A Appendix

A.1 Details on Methods of Counterfactual-Inspired Experiment

We closely follow our previous work [5] and hence often refer to specific sections of it in this Appendix.

A.1.1 Data Collection

Exclusion Criteria In order to acquire data of high quality from MTurk, we integrate five exclusion criteria. If one or more of these criteria is not met, we post the same HIT again:

- Maximal number of attempts to reach 100% performance in practice trials: 5
- Performance threshold for catch trials: two out of three trials have to be correctly answered
- Answer variability: at least one trial must be chosen from the less frequently selected side (to discard participants who only responded with "left" or "right")
- Time to read the instructions: at least $20\,\mathrm{s}$ ($15\,\mathrm{s}$ in the none condition)
- Time for the whole experiment: at least $90\,\mathrm{s}$ and at most $900\,\mathrm{s}$ (at least $40\,\mathrm{s}$, and at most $900\,\mathrm{s}$ in the none condition)

Minimize Biases To minimize a bias to either query image, the location of the truly maximally activating query image is randomized and participants have to center their mouse cursor by pressing a centered button "Continue" after each trial.

Expert Measurements The two first authors complete all 10 image sets in multiple conditions: At first, they label the query images for the Primary Object Baseline. Then they answer the none, synthetic or natural (counterbalanced between the two authors), mixed, and blur condition. Clicking through the trials several times means that they see identical images repeatedly.

A.1.2 Stimulus Generation

Model In line with previous work (e.g. Borowski et al. [5], Olah et al. [40]), we use an Inception V1 network [53] trained on ImageNet [12, 49]. For more details, see Sec. A.1.2 "Stimuli Selection - Model" in Borowski et al. [5].

Natural Images as Query and Reference Images The natural reference and query images are selected from a random subset of $599,552$ training images of the ImageNet ILSVRC 2012 dataset [49]. For each unit, we select those images that elicit a maximal activation. More specifically, we choose the very most activating images as the query images and the next most activating images as reference images and ensure no overlap between query and references images as well as between image sets. As we follow our work published in Borowski et al. [5], please see A.1.2 for more details on the sampling procedure. In total, we generate 20 different image sets per unit. In the presented data, we only use half of these sets.

Query Images For the query images, we use the 20 maximally activating images for a given unit. To produce the manipulated query images, a square patch of 90×90 pixels is placed on the unperturbed query image. The side length of a patch corresponds to 40% of a preprocessed image's side length. The position of the occlusion patch is chosen such that the manipulated image's activation for a given unit is minimal (maximal) among all possible manipulated images' activations. This maximizes the signal in the query images and means that patches of the two query images can overlap.

In a control experiment, we test whether the partial occlusions of the natural ImageNet images cause the manipulated images to lie outside the natural image distribution. If this was the case, the query images would fail to be representative of the network's activity for natural images. Here, we test how similar the response to the unperturbed and partially occluded images is. Specifically, we count how often there is an overlap of the top-5 predictions. If network activations were drastically different for the occluded than for the unperturbed images, we should find low agreement. However, we do find an agreement for $97.8,\%$ of all tested images. Therefore, the square occlusions only have a marginal effect on the network's overall activity/predictions.

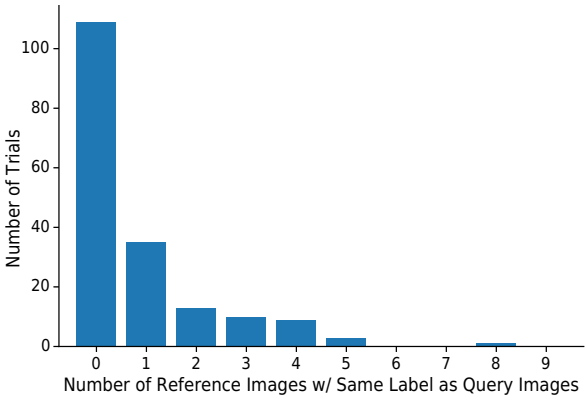


Figure 8: Distribution of the number of natural reference images that have the same label as the query image over the main trials used in the counterfactual-inspired experiment.

Reference Images: Natural Images In a control experiment, we test how often the label of the reference images coincide with the query image's label. If there was a high correspondence of these ImageNet labels, this could suggest that our experiment would rather reveal insights on how well humans would be able to *classify* images according to *labels* rather than to answer a counterfactual-inspired task based on the unit activations. Fig. 8 shows that the overlap of labels between query and reference images is low.

Reference Images: Blurred Images The blurred reference images are created by blurring all but one patch with a Gaussian kernel of size $(21, 21)$. This parameter choice allows participants to still get a general impression of an image, but not recognize details. Further, it is in line with work by Fong et al. [17]. The image choices are identical to the natural condition. Further — and just like for the query images — the position of the unblurred patch is chosen such that the manipulated image's activation for a given unit is maximal among all possible manipulated images' activations. Finally, the size of the unblurred patch is identical to the occlusion patch size: 40% of a preprocessed image's side length.

Reference Images: Synthetic Images from Feature Visualization Depending on the condition, we adjust the number of feature visualizations we generate: For the purely synthetic condition, we generate 9 visualizations, for the mixed condition, we generate 4 visualizations. As we follow our work published in Borowski et al. [5], please see A.1.2 for further details.

A.1.3 Baselines

Primary Object Baseline The Primary Object Baseline simulates that the more strongly activating manipulated image would be the one where the occlusion hides as little as possible from the most prominent object of the query image. To this end, the first two authors and the last author label all images. When doing so, they use a slightly modified logic: They select the image whose most prominent object is *most* occluded. If they cannot clearly identify a primary object in the image, they flag these trials, which are then treated differently in the analysis. For the analysis, the image choice is inverted again to counteract the inverted task that the authors responded to.

The performance reported in Fig. 4 is calculated by averaging over the three individual performances. Each individual performance itself is in turn estimated as the expectation value over random sampling for query images with no clear primary object. This analysis is in line with how the performance of MTurk participants is analyzed. An alternative option would be to take the majority vote of the three answers. When randomly sampling the choice for query images with no clear primary object, taking the majority votes and evaluating the expected accuracy, the performance would evaluate to 0.70 ± 0.02. Notably, 58 of all 180 trials are affected by the sampling as two or more authors responded with a confidence of 1 in 36 trials, and one author responded with a confidence of 1 while the other two gave opposing answers in 22 trials. This represents a fairly large fraction and reflects that many images on ImageNet have more than one prominent object [55, 3]. Consequently, there may not be a ground-truth for each trial in the Primary Object Baseline.

Saliency Baseline The Saliency Baseline simulates that participants select the image with a patch occluding the less prominent image region. To this end, we pass the unoccluded query image through the saliency prediction model DeepGaze IIE [29] which yields a probability density over the entire image. Next, we integrate said density over each of the two square patches. We then select the image with a lower value indicating that less important information is hidden by the occlusion patch.

A.1.4 Trials

Main trials For both the 3×3 and the POOL branch of each of the 9 layers with an Inception module, one randomly chosen unit is tested (see Table 1). These are the same units as in Experiment I of Borowski et al. [5].

Table 1: Units used as main trials in the 3×3 as well as the POOL branch in the counterfactual-inspired experiment. The numbers in brackets after each layer's name correspond to the numbering used in all our plots.

	Unit	
Layer	3×3	POOL
mixed3a (1)	189	227
mixed3b (2)	178	430
mixed4a (3)	257	486
mixed4b (4)	339	491
mixed4c (5)	247	496
mixed4d (6)	342	483
mixed4e (7)	524	816
mixed5a (8)	278	743
mixed5b (9)	684	1007

Instruction, Practice and Catch Trials The instruction, practice and catch trials are hand-picked by the two first authors. As a pool of units, the appendix overview of Olah et al. [40] as well as the "interpretable" POOL units used in Experiment I and all units used in Experiment II of Borowski et al. [5] are used. After generating all 20 reference and query image sets for these units, the authors select those units and image sets that they consider easiest (see Table 2).

Instruction Trial To explain the task as intuitively as possible, we construct an easy, artificial instruction trial (see Fig. 9 and 10): At first, we select a unit with easily understandable feature visualizations: The synthetic images of unit 720 of the POOL branch in layer 8 show relatively clear bird-like structures. From a popular image search engine, we then select an image[8] which not only clearly shows a bird but also other objects, namely a dog and water. To construct the minimally and maximally activating query images, we place the occlusion patches manually on the bird and dog.

Practice Trials In each attempt to pass the practice block, the trials are randomly sampled from a pool of 10 trials (see Table 2). Please note that unlike in any other trial type, participants receive feedback in the practice block: After each trial, they learn whether their chosen image truly is the query image of higher activation.

Catch Trials While all conditions with reference images use hand-picked easy trials (see Table 2), the none condition cannot rely on straight-forward clues from references. Therefore, we exchange the minimal query image with a minimal query image of a different, otherwise unused unit. This ensures that the catch trials in the none condition are also obvious.

A.1.5 Infrastructure

The online experiment is hosted on an Ubuntu 18.04 server running on an Intel(R) Xeon(R) Gold 5220 CPU. The experiment is implemented in JavaScript using jspsych 6.3.1 [11] and flask via

[8] `https://pixnio.com/fauna-animals/dogs/dog-water-bird-swan-lake-waterfowl-animal-swimming` released into public domain under CC0 license by Bicanski.

Table 2: Hand-picked unit choices for instruction, catch and practice trials in the counterfactual-inspired experiment.

Trial Type	Layer	Branch	Unit	Difficulty Level
instruction	mixed5a	pool	720	very easy
catch	mixed4e	pool	783	very easy
	mixed4c	pool	484	very easy
	mixed5a	3×3	557	very easy
practice	mixed4e	1×1	6	very easy
	mixed4a	pool	505	very easy
	mixed4e	pool	809	very easy
	mixed4c	pool	449	easy
	mixed4b	pool	465	easy
	mixed4c	1×1	59	easy
	mixed4e	1×1	83	easy
	mixed3a	1×1	43	easy
	mixed3b	pool	472	easy
	mixed4b	1×1	5	easy

Python 3.6. The generation of the stimuli shown in the experiment was completed in approximately 35 hours on a single GeForce GTX 1080 GPU. The calculation of all baselines required 8 additional GPU hours.

A.1.6 Amazon Mechanical Turk

MTurk participants To increase the chance that all MTurk participants understand the English instructions at the beginning of the experiment, we restrict access to workers from the following English-speaking countries: USA, Canada, Great Britain, Australia, New Zealand and Ireland.

Financial Compensation Based on an estimated duration and pilot experiments as well as a targeted hourly rate of US$ 15, we calculate the pay to be US$ 0.70 for the none condition and US$ 1.95 for all other conditions. MTurk participants whose data we include need a mean time of 209.64 ± 79.53 s and 396.87 ± 145.78 s for the whole experiment for the none condition and for all other conditions, respectively, which results in an hourly compensation of ≈ 12.02 US$/hour and 17.69 US$/hour, respectively. All MTurk participants who fully complete a HIT are paid, regardless of whether their responses meet the exclusion criteria. A total of US$ 1989.06 is spent on all pilot and real replication and counterfactual-inspired experiments.

Rights to Data We do not gather personal identifiable data from any MTurk participant. According to the MTurk Participation Agreement 3a [9], workers agree to vest all ownership and intellectual property rights to the requester (i.e., the authors of this study). Besides informing MTurk participants in the HIT preview about the academic and image classification nature of the experiment, we restate that "By completing this HIT, you consent to your anonymized data being shared with us for a scientific study." Further, we provide an email address, which some MTurk participants used to share feedback.

[9] `https://www.mturk.com/participation-agreement`, accessed on May 22nd, 2021

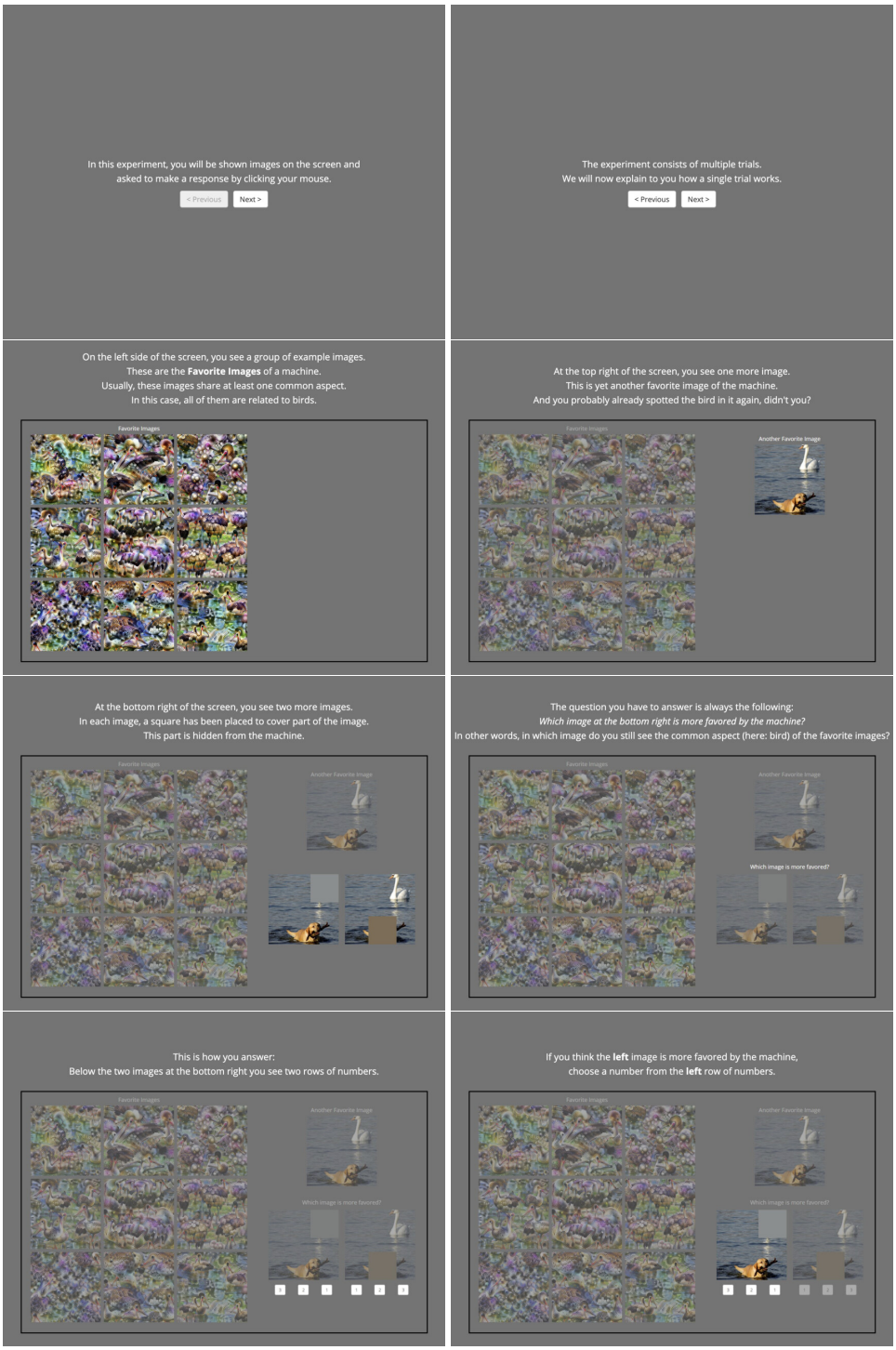


Figure 9: First eight instructions at the beginning of the counterfactual-inspired experiment.

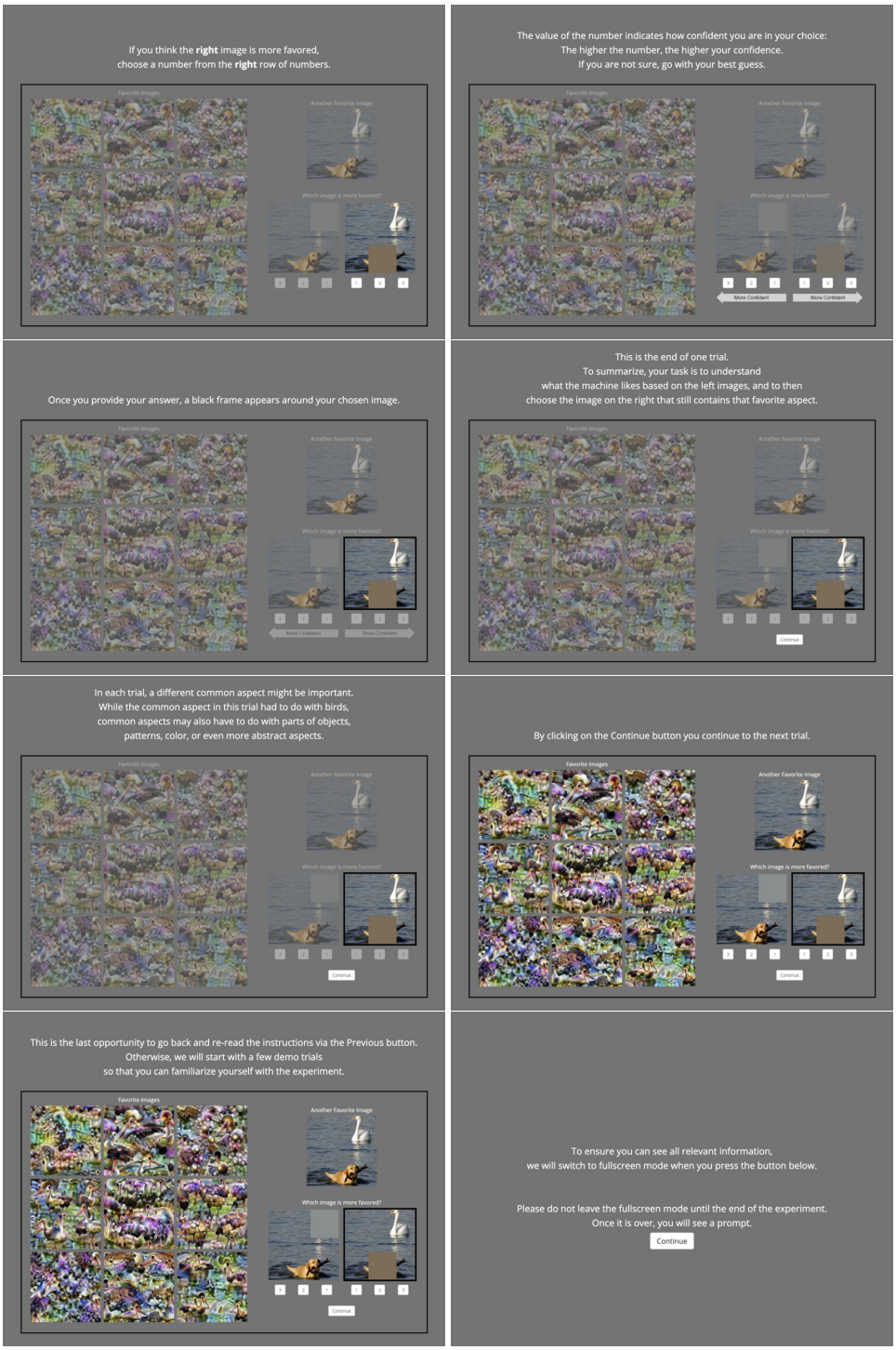


Figure 10: Second eight instructions at the beginning of the counterfactual-inspired experiment.

A.2 Details on Results of Counterfactual-Inspired Experiment

A.2.1 Different Query images

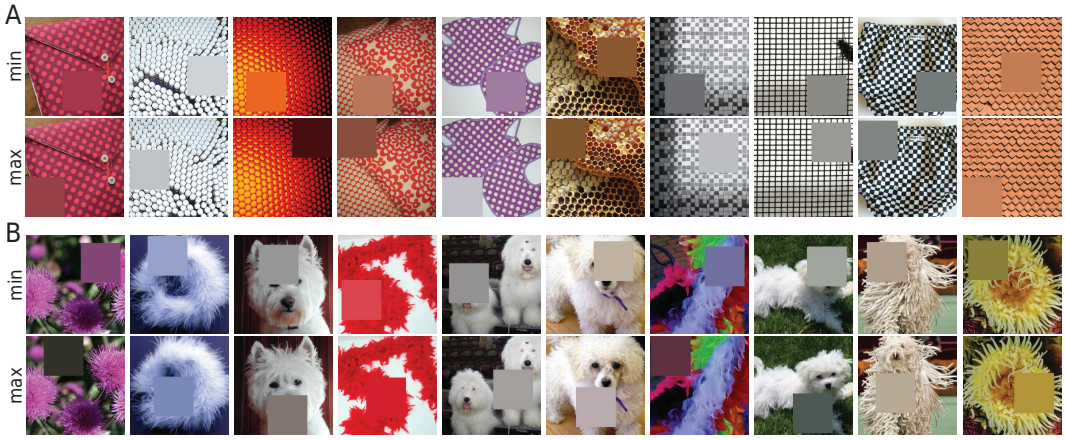
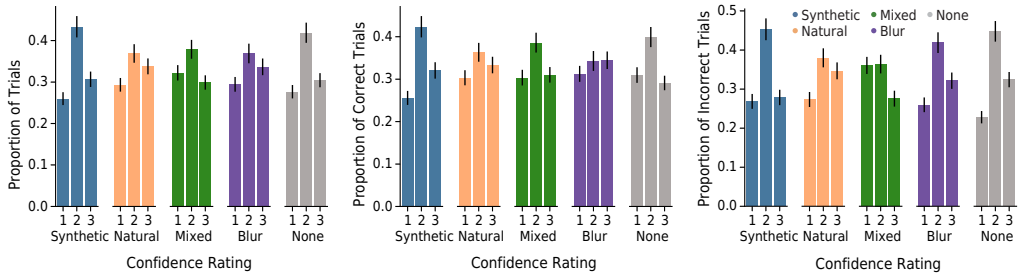


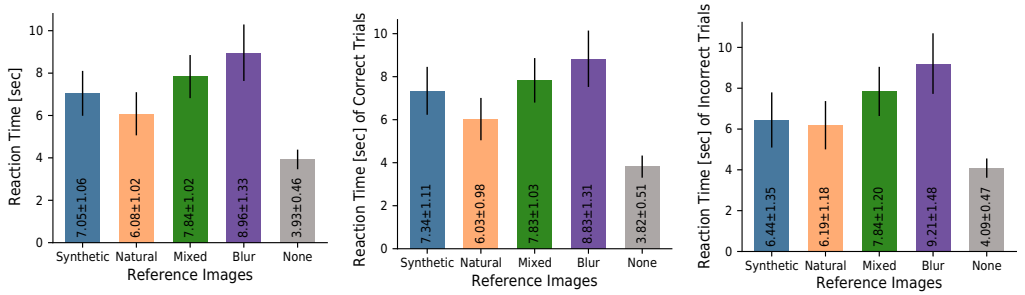
Figure 11: For each unit, we test 10 different image sets in the counterfactual-inspired experiment. The diversity of query images for layer 3 of the 3×3 branch (**A**), and layer 7 of the POOL branch (**B**) gives an intuitive explanation for varying performances.

A.2.2 Confidence Ratings and Reaction Times



(a) (b) (c)

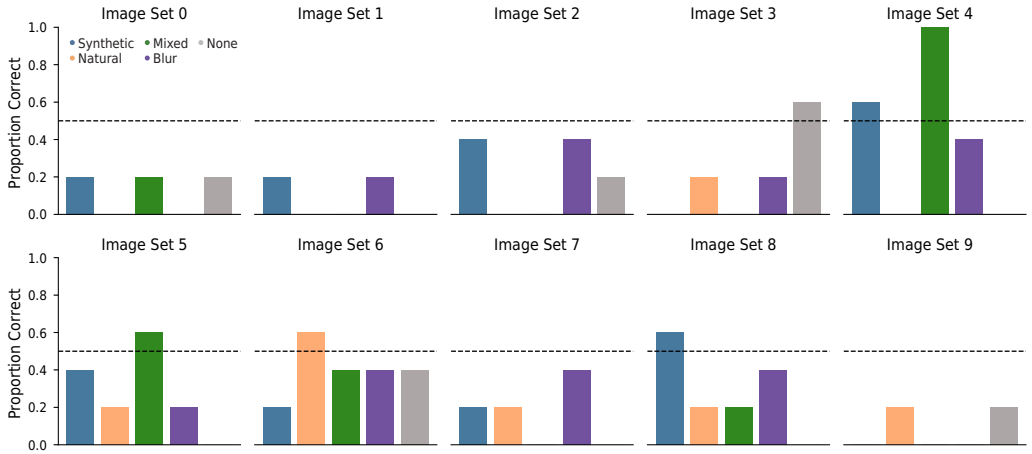
Figure 12: Confidence ratings of MTurk participants in the different reference conditions for (a) all, (b) only correct or (c) only incorrect trials of the counterfactual-inspired experiment.



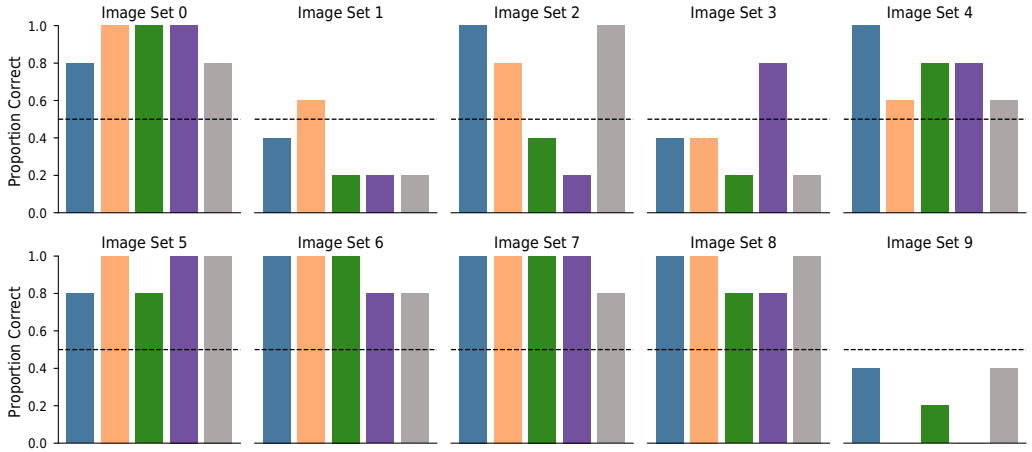
(a) (b) (c)

Figure 13: Reaction times of MTurk participants in the different reference conditions for (a) all, (b) only correct or (c) only incorrect trials of the counterfactual-inspired experiment.

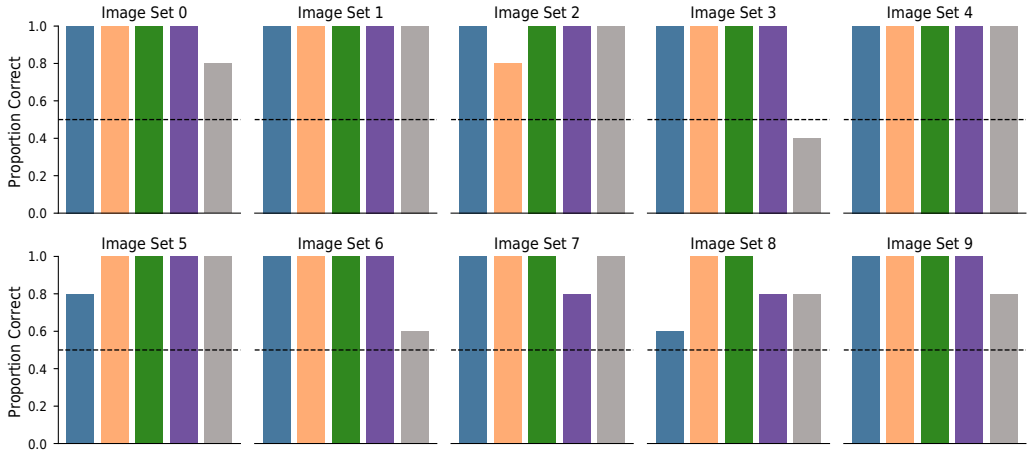
A.2.3 Performance per Image Set



(a) Difficult unit.



(b) Intermediate unit.



(c) Easy unit.

Figure 14: Performance in the counterfactual-inspired experiment split up by image sets and conditions for a difficult (layer 3, POOL), intermediate (layer 7, POOL) and easy unit (layer 8, POOL). Each bar shows the average over 5 MTurk participants.

A.2.4 Strategy Comparisons

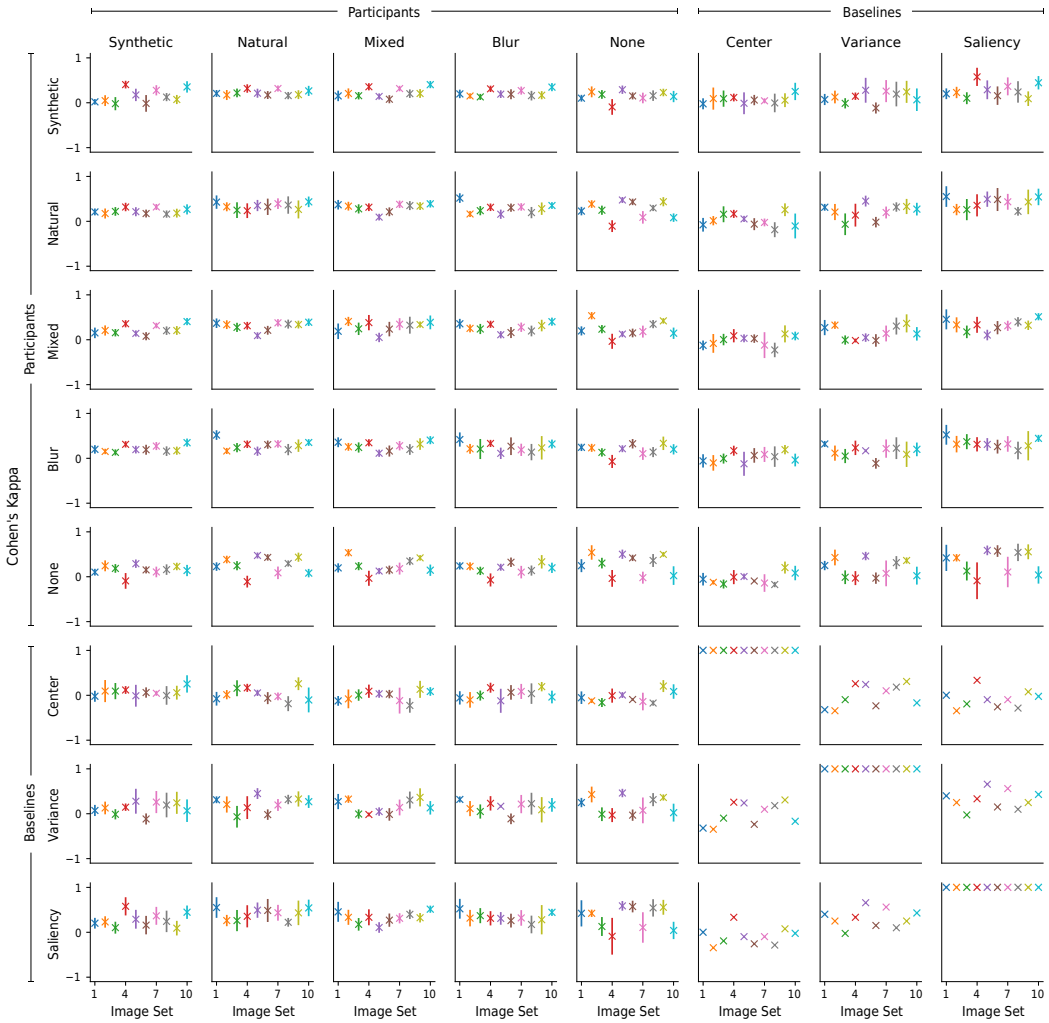
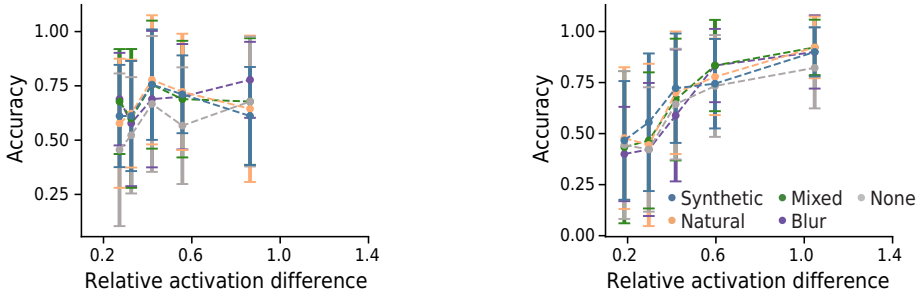


Figure 15: Cohen's kappa per image set in the counterfactual-inspired experiment (averages over participant-participant-, participant-baseline- or baseline-baseline-pairs). Error bars denote two standard errors of the mean.

A.2.5 Relative Activation Differences

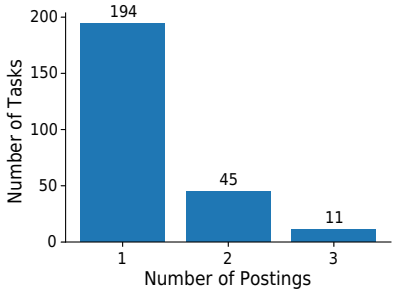


(a) 3×3 branch.

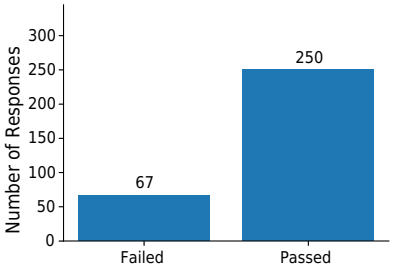
(b) POOL branch.

Figure 16: Accuracy in the counterfactual-inspired experiment as a function of the relative activation difference between the two query images for the (a) 3×3 branch and the (b) POOL branch. Here, the data points shown in Fig. 7 are summarized in 5 bins of equal counts; the plot shows the mean and standard deviation for each of the bins.

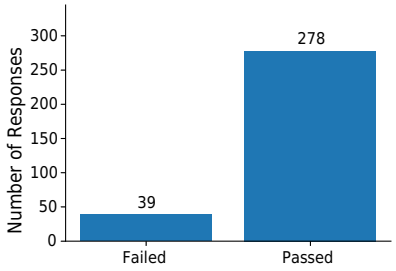
A.2.6 Exclusion Criteria



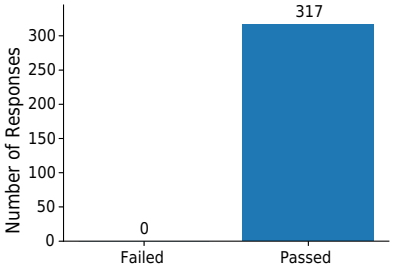
(a) Number of times a HIT is posted.



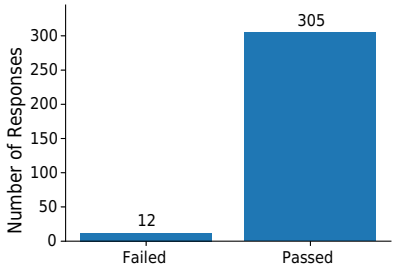
(b) All exclusion criteria.



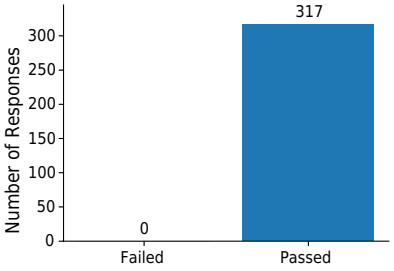
(c) Exclusion criterion: catch trials.



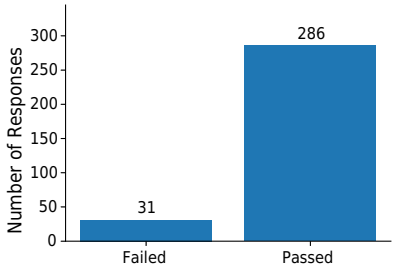
(d) Exclusion criterion: row variability.



(e) Exclusion criterion: instruction time.

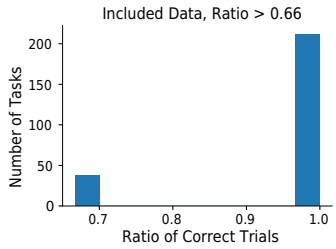


(f) Exclusion criterion: total response time.

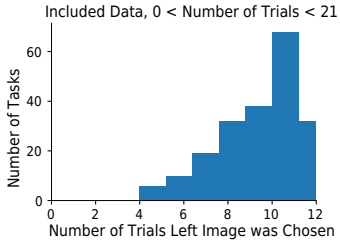


(g) Exclusion criterion: practice block.

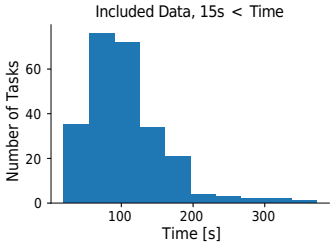
Figure 17: (a) Number of times a HIT is posted. To limit the financial risk, we limit the maximal number of times that a HIT can be posted at 3. (b-g) Distributions of MTurk participants that passed/failed the exclusion criteria in the counterfactual-inspired experiment on MTurk. Note that the sum of the counts of responses for the individual exclusion criteria in c-f is higher than the summary in b because a participant may have failed more than one exclusion criterion.



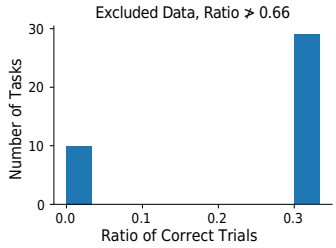
(a) Catch trials from included data.



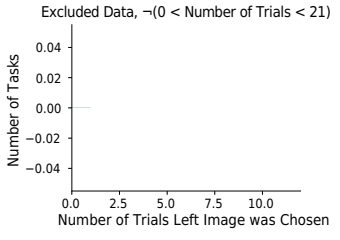
(b) Row variability from included data.



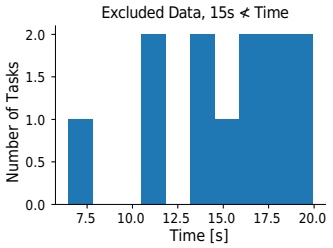
(c) Instruction time from included data.



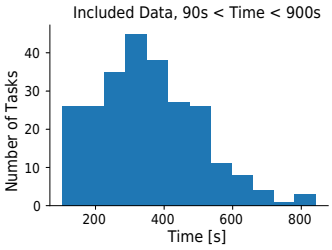
(d) Catch trials from excluded data.



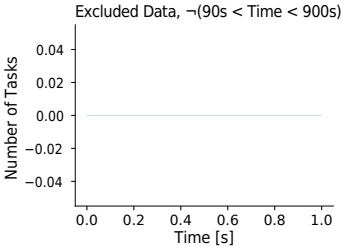
(e) Row variability from excluded data.



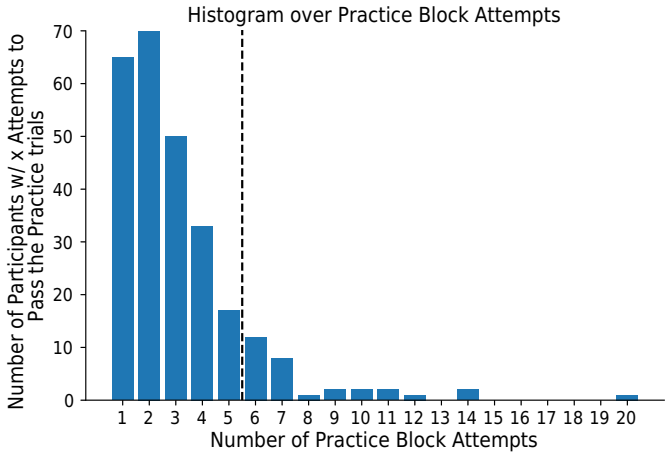
(f) Instruction time from excluded data.



(g) Total response time from included data.



(h) Total response time from excluded data.



(i) Practice Block Attempts: We include data from people who needed 5 or fewer attempts.

Figure 18: Distributions of the individual values controlled by the exclusion criteria in the counterfactual-inspired experiment on MTurk. For the first four criteria, a - c and g (d - f and h) show the data for the included (excluded) data. The final criterion in i shows a joint distribution.

A.3 Replication of the Main Result of Borowski et al. [5]

To check whether collecting data on a crowdsourcing platform yields sensible data in our case, we first test whether we can replicate the main finding of our previous human psychophysical experiment on feature visualizations [5]. In the latter, we found in a well-controlled lab environment that natural reference images are more informative than synthetic ones when choosing which of two different images are more highly activating for a given unit. Below, we report how we alter the experimental set-up to turn the lab experiment into an online experiment on MTurk and what results we find.

A.3.1 Experimental Set-up

While keeping as many aspects as possible consistent with our original study [5], we make a few changes: (1) We run an online crowdsourced experiment on MTurk, instead of in a lab. (2) Instead of testing the 45 units used in the original Experiment I, we only test one single branch of each Inception module, namely the 3×3 kernel size. This is a reasonable decision given that the main finding of the superiority of natural over synthetic images was present in all branches and that there was no significant difference per condition between different branches. (3) We exchange the within-participant design for a between-participant design, i.e. one MTurk participant does one condition only, namely either the natural or synthetic reference condition. This version is more suitable for short online experiments. (4) Instead of testing 10 participants in the lab, we test 130 MTurk participants per condition, i.e. 260 in total. This number of participants is estimated with an a priori power analysis using the SIMR package [23] to allow us to detect an effect half as large as the one observed in Borowski et al. [5] 80% of the time. Assumptions about variance, average performance, and effect size are chosen to be conservative relative to the original study because we expect MTurk participants' responses to be noisier.

One HIT on MTurk consists of 1 extensively explained instruction trial, 2 practice trials, and then 9 main trials that are randomly interleaved with a total of 3 catch trials. Each trial type is sampled from a disjoint pool of units: All participants see the same unit for the instruction trial; the catch trials are sampled from the same pool as in the original experiment, and the practice trials are the units that were used as interpretability judgment trials in [5], namely mixed3a, kernel size 1×1, unit 43; mixed4b, POOL, unit 504; mixed5b, 1×1, unit 17. A total of 13 participants see the same main trials that one lab participant saw. The order of the main and catch trials per participants is randomly arranged.

Exclusion Criteria If a participant's response does not meet one or more of the following criteria, which were determined before data collection, we discard it and post the same HIT again:

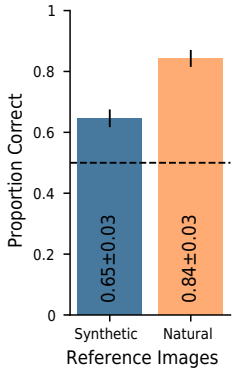
- Performance threshold for catch trials: two out of three trials have to be correctly answered
- Answer variability: at least one trial must be chosen from the less frequently selected side (to discard participants who only responded with "up" or "down")
- Time to read the instructions: at least $15\,\mathrm{s}$
- Time for the whole experiment: at least $90\,\mathrm{s}$ and at most $600\,\mathrm{s}$

MTurk compensation Based on an estimated and pilot experiment duration as well as an hourly rate of US$\$\,15$, we calculate the pay to be US$\$\,1.25$. We pay all MTurk participants who fully complete the experiment regardless of whether they succeed or fail in the exclusion criteria. The experiment without pilot experiments costs US$\$\,447$. MTurk participants whose data we include need a mean time of $220.70 \pm 71.58\,\mathrm{s}$ for the whole experiment, which results in an hourly compensation of ≈ 20.39 US$/hour.

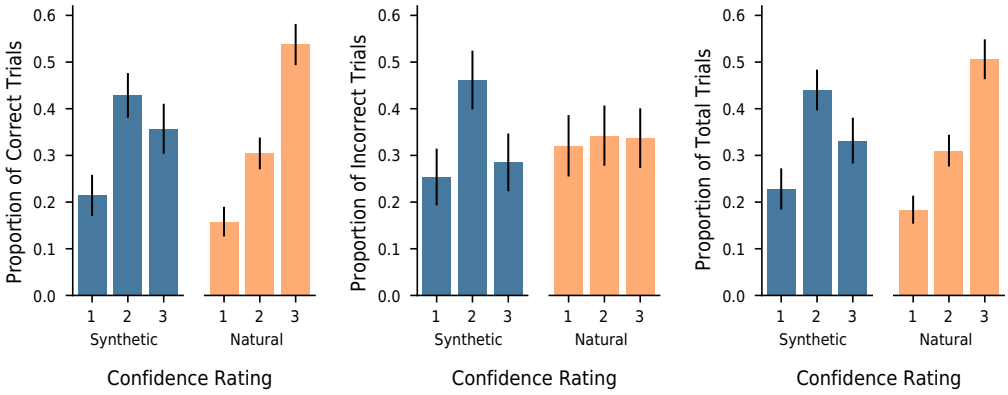
A.3.2 Results

MTurk participants achieve a higher performance when given natural than synthetic reference images: $84 \pm 3\,\%$ vs. $65 \pm 3\,\%$ (see Fig. 19a). Qualitatively, this result is the same as in the original Experiment I, see Figure 16 in Borowski et al. [5]. More precisely, the data shows a 1.35 (2.1) times larger odds (accuracy) difference for the replication. Compared to the lab data, MTurk participants seem more confident on the synthetic condition (see Fig. 19b-d), are faster in the synthetic condition (see Fig. 19e-g), and are about as fast in the natural condition (see Fig. 19e-g).

Fig. 20 shows that most participants passed the exclusion criteria. For more details on the number of postings per HIT and for more details on the MTurk participants' performance on the exclusion criteria, see 21.



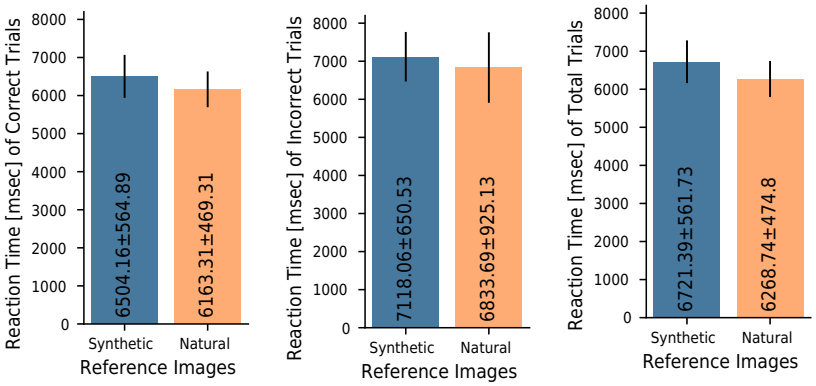
(a) Performance.



(b) Confidence ratings on correctly answered trials.

(c) Confidence ratings on incorrectly answered trials.

(d) Confidence ratings on all trials.

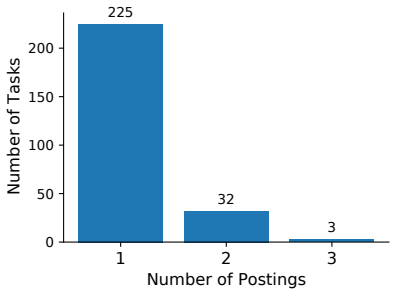


(e) Reaction time on correctly answered trials.

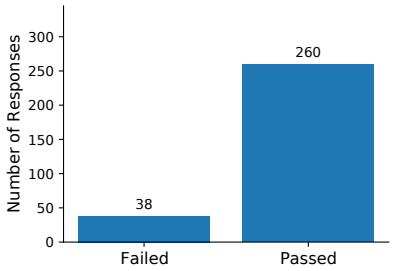
(f) Reaction time on incorrectly answered trials.

(g) Reaction time on all trials.

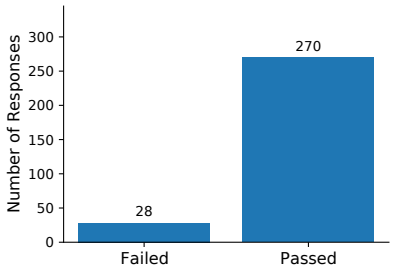
Figure 19: Results of the replication experiment of Borowski et al. [5] on MTurk for kernel size 3×3: task performance (a), distribution of confidence ratings (b-d) and reaction times (e-g).



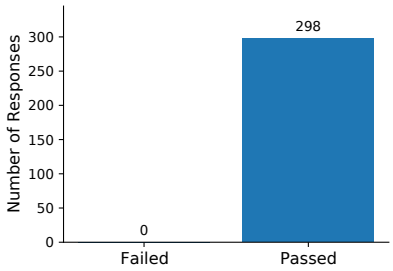
(a) Number of times a HIT is posted.



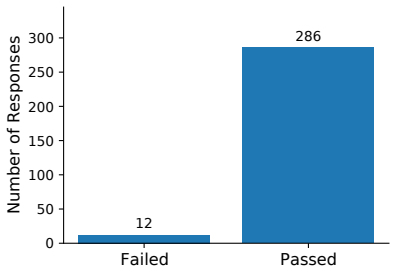
(b) All exclusion criteria.



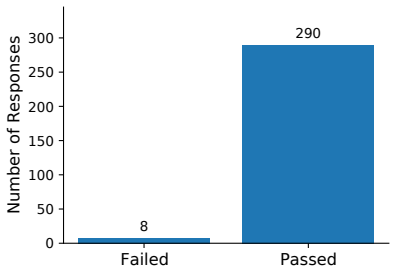
(c) Exclusion criterion: catch trials.



(d) Exclusion criterion: row variability.

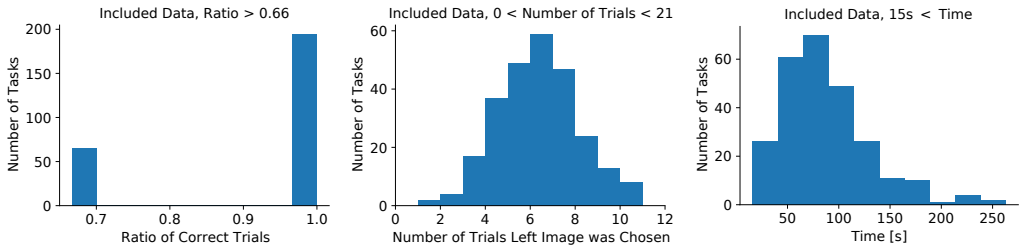


(e) Exclusion criterion: instruction time.

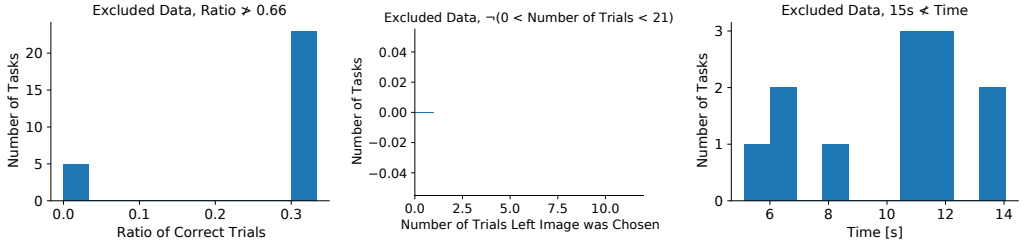


(f) Exclusion criterion: total response time.

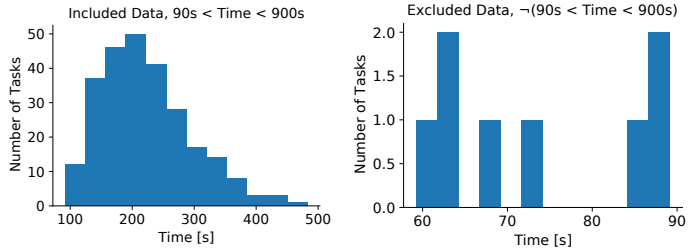
Figure 20: (a) Number of times a HIT is posted. (b-f) Distributions of MTurk participants that passed/failed the exclusion criteria in the replication experiment on MTurk. Note that the sum of the counts of responses for the individual exclusion criteria in c-f is higher than the summary in b because a participant may have failed more than one exclusion criterion.



(a) Catch trials from included data. (b) Row variability from included data. (c) Instruction time from included data.



(d) Catch trials from excluded data. (e) Row variability from excluded data. (f) Instruction time from excluded data.



(g) Total response time from included data. (h) Total response time from excluded data.

Figure 21: Distributions of the individual values controlled by the exclusion criteria in the replication experiment on MTurk. Figures a - c and g (d - f and h) show the data for the included (excluded) data.