## Supplementary methods

### *Participants*

Sixteen right-handed participants (9 female, mean age 24 years and 7 male, mean age 25 years) participated in the experiment for monetary reward. All participants signed an informed consent form and had normal, or corrected-to-normal, vision (mean left eye 2020, mean right eye 2030).

### *Stimuli*

We constructed a set of 12 Kanizsa-control pairs of different shapes: triangles, squares and pentagons (see Figure S1). Traditionally, controls are created by rotating the inducers outwardly. Although such controls retain the overall configuration of the inducers, they allow for Kanizsa recognition using low-pass spatial frequency filters (1). Moreover, the support ratio of the stimulus (the ratio between the physically specified side length and illusory side length) is obliterated in such controls. We therefore constructed controls in which one or more of these characteristics were optimally matched with their Kanizsa counterpart; keeping the shapes of the inducers intact (see 1, 2, 3, 4, 6, 7, 9, 10, 11 in Fig. S1), retaining the low spatial frequency characteristics of the global stimulus (see 1, 4, 5, 8, 12 in Fig. S1), as well as maximizing the support ratio of controls (see 1, 2, 3, 4, 5, 6, 8, 11 in Fig. S1) when compared to their Kanizsa counterparts. In cases where large inducers were rotated outwardly (see 1, 6, 7, 9 10, 11 in Fig. S1), we rotated the inducers around their center of gravity rather than around their "veridical" center, so as to further minimize differences between Kanizsas and controls in terms of their global spatial frequency characteristics.

As a result, the stimulus set contained large variations in terms of physical properties across stimulus instances, but had similar physical properties within any given Kanizsa-control pair. Because the decoding analyses involve single trial extraction of class membership that needed to carry over from one stimulus instance to the next in order to be able to work (i.e. the task was to classify stimuli based on the existence of surface information, irrespective of the physical features of the inducers or the shape of the configuration of the inducers), differences we observed in the Kanizsa-control dimension could not be explained by any single physical property, but were particular to differences resulting from perceptual integration. Put differently: neither the subjects nor the classifier could solve the task by using particular features of any of the inducers, the only way of solving the classification task was to perceptually integrate the features and establish the existence of surface information to determine class membership. Finally, the total region covered by Kanizsa figures (including inducers) was 9.4° by 8.5° degrees visual angle for triangles, 7.4° by 7.4° for squares and 7.7° by 7.7° for pentagons, keeping only the size of the illusory surface region approximately the same between shape types.

Masks were created by randomly rotating inducer elements from the Kanizsa and control images (see Fig. S2). There were 10 masks for each stimulus shape. Masks were picked randomly from these sets, but always matching masks to shapes, so that triangular Kanizsas and controls would be followed by triangularly organized masks, square Kanizsas and controls by square masks and pentagonal Kanizsas and controls by pentagonal masks. All stimuli and masks were generated using Adobe Illustrator CS6 (Adobe Systems Incorporated, San Jose, CA, USA).

### *Procedure and tasks main experiment*

All tasks were programmed in Presentation (Neurobehavioral Systems, Inc., Berkely, CA, USA) and displayed on a 19 inch CRT-monitor running at 100 Hz. Subjects participated in a total of three sessions. The first session was a training session to make subjects familiar with the task and the stimulus set. In this task, Kanizsa and control images were presented for 10 ms and participants were prompted to identify whether the image contained a surface. When they were able to perform this task with an accuracy of more than 90% they continued with the next task. In the second part of the practice session, participants performed a no-blink (long lag) version of the experimental task to determine if they were able to correctly identify black T1 and T2 targets amidst an RSVP of red distractors. Subject performance was computed as hit rate (the fraction of Kanizsa figures categorized as containing a surface) minus false alarm rate (the fraction of control figures categorized as containing a surface). If their hit rate minus false alarm rate exceeded .8 in both T1 and T2, they went on to the third and final part of the practice session. In this part they performed two versions of the experimental task to determine at what latency the T2 induced the largest attentional blink for that subject.

The task was the same as the experimental task, but did not include the masked conditions. The difference between these two versions was the inter stimulus interval (ISI). In the first version the ISI was 150ms (resulting in a short AB lag of 300 ms), while the second version had an ISI of 100ms (resulting in a short AB lag of 200 ms). If the participants were not able to perform adequately in one of the tasks or did not show a sufficiently strong AB, they were excluded from the rest of the experiment. Eight participants performed the EEG sessions at an ISI of 150ms (short AB lag: 300 ms) three participants did the task at an ISI of 100ms (short AB lag: 200 ms), and five participants were excluded after the first training session for not meeting one or more of the above criteria for inclusion.

Subsequently, subjects took part in two separate sessions on separate days, in which they performed the experimental task while their EEG was recorded (see main text for task details). At the end of each trial, a response screen would appear, asking subjects to indicate whether the first and/or second target contained a surface. Subjects gave two responses, the first for T1, the second for T2. Responses consisted of button presses using a two-button box attached to the right arm of the chair, left button indicating “I perceived a surface” and right button indicating “I did not perceive a surface”.  Each of the two sessions consisted of 9 blocks of this experimental task. Each block consisted of 24 Kanizsa and 24 control images for each of the four conditions resulting in 192 trials per block. Across both experimental sessions the participants performed a total 3456 trials of the experimental task.

In addition to the experimental blocks, they also performed a 1-back RSVP task, which was used to train the multivariate discriminant classifier. In this 1-back task, black images (Kanizsa and control) and red distractor images were displayed in an RSVP, interleaved with one another, at an ISI of 1000 ms (+- 50 ms jitter). Each image was displayed for 10 ms (see Fig. S3). Image type (Kanizsa or control) was randomized, with three randomly occurring repetitions in every ten black images. Participants were required to press a button every time a black image repeated itself while ignoring the red images. There was no relationship between stimulus type (Kanizsa or control) and image repetition, the task was purely intended to keep attention focused on the screen and the behavioral data was not analyzed further. Over both experimental sessions, participants performed 1152 trials of this task, split across 8 blocks.

### *Procedure and tasks of the masking control experiment*

Six subjects from the main experiment took part in the masking control experiment (see main text for task rationale), which consisted of one EEG session. The difference with the main experiment was that the attentional blink manipulation was not included and that the strong mask condition was replaced by a weak mask condition. The experiment was identical with respect to timing and response method. Prior to testing, each subject performed a staircase to titrate the contrast of the weak masks so that subject performance was the same as their performance in the unmasked short lag AB condition of the main experiment. After the staircase, subjects performed 9 blocks of the control experiment (1728 trials), while their EEG was recorded. In addition, they performed four blocks of the same RSVP 1-back task as in the main experiment (576 trials).

We used a double staircase procedure employing the weighted up-down method (2). Masks were presented on a white background. Contrast was adjusted by changing the intensity of the masks. One staircase started out at the minimum contrast, the other started at the maximum contrast. The staircase was updated only on trials with a Kanizsa figure: detection of the Kanizsa (hit) increased the difficulty ($S_{down}$), and indicating absence of a Kanizsa (miss) decreased the difficulty ($S_{up}$). The step size with which mask contrast was changed, was determined using the weighted rule $S_{up} * p = S_{down} * (1-p)$, in which $S_{up}$ is the upwards step size corresponding to a decrease of mask contrast, while $S_{down}$ is the downward step size corresponding to an increase of mask contrast, and p is the percentage correct onto which the staircase should converge. For example, if a subject had a hit rate of .7 in the short lag condition of the main experiment, the relationship between the two step-sizes would be $S_{up}$*.7= $S_{down}$*.3, rounding off to the nearest available values that fit given the available contrast levels (there were 20 available contrast steps between minimal and maximal).

The staircase ended after 12 reversals. The median reversal contrast for both staircases was used as starting point for mask contrast. During the experimental blocks, mask contrast was updated after each block, based on the behavioral performance in the previous block. The updating was done to keep the behavioral performance as close as possible to the unmasked short lag condition in the main experiment. Updating was rare, for four of the subjects mask contrast was adjusted only twice (on the first two blocks). One subject had mask contrast adjusted once (after the first four blocks) and one subject did not have the mask contrast adjusted at all.

### Behavioral Analysis

Where applicable, all reported statistical tests are double sided. Responses were scored as hits (Kanizsa correct) misses (Kanizsa incorrect) correct rejection (control correct) and FA (control incorrect). Hit, miss, correct rejection and false alarms. The hit rate was computed as the fraction of Kanizsa figures categorized as containing a surface, while the false alarm rate was computed as the fraction of control figures categorized as containing a surface. Behavioral performance was computed as hit rate minus false alarm rate for each of the conditions to determine how well they performed on the task. Repeated measures Analysis of Variance (ANOVAs) were used to detect main and interaction effects of the conditions.

### EEG data collection and preprocessing

EEG data was collected at 2048 Hz using a 64 channel ActiveTwo system (BioSemi, Amsterdam, the Netherlands). EEG data analysis was performed using Matlab (MathWorks, Inc., Natick, MA, USA), the EEGLAB toolbox (3) and custom written scripts to perform multivariate classification.

All data were first downsampled to 512 Hz and epoched between -500 and 1000 ms. Trials containing muscle artifacts were removed using an adapted version of the ft_artifact_zvalue muscle artifact detection function taken from the Fieldtrip toolbox (4). This function applies a frequency filter between 110 and 140 Hz and assigns a Z-value to each time point to ascertain the degree to which power values in that frequency range deviate from normality. Trials which contained Z-score outliers more than three deviations away from the absolute value of the minimum negative Z-value were discarded. Next, the data was high-pass filtered at 0.1 Hz. No low pass filtering was applied.

We did not apply baseline correction to the T2 data obtained from the main AB/mask experiment, as baseline correction introduces unwanted confounding effects on short lag versus long lag trials. There are two potential ways of performing baseline correction in our experimental design: (1) either one chooses a fixed baseline time window prior to a T2 target or (2) one applies a baseline that comes from a fixed time window prior to T1. Both approaches are problematic. The first approach only works when picking a clean baseline period before trial onset (so prior to T1), keeping the distance between baseline and T2 fixed (which would result in a different baseline time window depending on whether T2 was a short or long lag trial). However, this would have required an extremely long clean inter trial interval, which given the long trial sequence we already had was not feasible. When picking a baseline window that is closer to T2, the baseline period would overlap with T1 or with the T1-T2 lag period depending on whether it is a short or long lag T2. In that case, task-related activity during the baseline period would get introduced into the T2 period. The second approach is also problematic because the period between the baseline period and T2 onset would be different for short and long lag trials, allowing long lag trials to drift off more than short lag trials. We investigated this and confirmed that such a procedure indeed artificially boosts the short lag T2 signal when compared to the long lag T2 signal, counteracting a potential impact of the attentional blink. Instead, we therefore performed a 0.1 high-pass filter, which takes slow drifts out of the signal, similar to performing a baseline correction but does not have any of the aforementioned problems. However, we did perform baseline correction on the RSVP and T1 training sets, and on the masking testing set from the control experiment, because none of these carry the T2 specific baseline problems outlined above. When baseline correction was carried out, it was always applied on the period of -250 ms to 0 ms prior to stimulus onset.

Finally, we ran a number of control analyses to ascertain the influence of eye-blinks on the classification analysis, using both an Independent Component Analysis to remove eye-blink component as well as using a procedure to remove all trials containing eye-blinks altogether. Neither procedure had quantitative or qualitative effects on any of our classification results when compared to leaving the eye-blinks in, so we opted to retain the signal in its original form and not remove eye-blinks.

### *EEG MVPA analyses*

For each participant, we applied a backward decoding classification algorithm either using the independent RSVP data for training, using the T1 data for training, or using an 8-fold cross validation scheme (explained further down below). In all analyses, we trained a linear discriminant classifier to discriminate Kanizsa and control images using the raw EEG activity across electrodes as the features used for classification. Next, we computed classification accuracy of the classifier as the hit-rate (the fraction of Kanizsa figures that were classified as Kanizsa) minus the false alarm rate (the fraction of control figures that were classified as Kanizsa) for each subject, and for each of our conditions: T1, masked AB, unmasked AB, masked without AB, unmasked without AB. The procedure was executed for every time sample in a trial, yielding the evolution of classification accuracy over time for each of the conditions. All statistical tests were double-sided t-tests across subjects of classification accuracy (HR-FAR) against zero.

Because the classifier weights that result from the training procedure result from a backward model, they do not unambiguously reflect neural sources. They may have small amplitudes for electrodes containing the signal-of-interest, but also large amplitudes at electrodes not containing this signal, and may therefore result in both Type I and Type II errors. To mitigate this problem we obtained topographic maps by using a method recently described by Haufe et al. (5), in which the classifier weights are multiplied by the data correlation matrix (see Fig. S4, top left for classifier weights). This operation creates a correlation/class-separability map (see Fig. S4, top right) that generates interpretable neural sources for which nonzero activity is only observed at channels for which the task-related signal is both strong and highly correlated with the task, while at the same time minimizing the influence of potential artifacts.

We normalized both the weight and class/correlation separability maps across electrodes for each subject, to be able to compute topographic plots of condition averages across subjects. Fig. S4 bottom provides a direct comparison between classifier weights and the correlation/class separability map. Perhaps unsurprisingly, the effect of perceptual integration was strongly occipital in nature. Since the occipital electrodes yielded the highest classification accuracies and non-zero classifier weights (see Fig. S4, bottom), we restricted our initial analyses of the experimental conditions by using only occipital electrodes as features for classification (PO7, PO3, O1, Iz, Oz, POz, PO8, PO4, O2) to ensure that any effects we observed were not due to poor signal to noise ratio. Control analyses revealed that using all electrodes did not change any of the effects that we observed.

Next we used robust linear regression to characterize the relationship between peak accuracy of the classifier and behavioral accuracy at T1, using the 12 Kanizsa-control pairs as data points (Fig. S1). Robust linear regression guards against violations of assumptions that are required for standard regression, as well as the unwanted influence of outliers. This analysis underpins the validity of viewing peak classification accuracy as a neural measure for perceptual integration, evidenced by its strong predictive power of the behavioral response regarding surface perception.

To be able to compare the differential effect of the four T2 conditions under behavioral and neural measures of perceptual integration (HR-FAR), we entered the measurements into a large 2×2×2 ANOVA of measure (normalized behavioral / normalized neural), AB (yes/no) and masking (yes/no). The normalization step Z-scores the data, separately within the behavioral and within the neural matrix, subtracting their respective means and dividing by their respective standard deviations. It is important to realize that this normalization step does not change any of the statistics that result from the initial 2×2 (masking yes/no ×AB yes/no) ANOVA analyses. Whether entering the normalized or the non-normalized data into such an analysis, all F-statistics, p-values and all other aspects of the analysis remain the same. The only thing that changes when entering both of these normalized matrices into a large 2×2×2 ANOVA, is that any main effects of measure fall out because the measure means have been subtracted out.

The rationale for doing this is that we are not interested in main effects of measure, which is differentially affected by the signal to noise ratio for behavioral and EEG data. Rather, we want to know whether the pattern that we observe under behavioral and neural measures is the same or not, which can be obtained by looking at the interaction between measure (normalized behavioral / normalized neural) and the other factors. Whether one can regard normalized behavioral and neural measures as repeated measures of the same perceptual object, can best be understood by drawing an analogy. Let us say we want to know whether there is a differential effect of X on Y at night and during the day, but there is an overshadowing main effect on our measurements during daytime and nighttime that we are not interested in (simply because there is more light during the day, our measurement is affected by this). In such a case it would be valid to separately normalize the measurements during day and during night, removing the main day-night effect on our measurements to see if there is an interaction between factor X and moment of measurement (day/night). This is essentially what we do here by regarding the behavioral and neural measures of perceptual integration as repeated measures of the same thing, albeit with different overall averages. An interaction between that factor and the other factors shows that the underlying data pattern is not the same for the two measurements, which suggests that the experimental manipulations impact behavioral markers differently from neural markers

In the next analysis we looked at the degree to which a classifier would be able to determine class membership regarding high or low contrast on the one hand and high or low perceptual integration on the other, under masking and no masking conditions (collapsing across AB and no AB trials). Because any potential contribution of decision mechanisms was irrelevant in this analysis (subjects did not have to respond to feature contrast), we used an eight-fold training-testing algorithm. In this scheme, we first removed information about the order in which trials were acquired during the experiment by randomizing the order in which trials were stored on disk. Next, we split up the dataset into 8 equally sized subsets. Subsequently, a linear discriminant classifier was trained to discriminate between stimulus classes using 7/8 of the data, and was tested on the remaining 1/8 of the data, thereby ensuring independence of training and testing sets, repeating that scheme until all data was used for testing once, but never using the same data for training and testing in one train-test cycle. To obtain final accuracy scores we averaged across the 8 iterations. As before, the EEG activity at individual electrodes was used as features for classification and the cross-validation procedure was executed for every time sample in a trial, yielding the evolution of classification accuracy over time.

Finally, we wanted to determine the point in time at which neural signals could best explain our behavioral results. For this analysis, rather than controlling for the influence of decision mechanisms as we did initially, we now wanted to include this influence on classification accuracy. Therefore we used the T1 data as training set for the linear discriminant. Since decision mechanisms and conscious access are known to involve frontal cortex (6, 7), we went back to including all electrodes in this analysis. A control analysis confirmed that when training on T1, classification accuracy was indeed better for all electrodes when compared to restricting to occipital electrodes (Fig. S7 top and online methods, cf. S4 where the reverse is the case). All final analyses are therefore executed on T1 trained data, using all electrodes. Again, we performed 2×2 ANOVAs on the behavioral and neural data as before, and again we performed large 2×2×2 ANOVAs which includes the normalized behavioral and normalized neural data as a repeated measure (see. Fig S7, bottom for normalized responses).

To further fully characterize the moment in time at which the neural data are able to explain the behavioral data, we quantified the degree to which the neural data can serve as a model for the behavioral data using a goodness of fit on the behavioral data, taking the neural data as a reference (see main text for details). We computed this measure on the normalized neural and behavioral data, using the same rationale for normalization as before. Goodness of fit was calculated for every time point of the neural data, using a 40 ms moving average (we used a forward looking moving average to maintain liberal estimates of fit onsets). This was done separately for the masking factor, the AB factor, and for all data. The masking factor was computed by averaging accuracy scores across the AB conditions, the AB factor was computed by averaging across the masking conditions, and the total data (masking + attention + their interaction) was computed by averaging across pairs of values within each condition. Using this averaging procedure, the total number of points was kept constant for each estimation, while still being able to generate separate estimates for masking, AB and all data. However, given the uneven number of subjects (N=11), we could not create a balanced set when averaging within conditions for the total data. Therefore, the procedure was repeated 11 times for all fit types (masking, AB and total), leaving out a subject at each iteration to acquire an even number, and then averaging over the 11 resulting fits to obtain the final values.

1. Ginsburg AP (1975) Is the illusory triangle physical or imaginary? *Nature* 257(5523):219–220.

2. Kaernbach C (1991) Simple adaptive testing with the weighted up-down method. *Percept Psychophys* 49(3):227–229.

3. Delorme A, Makeig S (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Meth* 134(1):9–21.

4. Oostenveld R, Fries P, Maris E, Schoffelen JM (2011) FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological

data. *Comput Intell Neurosci* 2011:156869.

5. Haufe S, et al. (2014) On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage* 87:96–110.

6. Sergent C, Baillet S, Dehaene S (2005) Timing of the brain events underlying access to consciousness during the attentional blink. *Nat Neurosci* 8(10):1391–1400.

7. Dehaene S, Changeux JP, Naccache L, Sackur J, Sergent C (2006) Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends Cogn Sci* 10(5):204–211.

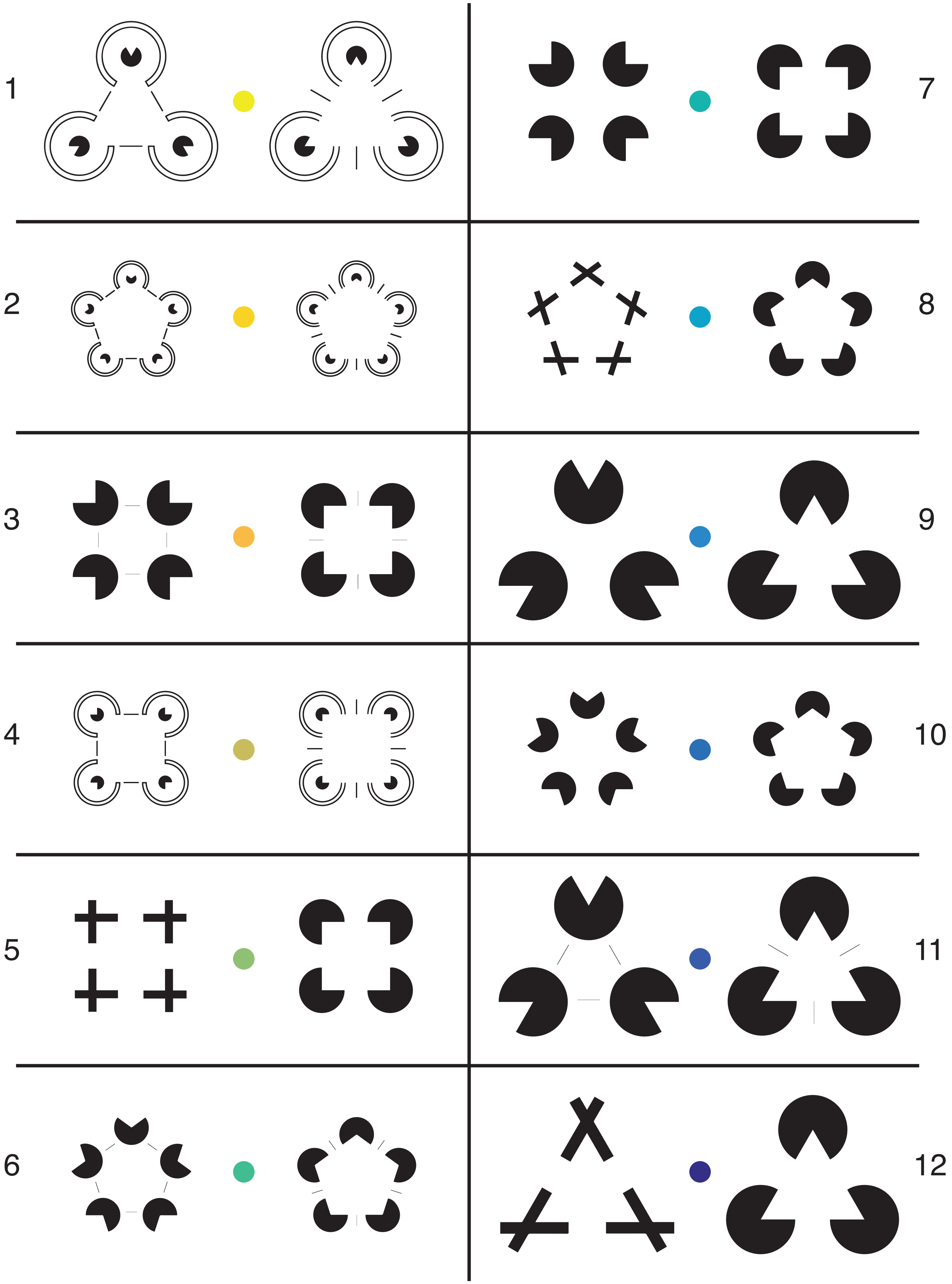
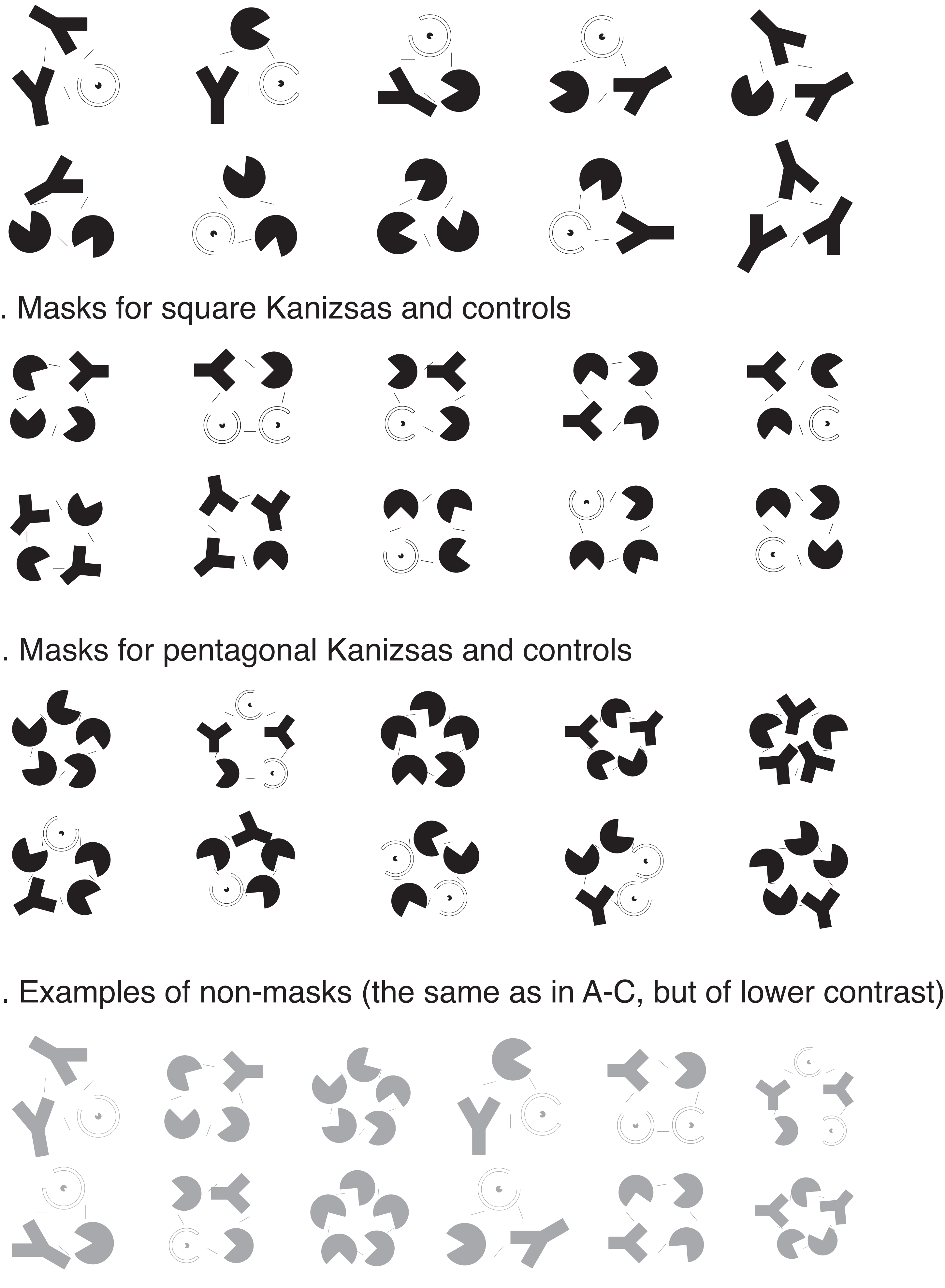


Fig. S1, the 12 Kanizsa-control pairs, see online methods for rationale behind stimulus design.



A. Masks for triangular Kanizsas and controls

B. Masks for square Kanizsas and controls

C. Masks for pentagonal Kanizsas and controls

D. Examples of non-masks (the same as in A-C, but of lower contrast)

Fig. S2, masks used during the experimental tasks.

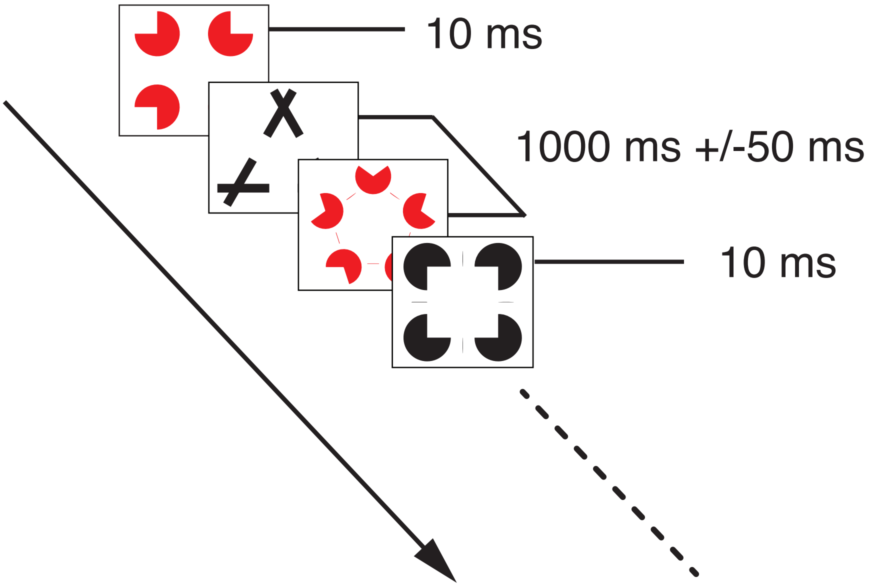


Fig S3. Independent RSVP task that was used to train the EEG classifier. Subjects were required to press a button whenever a black target would repeat (regardless of whether this target contained a Kanizsa or not), while ignoring the red distractors.

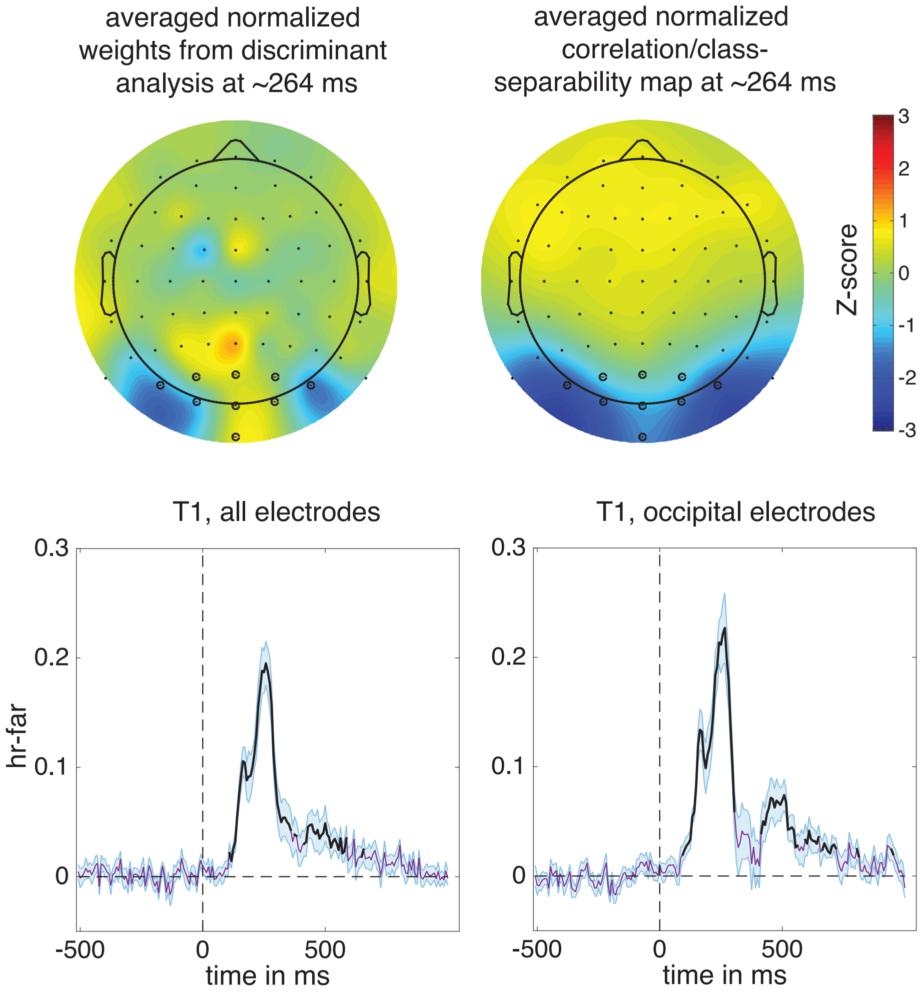


Fig. S4. Classifier weights when training on the 1-back RSVP task (top left) and the correlation class separability map (top right) at 264 ms. Since the signal is clearly occipital in nature, we compared T1 classification accuracy for all electrodes (bottom left) to classification accuracy for only the occipital elecrtrodes (PO7, PO3, O1, Iz, Oz, POz, PO8, PO4, O2, black dots in the top maps). Since the occipital electrodes result in superior performance, we used the occipital electrodes for our initial analyses (Fig. 2, Fig. 3 and Fig. 4).

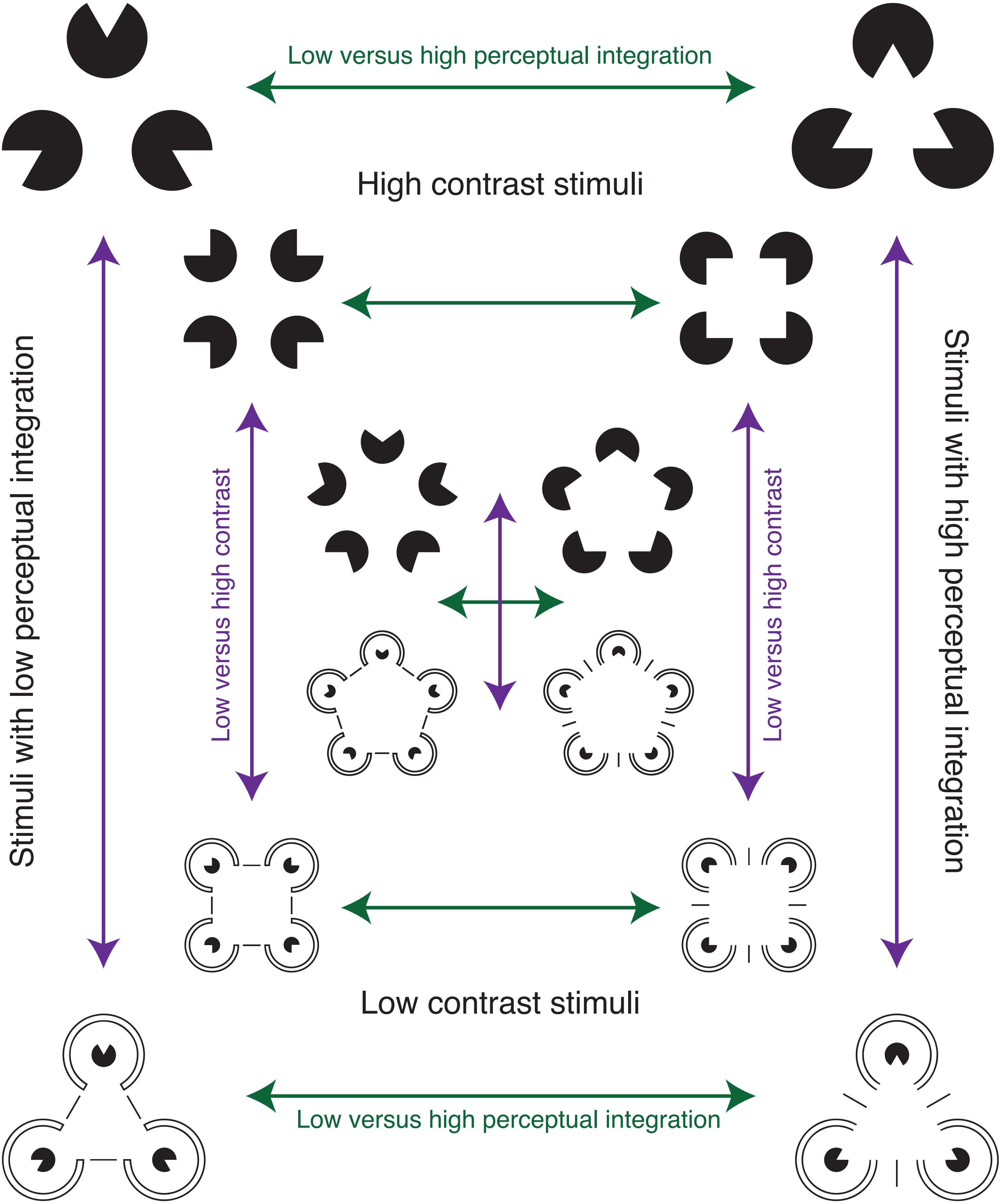

Low versus high perceptual integration
High contrast stimuli
Stimuli with low perceptual integration
Stimuli with high perceptual integration
Low versus high contrast
Low versus high contrast
Low contrast stimuli
Low versus high perceptual integration

Fig. S5. Contrast detection versus perceptual integration. Stimuli used in the masking control analysis belonging to Fig. 3. Stimulus design was such that one could compare either in the contrast dimension or in the perceptual integration dimension, while collapsing orthogonally over the other dimension.

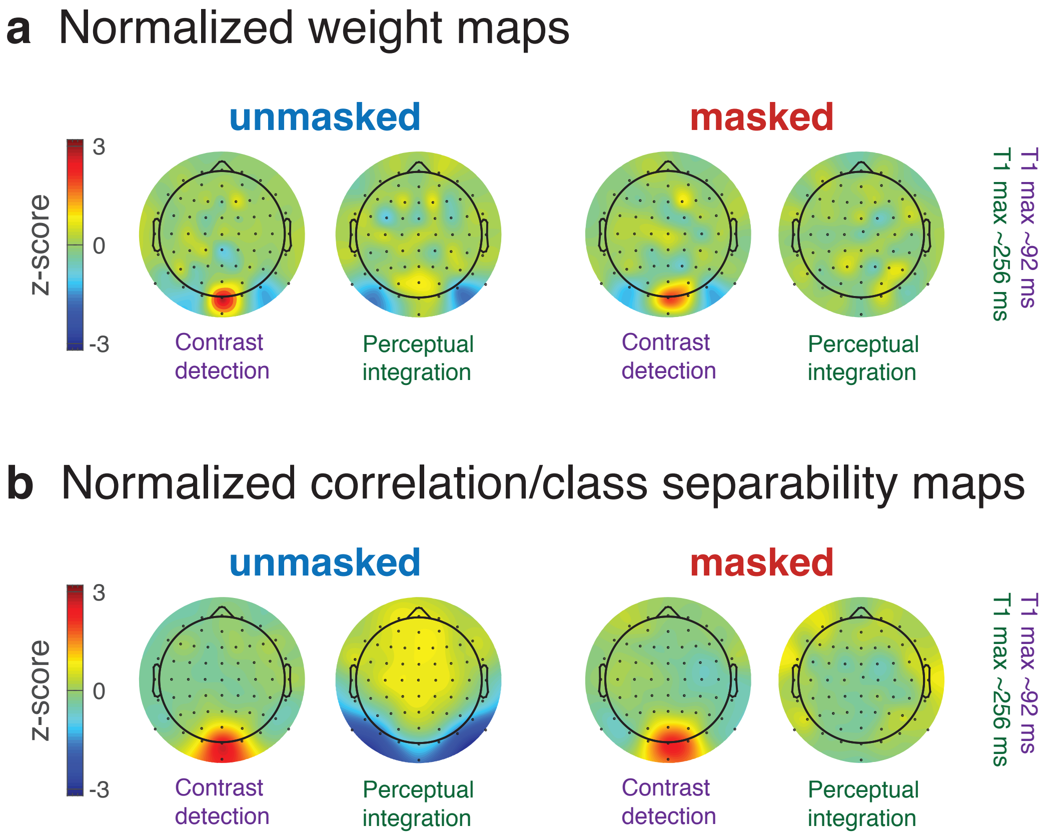


Fig S6. (a) Weight maps from classification of contrast detection and perceptual integration. (b) Corresponding correlation/class separability maps obtained by multiplying the weight maps with the correlation matrix of the latent factors.

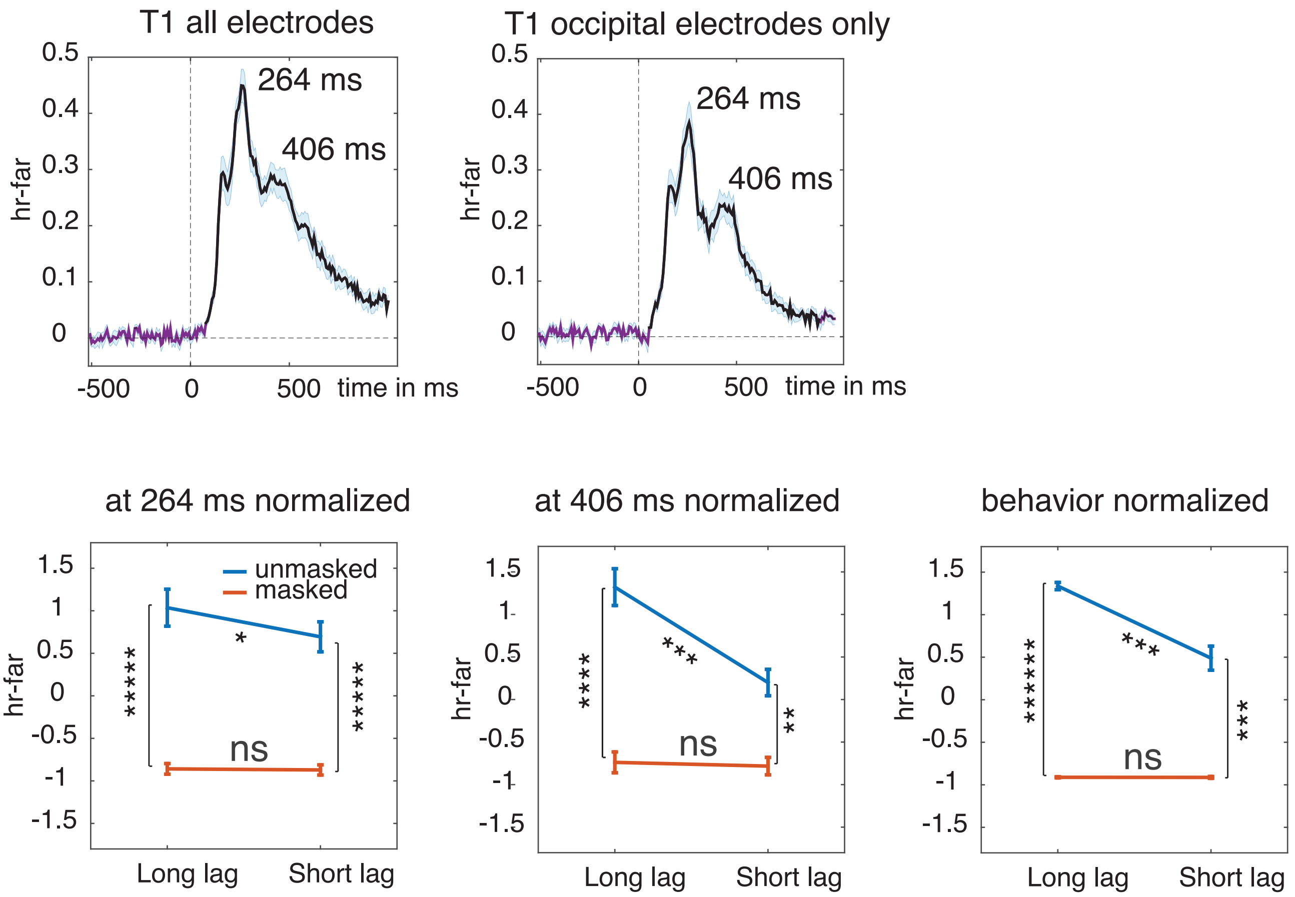


Fig. S7. Classification accuracy for all electrodes and occipital electrodes when training and testing on T1 (8-fold leave one out procedure). Given the contribution of decision mechanisms to the response, we now see a slight enhancement when using all electrodes over occipital electrodes only (top panels, cf. Fig. S4). Bottom panels shows graphs for the normalized responses when training on T1 at 264 ms, 406 ms, and when compared to behavior.