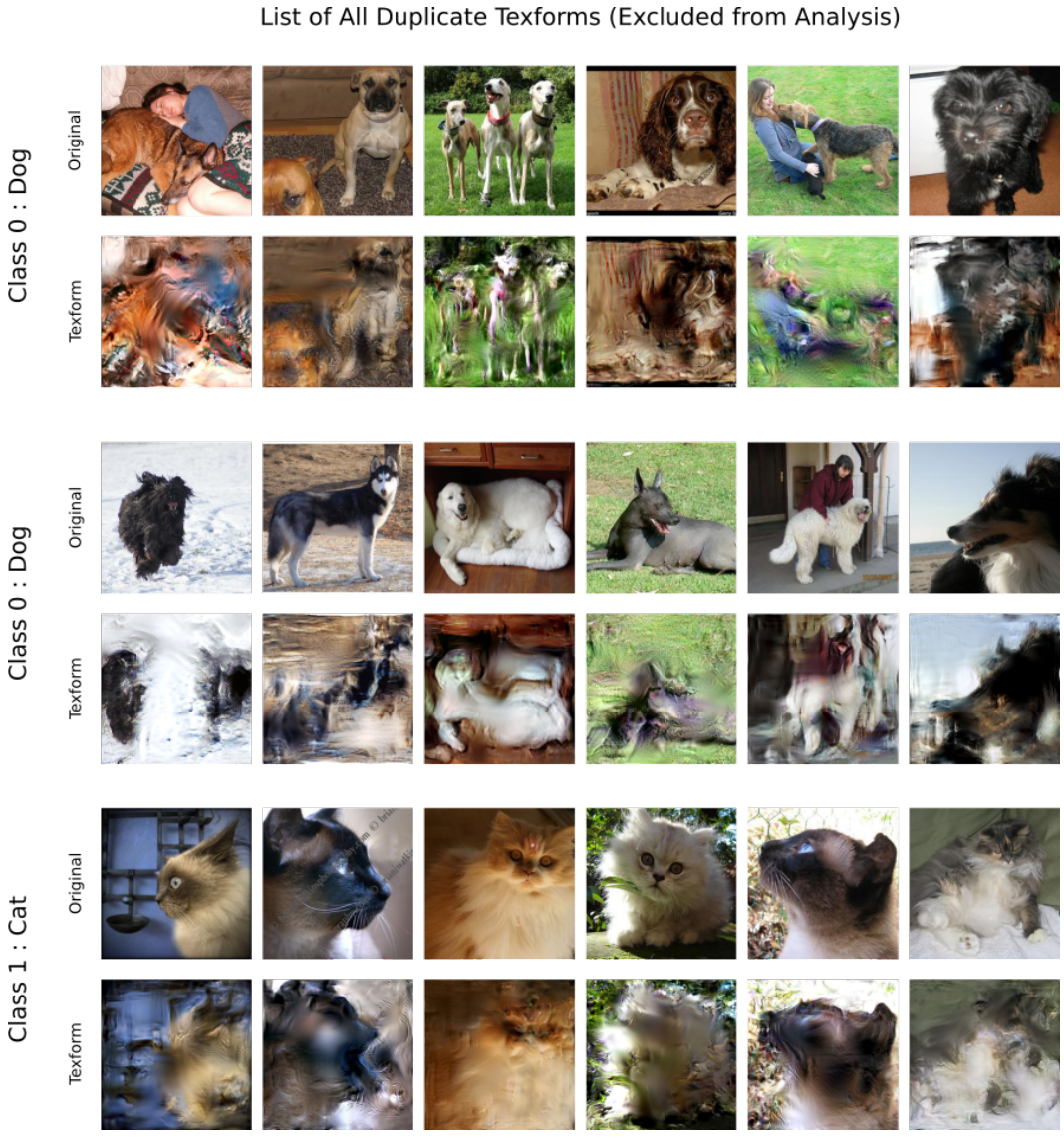

Figure 11: Duplicates are images that even though they were initialized with two different random noise images, they converged to the exact same image (Mean Square Error between synthesized samples is equal to zero). These stimuli were *excluded* from our analysis and represent only 2% $(18/894)$ of the used texform stimuli.

A.4 SAMPLE STIMULI

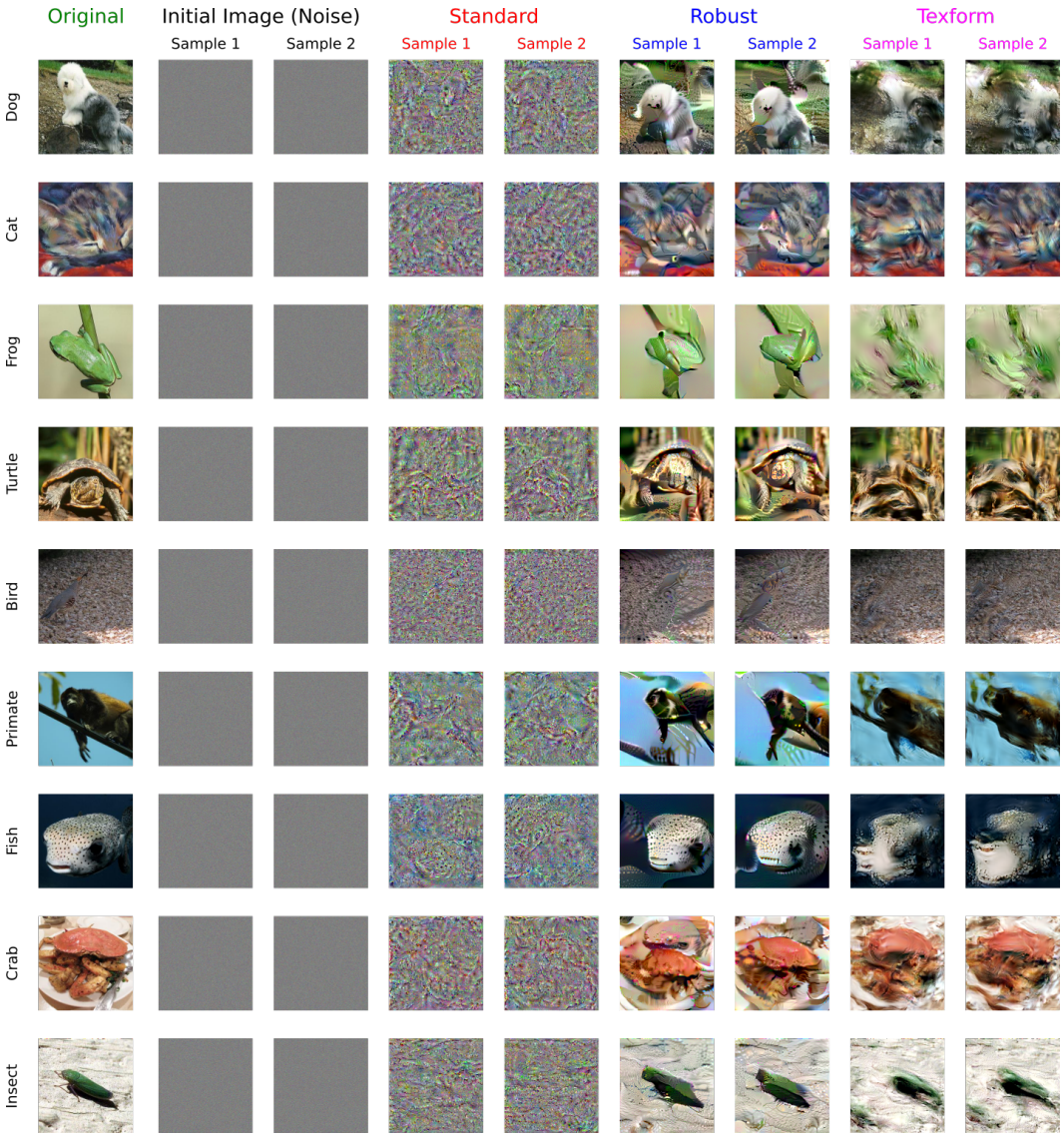

Figure 12: A collection of sample stimuli for each image class used in our experiments.

A.5 SYNTHESIS VARIATIONS

In this sub-section we show a collection of different synthesized samples using different reference images, and also using different starting images. If the transformations undergoing the texform model and the adversarially robust model are similar, then the resulting synthesis outputs should look similar if the output of one model is used as the input to another (See inset 2). A similar effect should occur if the starting image for the texform model is robust stimuli and vice-versa (See inset 3). We see this effects qualitatively holds even more so for the Turtle than the Cat image. Overall there are striking low-frequency structural similarities across all images in the last 2 columns. However, further psychophysical experiments are needed to test the rates of discriminability of such images as a function of retinal eccentricity to establish a more precise relationship between them.

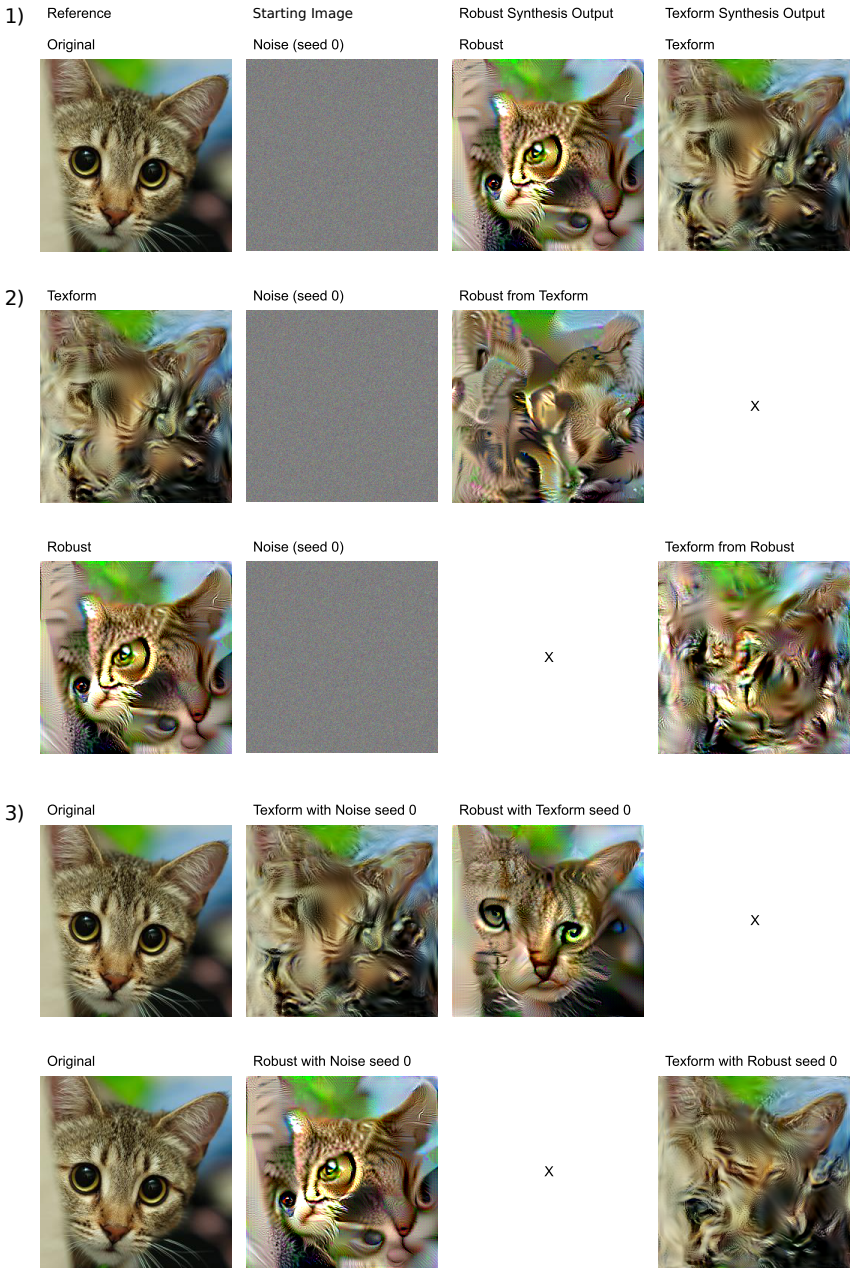

Figure 13: Robust and Texform synthesis variations of a cat with different reference images and starting images.

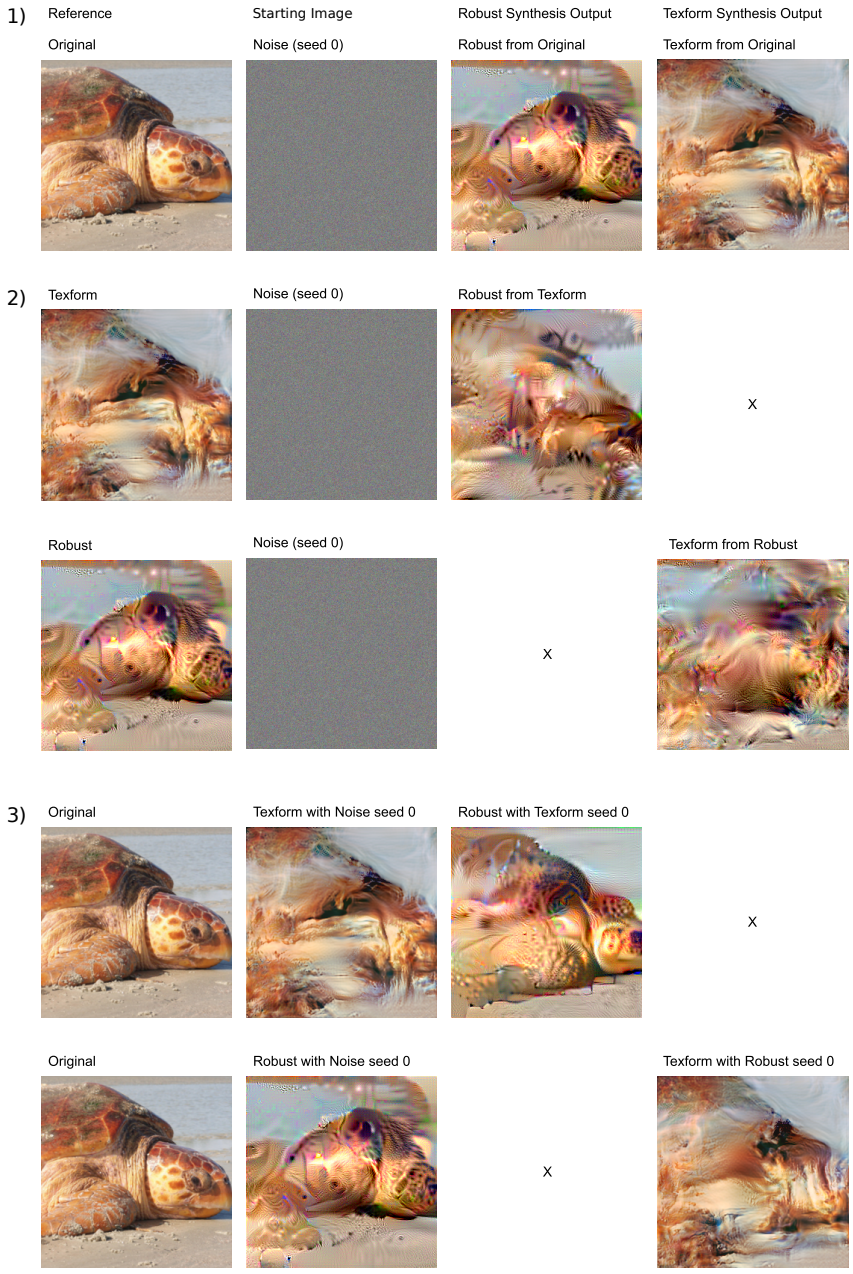

Figure 14: Robust and Texform synthesis variations of a turtle with different reference images and starting images.

A.6 ETHICS STATEMENT, ADDITIONAL METHODS & SINGLE OBSERVER RESULTS

All 12 human subjects involved in this research willfully participated in this experiment via explicit consent during each session of this experiment. Our experimental design was reviewed and approved by an Institutional Review Board (IRB).

Subjects: We used a total of 12 human subjects that consisted of undergraduates from (To Be Disclosed) University. Subjects were paid a fee of \$20 per hour to complete the experiment in a total of 6 hours over anywhere between 2 to 6 days where observers performed a maximum of 2 hours of psychophysics per day. Our experiments had an approved IRB protocol from (To Be Disclosed) University. Human participants were all tested with a Snellen eye-chart and had at least 20/20 visual acuity and had either no visual correction or contact lenses to correct their vision. All participants were naive to the experiment (*i.e.* no participants were the experimenters), in all cases, subjects were not familiar with the concepts of either visual metamerism or adversarial images. No participants with eye-glasses were used in our experiments.

Apparatus: Experiments were ran on an Ubuntu-Linux Machine version 14.04.5 LTS with MATLAB 2015a's Psychtoolbox version 3.0.14. A chin rest was used so that observers can view stimuli on a screen placed at 50 cm distance from their eyes. We used a 34 inch diagonal 75 Hz LCD monitor, that measured 80cm width and 34 cm height with a visual display resolution of 3440 pixels width by 1440 pixels height. From here the total degrees of visual angle was computed via:

$$\theta = 2 \times \text{atan}(17.0/50.0) \times 180.0/\pi \tag{8}$$

And the degrees of visual angle subtended by the stimuli is computed by multiplicating the proportion of pixels subtended by the stimuli with respect to the monitor:

$$\theta_{\text{Stimuli}} = \theta \times 256/1440 = 6.67 \tag{9}$$

In the rest of this sub-section we plot the single observer results where the trends observed in Figure 5 still hold true at the individual per-observer level. Each participant saw 72 trials of the oddity task for every stimuli condition and eccentricity (i.e. robust synthesized vs synthesized at 5 degrees, robust synthesized vs original at 5 degrees, etc ...). On the 2AFC matching, they saw 80. Errorbars in each plot were computed via a 10,000 sample bootstrapping and represent the 95% confidence interval.

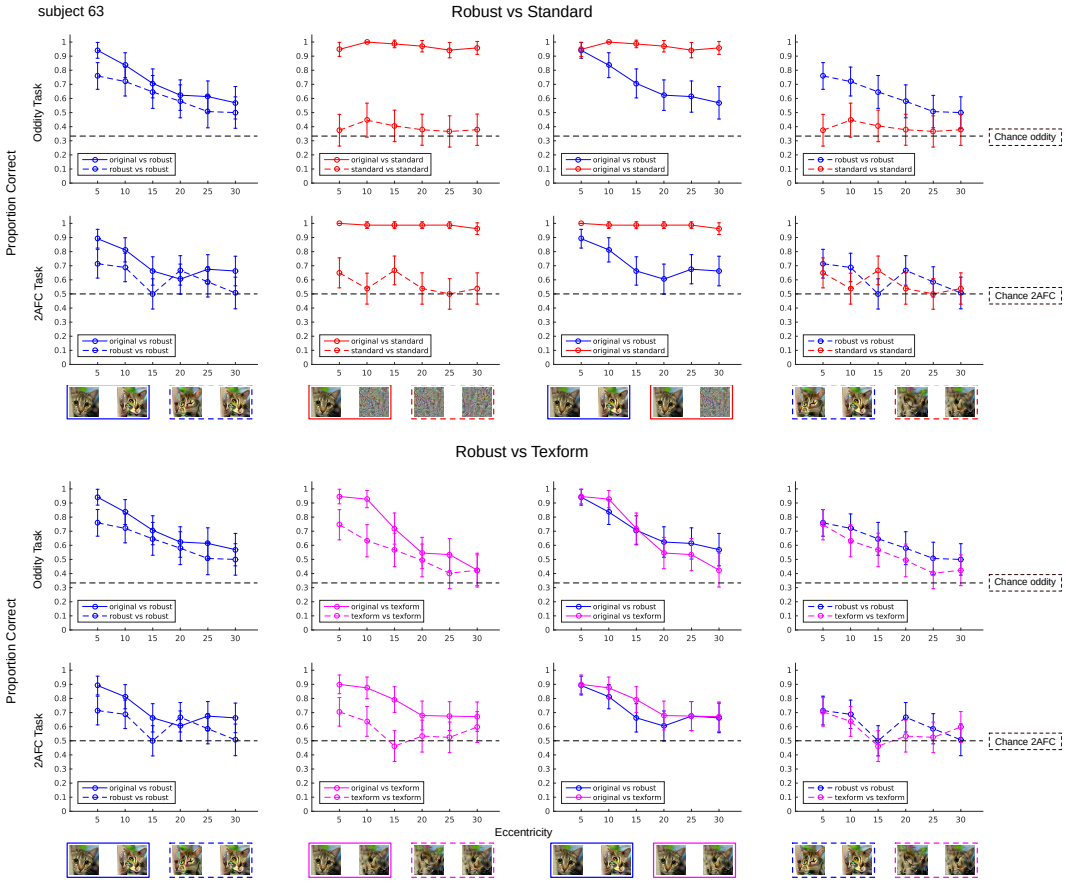

Figure 15: Subject 63

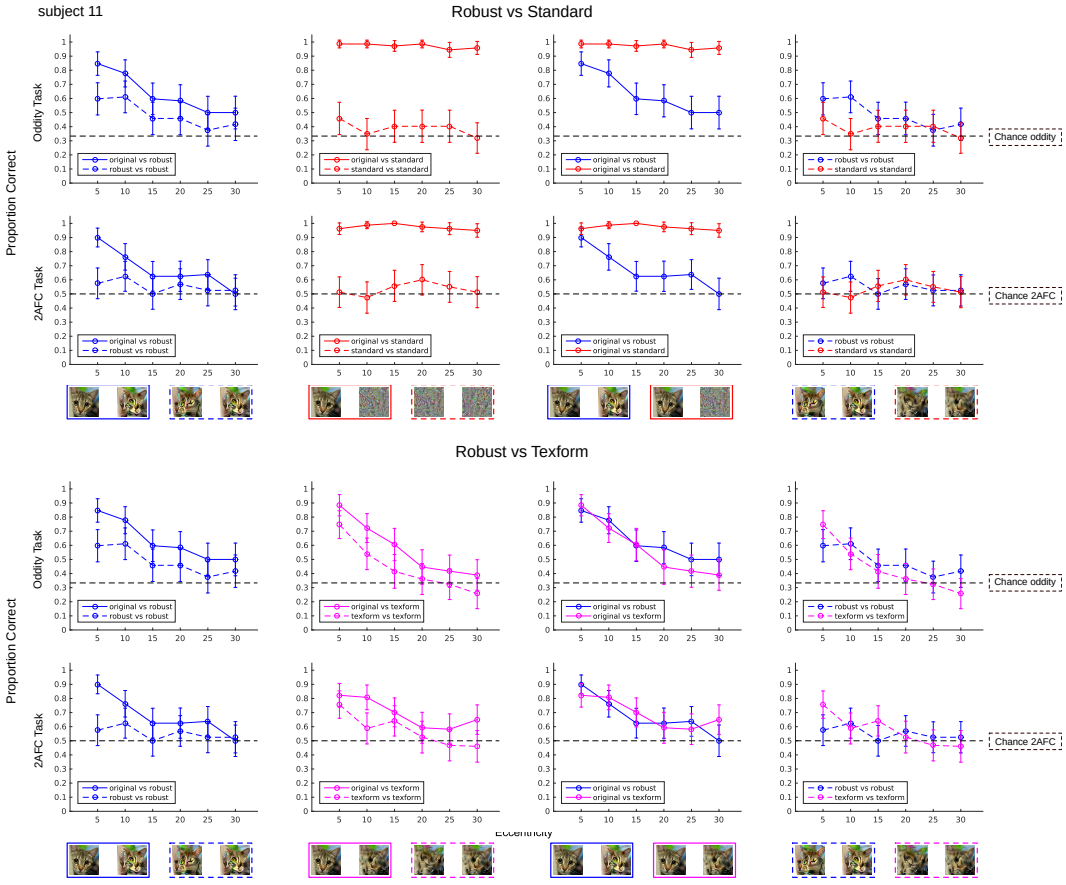

Figure 16: Subject 11

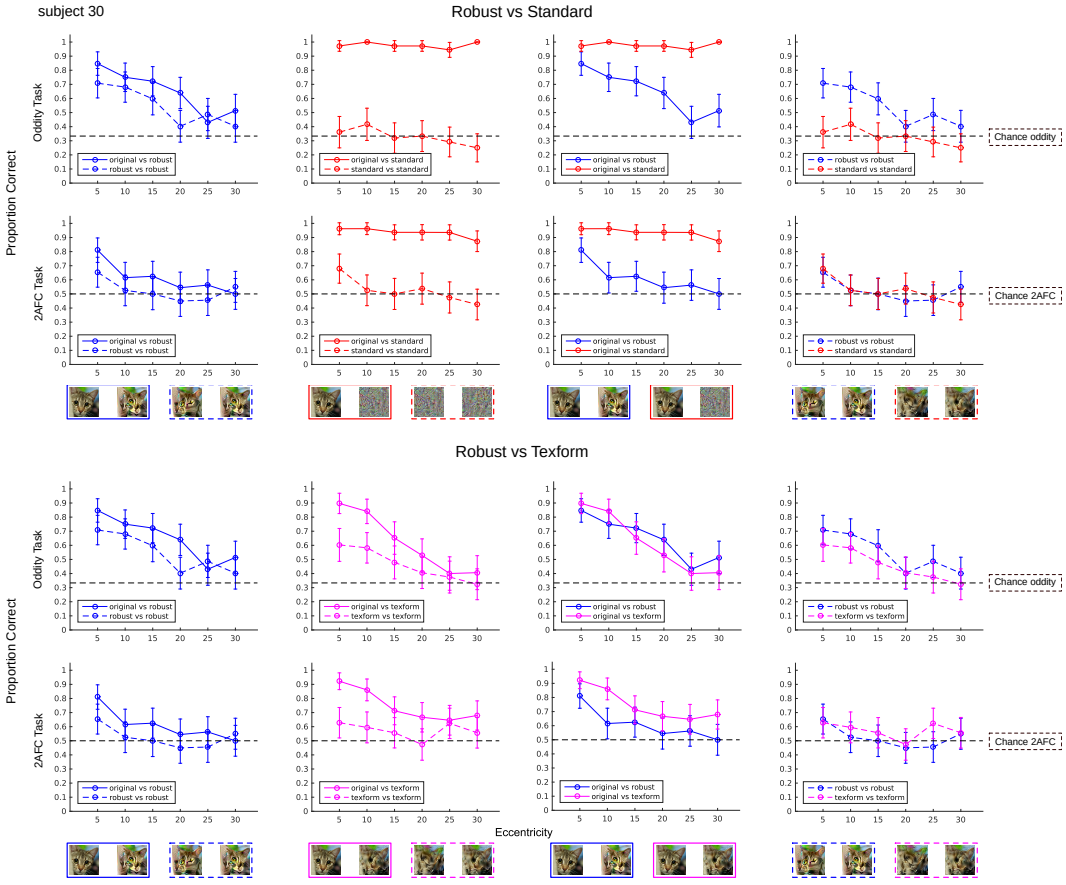

Figure 17: Subject 30

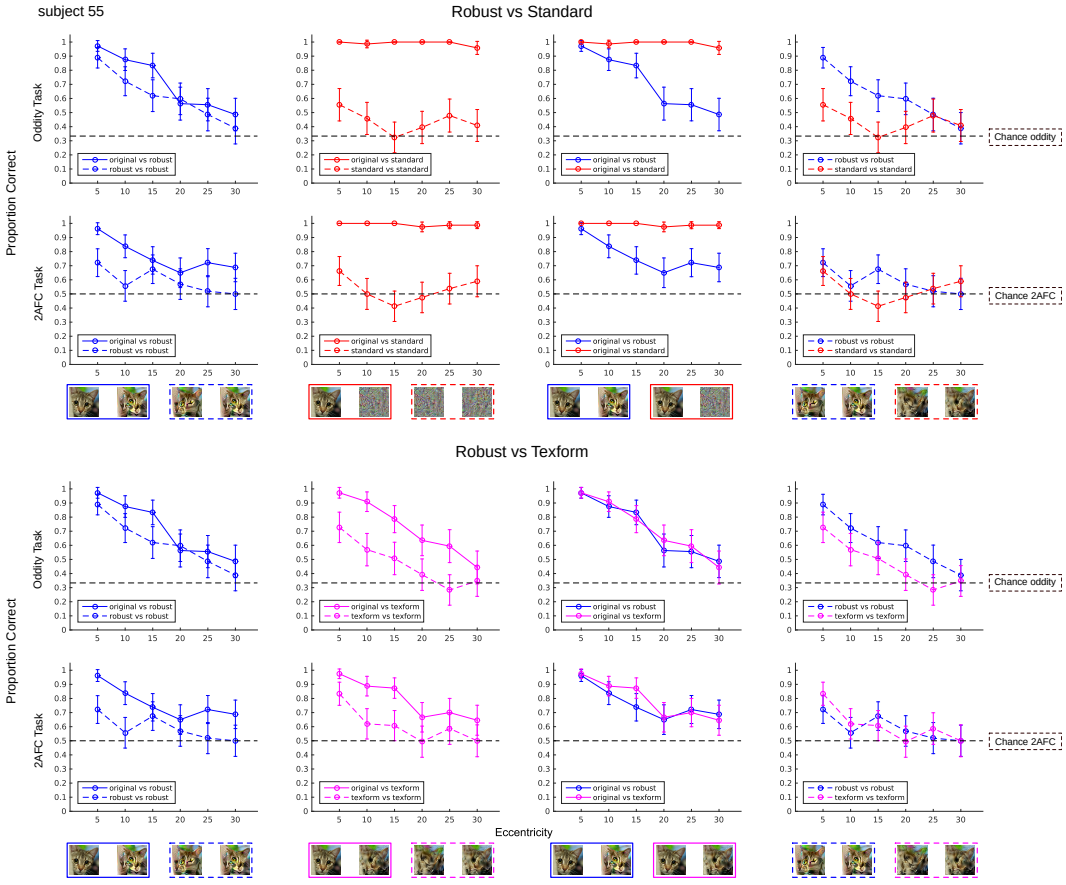

Figure 18: Subject 55

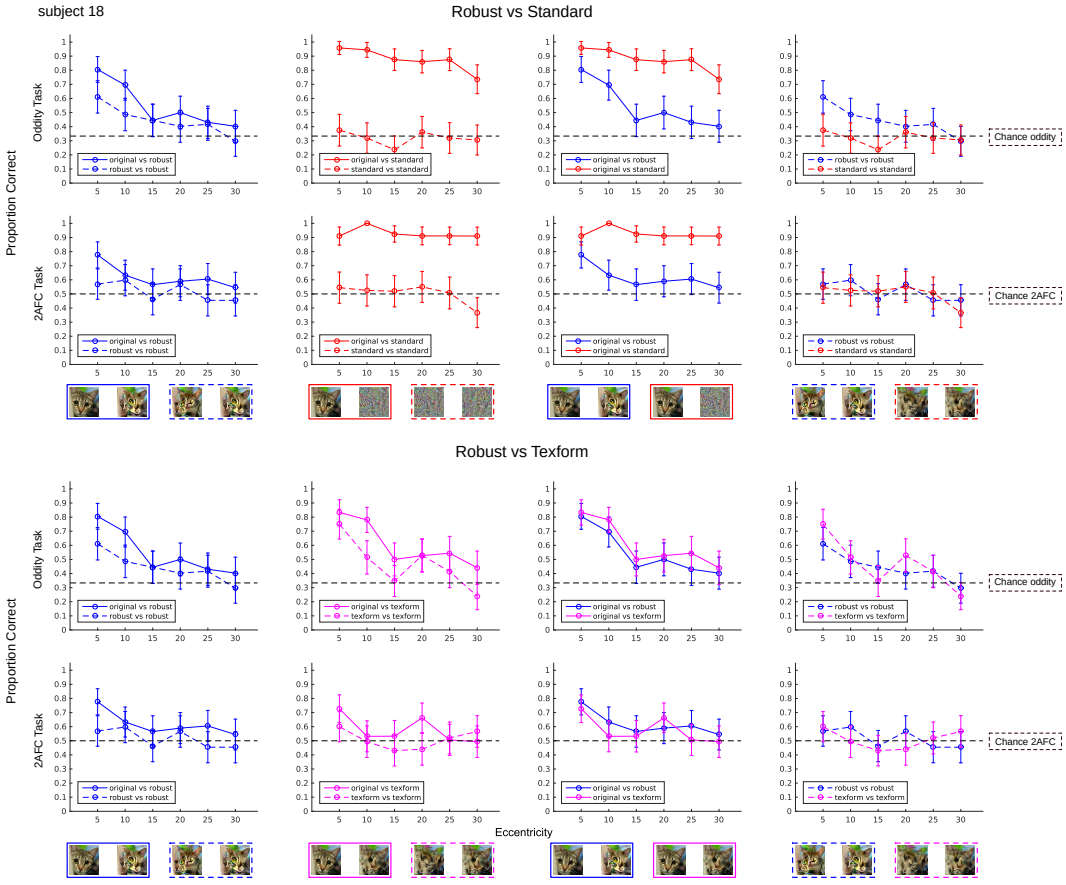

Figure 19: Subject 18

Figure 20: Subject 34

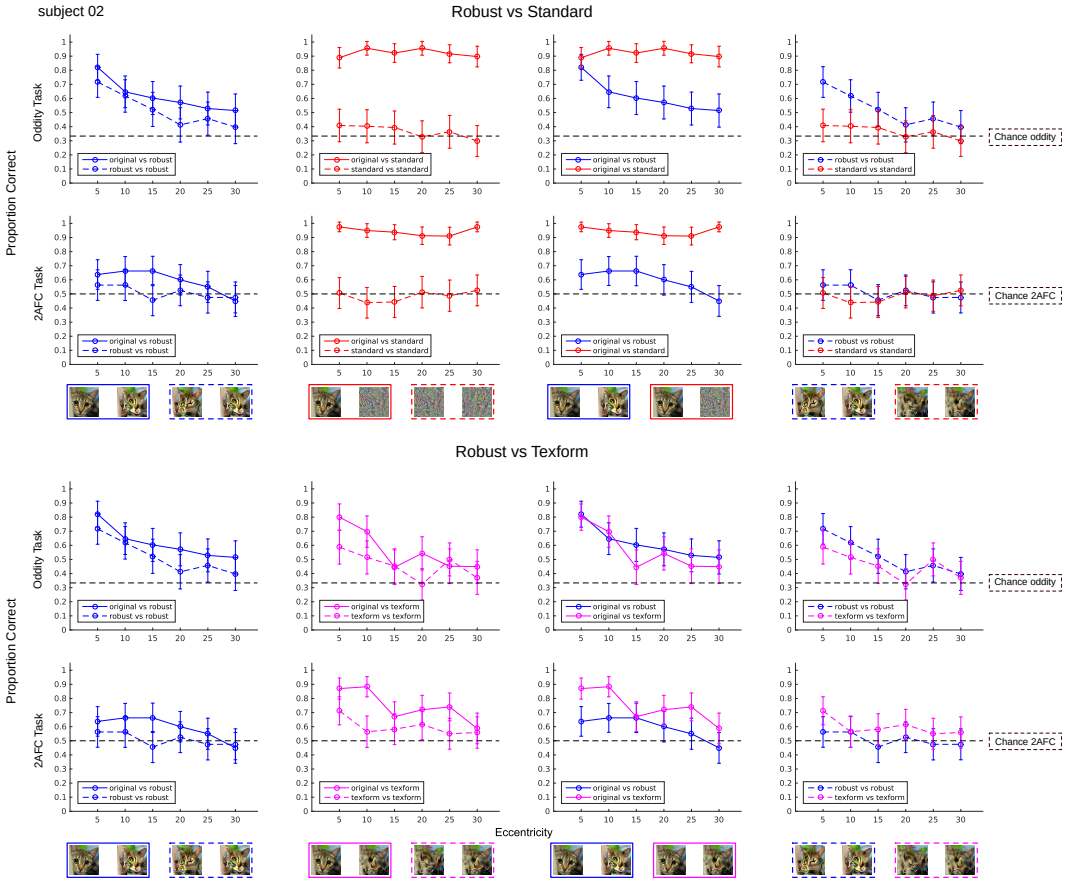

Figure 21: Subject 02

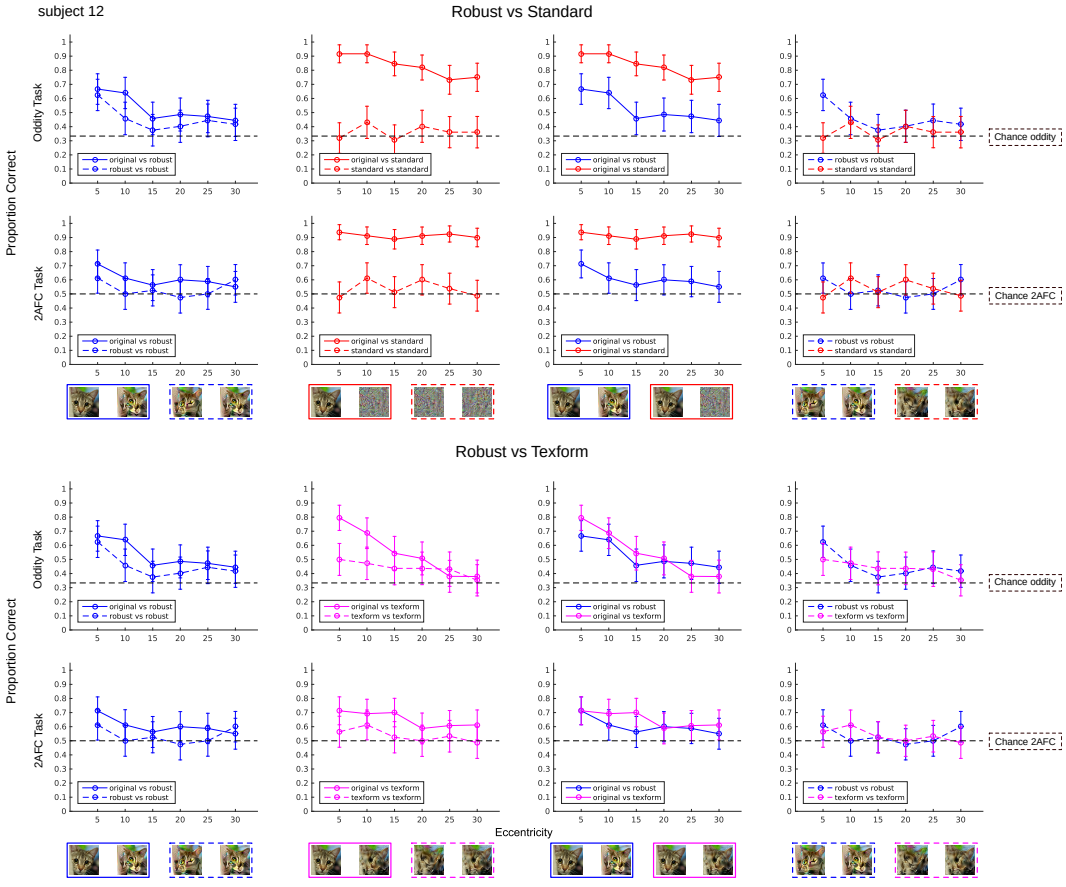

Figure 22: Subject 12

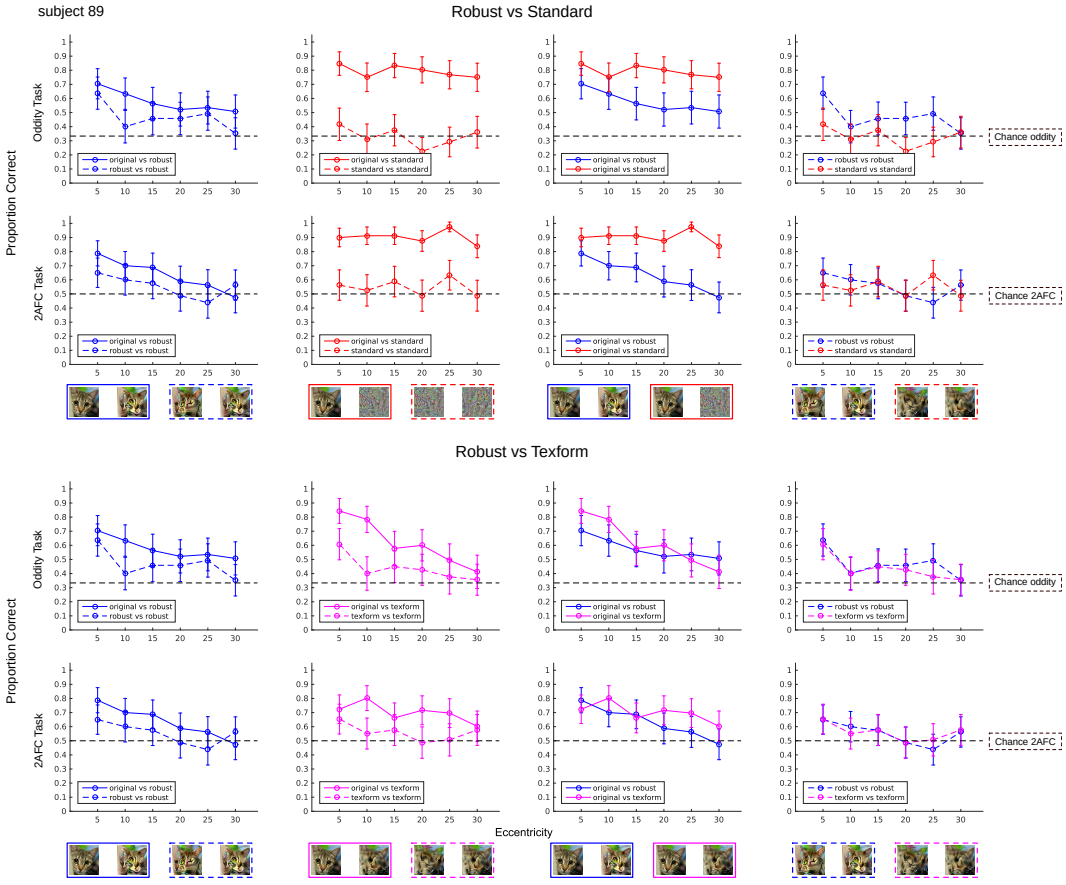

Figure 23: Subject 89

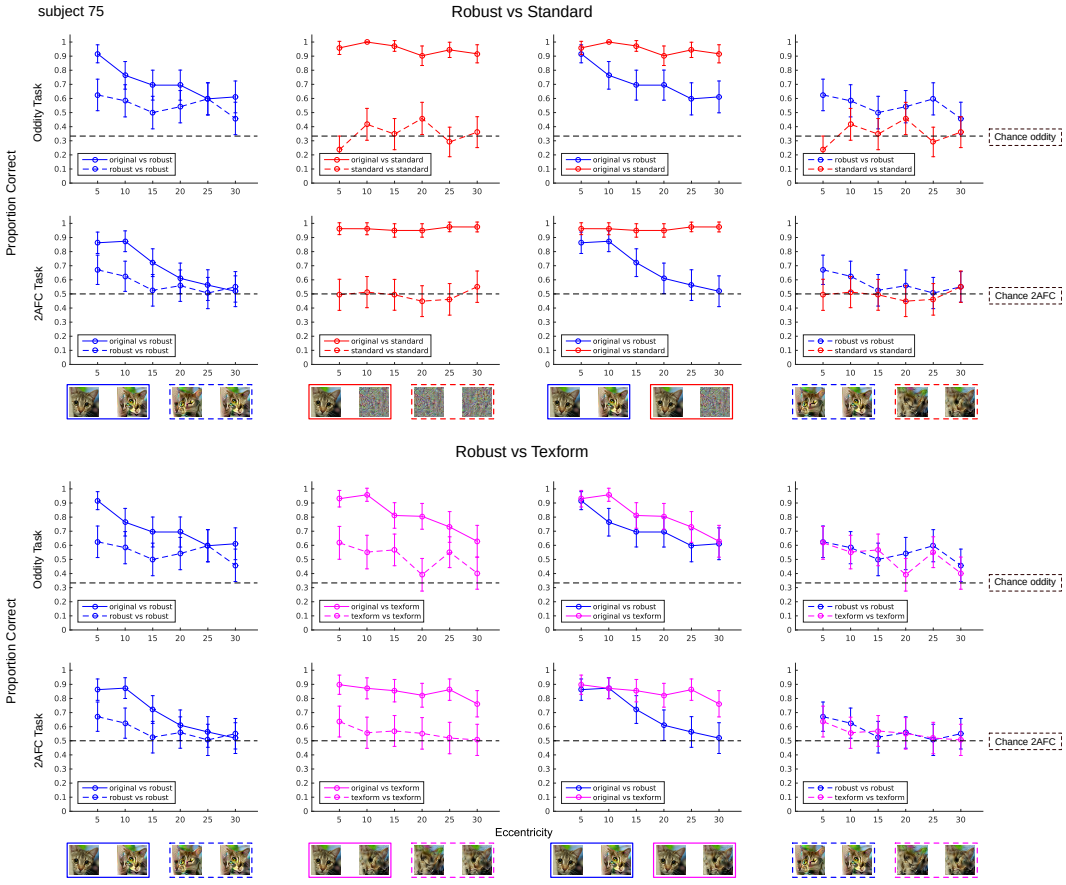

Figure 24: Subject 75

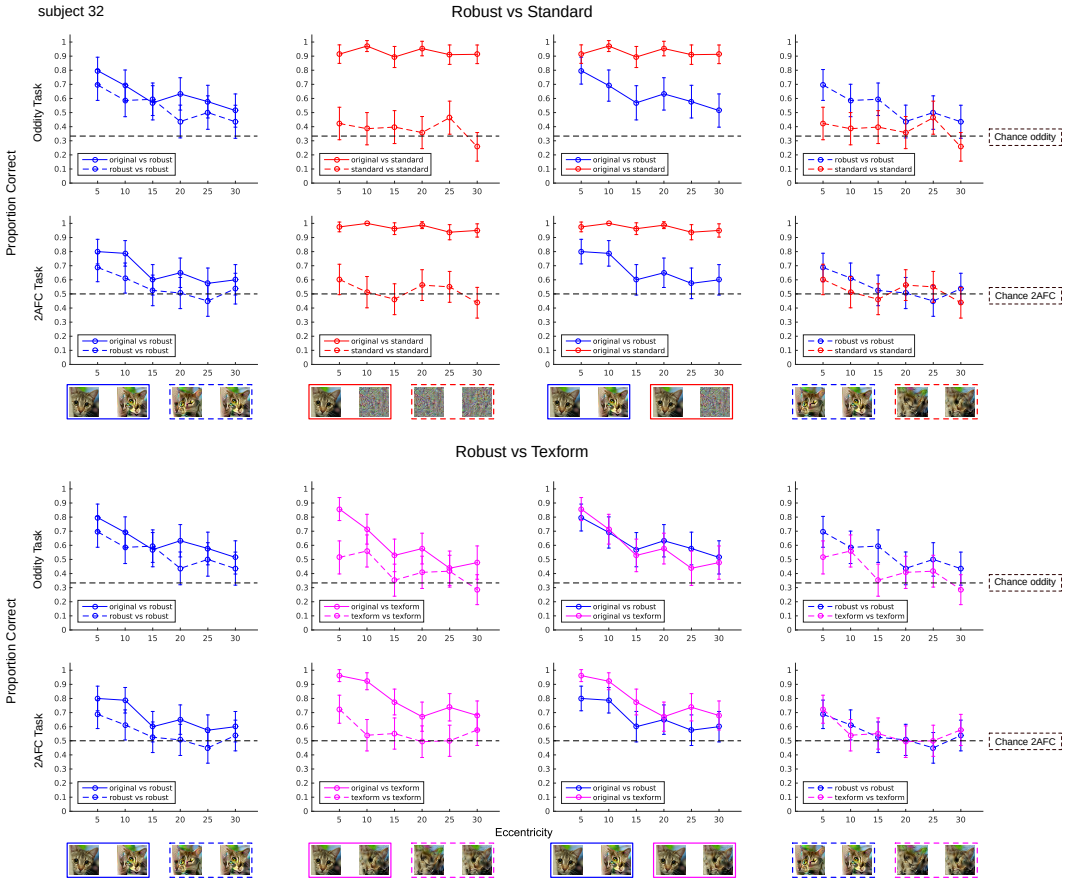

Figure 25: Subject 32

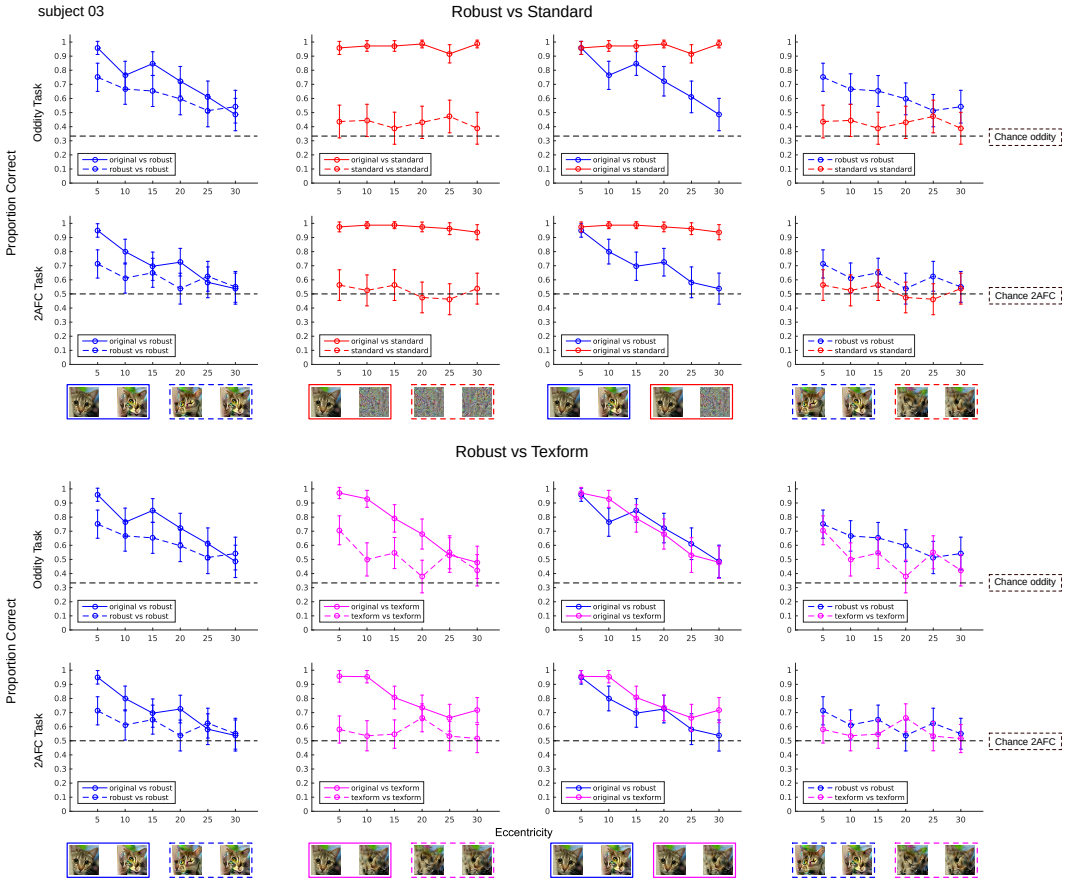

Figure 26: Subject 03