

Full Length Article

Spatio-temporal visual statistical learning in context

Dominik Garber, József Fiser*

Department of Cognitive Science, Center for Cognitive Computation, Central European University, Quellenstraße 51, 1100 Vienna, Austria

ARTICLE INFO

Keywords:

Spatio-temporal visual information
 Unconscious inference
 Perceptual biases
 Context dependent learning

ABSTRACT

Visual Statistical Learning (VSL) is classically investigated in a restricted format, either as temporal or spatial VSL, and void of any effect or bias due to context. However, in real-world environments, spatial patterns unfold over time, leading to a fundamental intertwining between spatial and temporal regularities. In addition, their interpretation is heavily influenced by contextual information through internal biases encoded at different scales. Using a novel spatio-temporal VSL setup, we explored this interdependence between time, space, and biases by moving spatially defined patterns in and out of participants' views over time in the presence or absence of occluders. First, we replicated the classical VSL results in such a mixed setup. Next, we obtained evidence that purely temporal statistics can be used for learning spatial patterns through internal inference. Finally, we found that motion-defined and occlusion-related context jointly and strongly modulated which temporal and spatial regularities were automatically learned from the same visual input. Overall, our findings expand the conceptualization of VSL from a mechanistic recorder of low-level spatial and temporal co-occurrence statistics of single visual elements to a complex interpretive process that integrates low-level spatio-temporal information with higher-level internal biases to infer the general underlying structure of the environment.

1. Introduction

Understanding how we learn representations of the significant structures in our sensory environment is one of the key challenges in uncovering how the mind works. The field of Statistical Learning (Aslin, 2017; Santolin & Saffran, 2018) addresses this question directly across various sensory modalities (Isbilen & Christiansen, 2022; Turk-Browne, 2012), most prominently in vision (Fiser & Lengyel, 2022). In visual research, the field is traditionally divided into spatial and temporal Visual Statistical Learning (sVSL and tVSL, respectively). In sVSL, the statistical regularities to be learned are found exclusively in the spatial relationships of simultaneously presented elements within a scene, with no persistent temporal structure across the sequence of scenes (Fiser & Aslin, 2001) (Fig. 1 A). In contrast, in tVSL, statistical regularities can be acquired only through temporal associations between elements of sequentially presented scenes, as each individual scene consists of a single element and thus carries no meaningful spatial information (Fiser & Aslin, 2002a; Kirkham, Slemmer, & Johnson, 2002) (Fig. 1 B).

While the separate investigation of spatial and temporal regularities is useful for gaining an initial understanding of representational learning, it stands in contrast to real-world experience, where spatial and temporal regularities are always intertwined. Indeed, several

studies have suggested that humans interpret visual input through a combined processing of spatial and temporal information (Gepshtein & Kubovy, 2000; Hochberg, 1968; Johansson, 1973; Rolls, 2012; Stone, 1998; Wallis & Rolls, 1997). Moreover, spatio-temporal regularity and stability have been consistently cited as defining features of objects and object cognition (Baillargeon, 2008; Piaget, 1954). If representational learning does, in fact, depend significantly on the interaction between spatial and temporal regularities, then relying on experimental approaches that isolate either domain risks seriously limiting our understanding of how such learning functions.

The processing of spatial and temporal statistics interacts not only with each other during learning but also with various inherent biases represented in the brain. These biases reflect expectations based on both momentary factors (Wade, Spillmann, & Swanston, 1996) and long-term knowledge (Sun & Perona, 1998), shaped by actual sensory input and extensive experience. They can influence both ongoing perception (Carlson, 1962) and learning (Liu, Dolan, Kurth-Nelson, & Behrens, 2019). Yet, very few studies have explored the interaction between ongoing visual statistical learning (VSL) and these inherent biases (Lee, Liu, & Lu, 2021).

The present work addresses the two aforementioned issues -the synergy between spatial and temporal visual statistical learning (sVSL

* Corresponding author.

E-mail address: fiserj@ceu.edu (J. Fiser).

and tVSL), and the interaction between VSL and internal biases- by focusing on three intertwined questions.

1. How primary is the strength of statistical information (and its accompanying noise) in determining the outcome of statistical learning? Existing studies in the field often limit spatial co-occurrence and temporal transitional probabilities to perfect correlations, with only anecdotal reports on learning under imperfect, noisy conditions.
2. At what level of abstraction can the interchange between sVSL and tVSL occur? For example, can humans extract spatial structures by inferring them from purely temporal evidence -without ever directly observing those spatial structures- or are spatial and temporal learning processes kept separate until relatively late stages of the knowledge representational hierarchy?
3. To what extent do inferential biases originating from higher-level knowledge -such as the general direction of stimulus motion or the presence of specific occlusions- reshape which statistical patterns are actually learned during statistical learning, even when all are equally available?

To investigate these questions, we developed a new spatio-temporal VSL (stVSL) paradigm, in which spatially defined patterns (i.e., fixed spatial arrangements of novel shapes) move in and out of the observer's view (Fig. 1C). This paradigm allows for the systematic manipulation or removal of temporal and spatial statistics under different conditions involving motion-induced or occlusion-based biases, enabling direct comparisons of learning outcomes across scenarios.

Using this paradigm across seven experiments, we obtained three main results that clarify the process of complex statistical learning in context. First, as a baseline, we confirmed that participants can learn the underlying static spatial structure of an environment even when it is embedded within the dynamic spatio-temporal input of our experimental design. Second, we found that introducing, first, a temporal structure and, next, a perceived coherent direction of movement across subsequent scenes progressively increases the strength of statistical learning of spatial co-occurrences. Third, by adding occlusion to the dynamic scenes with a motion-direction bias, we observed that the complex interplay between temporal and occlusion-based biases does

not merely enhance spatial statistical learning; it also: a) jointly alters which statistics are learned; b) enables purely temporal statistics to give rise to the learning of spatial structures; and c) goes beyond influencing choice preferences to actively boost the learning of statistics that are congruent with the configuration of biases.

Overall, these results suggest that visual statistical learning does not simply keep tab of low-level spatial and temporal co-occurrence statistics in the scenes and sums them up but acts as a sophisticated integrative process that combines low-level spatial-temporal information and various higher-level internal biases to develop a compatible internal representation of the underlying structure of the environment.

2. Rationale of the experimental paradigm

In nature, spatial and temporal statistics are typically present in a dynamically intertwined manner. For example, as animals move, their sub-parts (limbs, head, torso, etc.) change their relative positions to one another, and each part changes its appearance over time as the animal changes its global position. Yet, each part, as well as the animal as a whole, can be reliably identified across time. Moving rigid objects also provide the observer with temporal structure that can be integrated to comprehend a form which, in principle, can be described purely in spatial terms. For instance, viewing an object from different angles and integrating successive visual snapshots can help infer its three-dimensional shape. This latter problem has been explored using the "trace learning rule" (Wallis & Rolls, 1997), which relies critically on the temporal integration of successive spatial patterns to support the inference of invariant, purely spatial object structures. While such scenarios are reduced in complexity, and thus easier to control in experimental settings, they still preserve the essential elements of combined spatial and temporal statistics. The present study investigates this latter type of spatio-temporal processing of visual input, where temporal coherence is used to extract stable spatial patterns.

In the classical spatial VSL studies, the stimuli were individual NxN grid-based scenes with simple shapes in some of the grid cells, and these scenes were presented in a randomized fashion during familiarization (Fiser & Aslin, 2001, 2002b, 2005; Lee et al., 2021). The simple shapes were positioned within the scenes according to some rules of co-occurrence of shape pairs, triplets or quadruplets defined by the

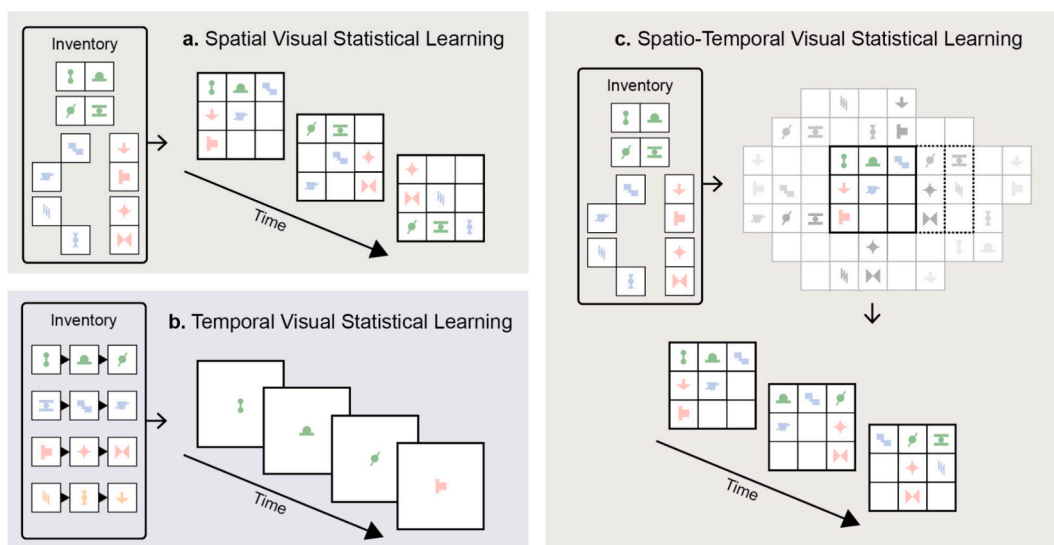


Fig. 1. - VSL paradigms: Panel (a) shows the standard spatial visual statistical learning setup. Panel (b) shows the standard temporal visual statistical learning setup. Panel (c) shows the new spatio-temporal visual statistical learning setup (stVSL). There, the visual scenes are conceptualized as part of a larger visual environment, populated with the pairs of the inventory. Participants only see a 3×3 snapshot at a given time, akin to the 3×3 scene used in spatial VSL. However, the following snapshot is given by moving the shapes under the aperture by one grid cell, making the succession of snapshots temporarily dependent on each other, as compared to them being identical and independently distributed (i.i.d.) as in the spatial VSL setup.

underlying structure of the environment. These individual scenes can be considered as randomly selected snapshots or glimpses of a small segment of a large grid-like environment populated with the spatial patterns defined by the underlying pair structures (Fig. 1 A). By generalizing this concept, we designed a new spatio-temporal VSL paradigm, in which instead of sampling random small scenes from this environment, observers viewed a small part of a large environment built on the underlying pairs structure through an $N \times N$ -sized grid-shaped aperture while the large landscape kept shifting around under the aperture in the horizontal or vertical direction with a discrete step size of one cell and paused a bit after each step (Fig. 1C). This stimulus presentation can be conceptualized as if many scenes of the classical sVLS were laid down next to each other to make up a large surface representing the environment, and observers passively but dynamically explored this environment experiencing both dynamic and occlusion-based effects beyond mere statistical co-occurrences of shape items.

Due to the periodic shifting of the environment under the aperture with a step size of one cell, observers saw not only the complete 3-by-3 tiling scenes themselves through the aperture but also, in two-thirds of the time, partial mixes of two such scenes. These cases led to partial presentations of the constituting shape pairs of the inventory when only one element of some pairs was visible. These partial presentations created noise as they introduced extra uncertainty about the true underlying structure of the input by providing evidence against the spatially fixed co-occurrence of the two elements of the pair. At the same time, this uncertainty could be counterbalanced by hypothesizing that observers make unconscious inferences based on the overall temporal coherence of the entire scene under the aperture over time due to the movement: "Although I see one half of a pair right now, this is because of the restriction of the aperture and the other half of that pair will reliably become visible after the next movement". Thus, the paradigm allows testing the intricate three-way interactions between information about temporal and spatial statistics and the effects of additional constraints such as the edge of the aperture, occlusion, and motion coherence.

Using this paradigm, we conducted seven experiments in which we systematically varied: (1) the amount of uncertainty in the input regarding the underlying spatial and temporal structure of the stimuli, and (2) the strength of the inductive bias due to perceived global motion and the occlusion of shape elements in the scene -thereby assessing the effects and interactions between these factors during statistical learning (Table 1).

3. Experiments 1a,b: Extracting the underlying spatial structures from static and dynamic scenes

The interaction between spatial and temporal statistical structures during visual learning is unclear since adding temporal changes to the display, even if they are structured, could both help and hinder the process of extracting the underlying spatial structure of the presented scenes. Specifically, the previously discussed partial presentation of the pairs during transitions -which define the underlying structure in this paradigm- may impede statistical learning, as it reduces the spatial conditional probability between elements of each pair. To clarify the nature of this interaction within the standard VSL paradigm, we tested participants in the online version of the new stVSL paradigm in Experiment 1a using the same training structure and tests as in the classic sVSL study (Fiser & Aslin, 2001), while letting the scenes unfold over time. For controlled comparison, we also directly replicated online the classical experiment of Fiser and Aslin (2001) (Experiment 1b).

3.1. Participants

40 participants gave informed consent prior to the experiments. 20 (6 female, mean age 25, $SD = 6.5$) for Experiment 1a and 20 (7 female, mean age = 24.7, $SD = 5.5$) for Experiment 1b. The sample size was based on the original study by Fiser and Aslin (2001). Participants were

Table 1

Summary of experiments in the present study. Conditions in the "Presentation Mode" column: i.i.d. – "Independent and Identically Distributed" condition, showing a sequence of individual scenes made of a 3×3 grid, with shapes in some grid cells. Scenes have abrupt onsets and offsets, with no relationship between the content of consecutive scenes. Temporal – Same as i.i.d., but consecutive scenes depict what would be visible when the 3×3 aperture shifts one grid step across an underlying larger field tiled with 3×3 small scenes. Temporal + Animation – Same as Temporal, but instead of abrupt onsets and offsets, an animation simulates the continuous shift of the underlying field. Conditions in the "Attention Check" column: + indicates that Attention check is applied during familiarization. – indicates that Attention check is not applied during familiarization.

Experiment	Purpose/Question	Presentation Mode	Attention check
1	a Proof-of-Concept of stVSL paradigm	Temporal + Animation	–
	b Online replication of basic VSL	i.i.d.	+
2	a The effect of spatial noise, temporal coherence, and perceived motion in stVSL	Temporal + Animation	–
	b	Temporal	–
	c	i.i.d.	+
3	a The limit of learning spatial structure from temporal information	Temporal + Animation + Occlusion	+
	b Interaction of global motion and occlusion biases	Temporal + Occlusion	+

recruited via prolific.co and received £ 2.5 for their contribution. For all experiments reported in this paper, all participants conducted the experiment at home on a laptop or desktop computer, and all participants had normal or corrected-to-normal vision. All experiments reported in this paper were approved by the Hungarian United Ethical Review Committee for Research in Psychology (EPKEB).

3.2. Materials

The stimuli were taken from (Fiser & Aslin, 2001) and consisted of 12 abstract black shapes on a white background. The shapes were grouped to form six pairs (two horizontal, two vertical, and two diagonal) randomly for each participant. 144 scenes were created by placing one horizontal, one vertical, and one diagonal pair in a 3×3 grid without any segmentation cues. The maximum horizontal and vertical extension of each shape was 50 % of the size of one grid cell.

Due to the nature of online experiments, the visual angle of the individual shapes was not exactly the same for all participants, as they used different devices with varying screen sizes and resolutions. However, since the image size was fixed (the 3×3 grid extended over 600×600 pixels and was centered in the middle of the computer screen) with the viewing distance varying between 40 and 80 cm, and the stimuli were simple black shapes clearly presented in the middle of white cells, the variations in the conditions of the individual setups did not significantly modulate the visibility of the perceptual input.

3.3. Procedure

Participants in both experiments first passively observed scenes during the familiarization phase before completing the test phase. For the familiarization phase, participants received only minimal instructions, stating that they should pay attention to what was happening on the screen as they would be asked simple questions about it later. The pair structure of the scenes was not mentioned.

In Experiment 1a, the 3×3 scenes moved in and out of the screen by one row or column at a time, depending on the direction of the motion,

starting from one completely visible scene in the first frame of the experiment. At each step, a segment of a new scene moved in, and a segment of the old one moved out. Since there were no segmentation cues at the scene boundaries between the abutting scenes of the underlying plane, the participants perceived the entire sequence as one continuous stream of randomly scattered shapes jointly changing the direction of movement at different “random” points in time. The change in the direction of movement was not truly “random” but always occurred at the multiples of three (after 6, 9, or 12 steps), so that at no point in the movie there were parts of more than two underlying 3×3 scenes shown in the aperture. Each movement took 0.5 s and was animated as a constant speed translation along the horizontal or vertical axis, while the image stood still for two seconds between two successive movements. The 0.5-s duration for translation was fast enough so that even if participants could not perceive the actual shapes, they could automatically track their position during the translation. No two identical pairs would ever be visible simultaneously within the aperture. Participants saw left, right, up, and down movements with occasional changes of the movement direction from one direction to one of the perpendicular directions (e.g., from horizontal move to either up or down) but never a complete reversal of direction (e.g., from up to down). Overall, all participants saw left, right, up, and down movements for the same number of steps.

In Experiment 1b, participants saw the scenes sequentially, each for two seconds with a one-second inter-trial interval, and the order of scenes was randomly chosen for each participant. In this “Independent and Identically Distributed” (i.i.d) condition, there is no link between the content of two consecutive scenes. Since the static input in Experiment 1b could be less engaging than the dynamic input in Experiment 1a, a simple attention-check feature was included in Experiment 1b to ensure that participants remained engaged with the task. In an attention-check trial, text appeared in the central cell of the grid, prompting participants to press the spacebar. Simultaneously, five black squares appeared in randomly chosen cells of the grid. The attention check disappeared every 2 s and reappeared after 0.5 s. The number of times the attention check was shown before the space bar was pressed was recorded as the indicator of the participant’s attention level.

Individual scenes were seen for a longer duration in Experiment 1a than in 1b since they moved in and out of sight over several steps. Therefore, we used only half of the original 144 scenes (balanced for pair frequency and co-occurrence) in Experiment 1a. The amount of exposure was still not entirely identical in the two experiments as the familiarization phase took nine minutes vs. seven minutes and 12 s in Experiment 1a and Experiment 1b, respectively. In addition to the difference in the number of scenes and exposure durations, two other aspects prevented a direct comparison between the two experiments: the full vs. full+partial presentation of each scene and the added effect of motion in Experiment 1a. However, all these differences were inconsequential as the goal of the study was not to compare the two setups quantitatively but to test whether participants can implicitly learn the pair structures in the dynamic stVSL setup.

The test phase was identical for both experiments. It consisted of 36 2-alternative forced choice (2AFC) trials. In each trial, participants saw a real pair and a foil pair after each other (randomized order, two-second presentation, and one-second inter-stimulus-interval) and indicated which of the two was more familiar based on the familiarization phase by pressing “1” or “2” on the keyboard. Overall, six foil pairs, two horizontal, two vertical, and two diagonal ones, were created by recombining shapes from different pairs of the familiarization phase. Each real pair was tested once with each foil pair. After the test phase, participants answered a series of open questions to assess their explicit knowledge of the pair structure and their previous experience with similar experiments (see Supplementary Materials for details).

Beyond the standard frequentist statistics analyses, we used Bayes Factors (BF) in this paper to assess the strength of the observed effects. These BFs were calculated by using the *BayesFactor* R package (Rouder,

Morey, Speckman, & Province, 2012). Following a conservative approach, we counted our results as significant if our criteria for both p -values (<0.05) and BF (>3) were met.

3.4. Results and discussion

For both experiments, two participants were excluded for gaining verbalizable explicit knowledge of the pairs by the end of the practice session, since we were interested in implicit learning of the structures. The remaining participants in both experiments performed significantly above the chance level of 50 % (Experiment 1a: $M(SE) = 57.7(2.4)$, $t(17) = 3.18$, $p = 0.005$, $d = 0.75$, $BF = 8.8$. Experiment 1b: $M(SE) = 56.5(2.5)$, $t(17) = 2.60$, $p = 0.019$, $d = 0.61$, $BF = 3.2$. see Fig. 2).

These results show that participants successfully learned the spatial structure in both the classic i.i.d. and the novel spatio-temporal VSL setup, confirming that despite the increased noise due to the partial presentation of scenes and the effect of motion, participants could implicitly learn the pair structures in the dynamic stVSL setup. This confirmatory result is significant because the general finding in the literature is that statistical learning is very sensitive to the conditional probabilities between elements within a chunk: even small deviations from $\text{Prob}(1.0)$ can severely affect learning. As Experiment 1b was a direct online replication of the previously lab-based spatial VSL experiment by (Fiser & Aslin, 2001), the close similarity between the results of Experiment 1b and that of the original study also suggests that the sVSL paradigm can be successfully transferred to an online environment for data collection.

4. Experiments 2a-c: The role of temporal coherence and perceived motion in learning spatial visual structures

After demonstrating in Experiment 1a that learning is possible in the spatio-temporal VSL (stVSL) setup, we asked how the different aspects of the input information contributed to learning in this paradigm. In Experiment 2a, we introduced pairs with different levels of spatial noise (quantified by the number of partial presentations) to the stVSL paradigm to test whether the strength of learning is a simple function of the amount of this type of noise. In Experiment 2b, we removed the animation of motion while leaving the temporal coherence of subsequent frames intact to assess the effect of temporal coherence across scenes in the absence of obvious motion cues. In Experiment 2c, we eliminated both motion cues and temporal coherence to provide experimental evidence for their joint effect on learning spatial patterns.

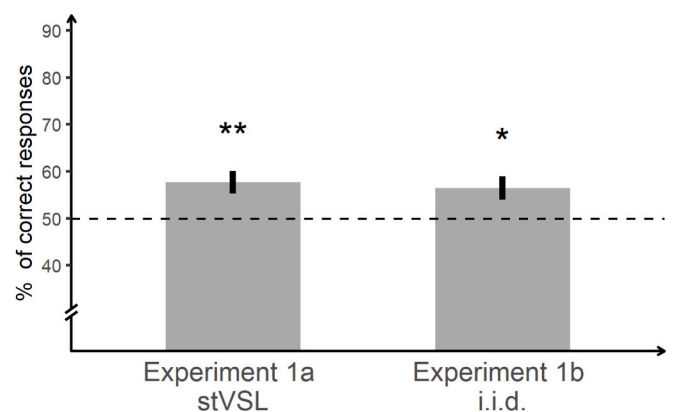


Fig. 2. - Results of Experiments 1a and 1b. The y-axis represents the participants’ mean performance on the 2-alternative forced choice (2AFC) trials, used as the measure of learning of pairs embedded in the familiarization stream. Error bars represent the standard error. The dotted line indicates the chance level of 50 %. Stars represent the significance of the difference from chance. * $p < 0.05$; ** $p < 0.01$.

4.1. Participants

267 participants gave informed consent prior to the experiments. 88 (39 female, mean age 26.9, SD = 8) for Experiment 2a, 89 (31 female, mean age 33.4, SD = 10.9) for Experiment 2b, and 90 (30 female, mean age 26.2, SD = 8.9) for Experiment 2c. The sample size was chosen to achieve a power of about 80 % for three parallel comparisons (i.e., $\alpha = 0.166$) assuming medium effect sizes (Cohen's $d = 0.5$) and rounded up to account for expected exclusions. Participants were recruited via [prolific.co](https://www.prolific.co) and received £2.5.

In Experiment 2a, three participants were excluded from the analysis due to response bias, and another 15 participants were removed for acquiring explicit knowledge of the structure of the task. *Response bias* was defined as the proportion with which participants used one of the two response keys ("1" and "2"), and participants who were 2.5 SD away from the mean were excluded. In Experiment 2b, six participants were removed from the analysis due to response bias, one participant was removed due to failing the attention checks (response time to attention check >3 SD), and 10 participants were removed for acquiring explicit knowledge of the structure of the task. In Experiment 2c, two participants were removed from the analysis due to response bias, two additional participants due to failing attention checks, and 10 participants were due to acquiring explicit knowledge of the structure of the task. For data of explicit participants, see the Supplementary materials.

4.2. Materials and procedure

The materials were identical to those in Experiment 1a. The general procedure of Experiment 2a was identical to Experiment 1a, with the exception of the specific movement directions. Participants no longer saw balanced movement enforced by the same number of shifts to all directions, but instead, they were randomly assigned to one of two conditions, having more horizontal or more vertical movement on average. In the horizontal condition, 75 % of the movements were along the horizontal axis, with equal amounts of movements to the left and to the right, and 25 % of the movements were along the vertical axis, with equal amounts of movements up and down. This was implemented by changing the direction of movement after 9, 12, 15, or 18 steps for horizontal movement and after 3 or 6 steps for vertical movement. As a result, horizontal pairs had more partial presentations than vertical pairs, as they were shown partially only during horizontal movement. In the vertical condition, the pattern of movement and, consequently, the occurrence of partial presentations were reversed. For the sake of clarity and simplicity, we will use the term *parallel* pairs for pairs aligned with the predominant movement (e.g., horizontal pair in the horizontal condition) and *orthogonal* pairs for pairs with orientation orthogonal to the predominant movement (e.g., vertical pairs in the horizontal condition) to refer to both conditions simultaneously.

Having a design with parallel and orthogonal pairs resulted in differences in the amount of partial presentation, that is, in noise across the three types of pairs of the inventory. In both the horizontal and vertical conditions, the diagonal pairs had the overall highest number of partial presentations since they were shown partially during both horizontal and vertical movement. The conditional probabilities of the shapes of one pair (the probability of one shape of a pair being visible when the other shape is visible) were 0.6 for diagonal, 0.75 for parallel, and 0.0916 for orthogonal pairs.

The only difference between Experiments 2a and 2b was that the animated motion seen in Experiment 2a, showing a smooth transition of the shapes during scene change was removed in Experiment 2b (Fig. 3). As a result, the temporal coherence across scenes was identical in 2a and 2b, i.e., the static pictures at the pauses followed the same sequence in the two experiments. However, the strong perceptual cue to temporal coherence given by the smooth motion pattern was present only in Experiment 2a, allowing a separation of the effects of the spatio-temporal co-occurrence statistics itself and the effect of perceived

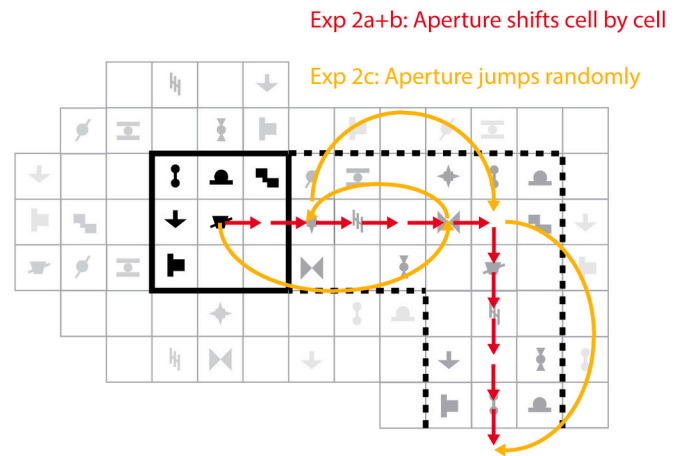


Fig. 3. Presentation Modes in Experiments 2a-c. This graphic visualizes the temporal relationship of subsequent spatial patterns in Experiments 2a-c. In Experiments 2a and 2b, visualized in red, the visual aperture shifts by one cell at a time, leading to a sequence of temporally coherent scenes. In Experiment 2c, visualized in yellow, the visual aperture visits the same overall parts of the environment, leading to the same number of partial presentations of pairs. However, as visualized with the yellow arrows, the order of visual scenes is random and, therefore, not temporally coherent. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

coherent motion of the scene. Since the statically presented inputs in Experiment 2b could be less engaging, the same attention check was included during the familiarization phase as in Experiment 1b.

Compared to Experiment 2b, the additional difference in Experiment 2c was the removal of temporal coherence of the stimuli presentation (Fig. 3). Participants still saw the same still images as in Experiments 2a and 2b between movements. However, instead of moving them in and out of the grid by one cell at a time, the scenes were temporally shuffled and shown in random order. This manipulation achieves exactly the same level of spatial noise in the visual input as in Experiments 2a and 2b but with an i.i.d. presentation instead of a temporally structured one. As before, the same attention check was included during the familiarization phase of Experiment 2c as in Experiment 1b.

4.3. Results

We performed one-sample t -tests for each of the three experiments both individually for the different pair types (parallel, orthogonal, and diagonal) and collapsed over pair types. In the collapsed analysis, significant learning occurred in all three experiments (Exp 2a $M(SE) = 56.7$ (1.1), $t(69) = 5.96$, $p < 0.001$, $d = 0.71$, $BF = 136,989$; Exp 2b $M(SE) = 53.4$ (1.2), $t(71) = 2.84$, $p = 0.012$, $d = 0.33$, $BF = 5.2$; Exp 2c $M(SE) = 52.3$ (1.2), $t(75) = 2.18$, $p = 0.032$, $d = 0.25$, $BF = 1.2$). All significance tests reported throughout the present study are corrected for multiple comparisons as appropriate using the Holm-Bonferroni method (Holm, 1979).

The analysis of data in Experiment 2a for different pair types showed that participants' performance for the orthogonal ($M(SE) = 57.1$ (2.3), $t(69) = 3.17$, $p = 0.018$, $d = 0.38$, $BF = 12.2$) and the diagonal ($M(SE) = 58.7$ (2.0), $t(69) = 4.4$, $p < 0.001$, $d = 0.53$, $BF = 535$) pairs were significantly different from chance, while the performance for the parallel pairs was not: $M(SE) = 54.4$ (2.0), $t(69) = 2.2$, $p = 0.157$, $d = 0.26$, $BF = 1.2$. In Experiment 2b, the same separate analysis showed that the performance for all of the three pair types failed to deviate significantly from chance: parallel ($M(SE) = 52.9$ (2.2), $t(71) = 1.3$, $p = 0.006$, $d = 0.15$, $BF = 0.29$), orthogonal ($M(SE) = 51.2$ (2.1), $t(71) = 0.5$, $p = 0.999$, $d = 0.06$, $BF = 0.15$), diagonal ($M(SE) = 56.0$ (2.2), $t(71) = 2.7$, $p = 0.064$, $d = 0.32$, $BF = 3.5$). In Experiment 2c, performance for neither of

the three pair types was significantly different from chance: parallel ($M(SE) = 48.9(2.0)$, $t(75) = -0.55$, $p = 0.999$, $d = 0.06$, $BF = 0.146$), orthogonal ($M(SE) = 53.4(1.95)$, $t(75) = 1.75$, $p = 0.338$, $d = 0.2$, $BF = 0.54$), diagonal ($M(SE) = 55.5(2.3)$, $t(75) = 2.4$, $p = 0.112$, $d = 0.276$, $BF = 1.86$) (see Fig. 4).

For further analysis, we entered Experiment 2a, 2b and 2c into one 3×3 mixed-ANOVA with pair type (parallel, orthogonal, diagonal) as a within-subject factor and experiment (2a, 2b, 2c) as a between-subject factor. The results showed a significant main effect of pair type ($F(2, 430) = 3.74$, $p = 0.025$, $\eta^2 = 0.012$) and of experiment ($F(2, 215) = 3.22$, $p = 0.042$, $\eta^2 = 0.012$) but no significant interaction ($F(4, 430) = 0.62$, $p = 0.649$, $\eta^2 = 0.004$). Tukey's SHD post-hoc tests showed that the diagonal pairs were learned significantly better than the parallel

pairs ($p = 0.020$, $BF = 3.6$) and that learning in Experiment 2c (i.i.d. condition) was significantly worse than in Experiment 2a ($p = 0.043$, $BF = 3.2$) (see Fig. 4).

4.4. Discussion

The overall pattern of the results suggests first that while participants learned pairs in all three experiments, there was a significant modulation in the amount of learning from the temporally coherent condition (Exp. 2a) to the condition without perceived motion (Exp. 2b) and to the setup without temporal coherence (Exp. 2c). This gradual modularity effect is direct experimental evidence that participants use global temporal regularities to learn local spatial patterns from passive exposure and it furthermore supports the idea that the perceived motion is an important general cue to temporal regularity.

The second surprising observation is that beyond showing a significant general sensitivity to motion and temporal coherence, the amount of learning did not directly reflect the local spatial statistical structure of the stimuli. Participants learned the diagonal pairs better than either the parallel pairs or orthogonal pairs despite their lower conditional probability (higher statistical noise) and despite being less aligned with the direction of motion. This finding suggests the existence of substantial additional factors influencing this kind of implicit statistical learning beyond the co-occurrence statistics of elements.

5. Experiment 3a: Full generalization from purely temporal statistics to static spatial structures

The previous set of experiments established that implicit learning of spatial statistics present in the sensory input is profoundly affected by factors beyond the spatial co-occurrences of elements, such as temporal coherence and perceived motion, yet it did not clarify the extent of these effects. In Experiment 3a, we focused on the first of the two findings of Experiments 2a-c and directly compared the effect of spatially and temporally conveyed information on learning spatially defined static patterns. In particular, we modified the experimental design and tested whether participants would be able to learn static spatial structures of the environment when they never see these spatial patterns during familiarization, but they need to infer their existence implicitly from dynamically presented partial information. In addition, we explored the nature of the interaction between temporal and spatial statistics and the general inductive biases that emerge from detecting global motion and occlusion.

5.1. Participants

132 (60 female, mean age 27.1, $SD = 11.1$) participants gave informed consent prior to the experiments. The sample size calculation built on the previous experiments but assumed a lower alpha level of 0.005 to account for the higher number of multiple comparisons (more tests in this experiment) and assumed a higher rate of exclusions. Participants were recruited via prolific. co and received £ 2.3.

Prior to analysis, two participants were removed for response bias (bias for one of the two responses buttons $>2 SD$), two participants were excluded for failing the attention check (having the attention check message appear at least 10 times over all three instances of the attention check), and 18 participants were removed for acquiring explicit knowledge of the structure of the task (for data of explicit learners see Supplementary Materials).

5.2. Materials

This experiment used the same materials as the previous experiments, but with one important difference. Throughout the experiment, a rectangular static occluder strip image with $1/f$ -noise content and 50 % of the width and height of the grid cells was placed over the middle row

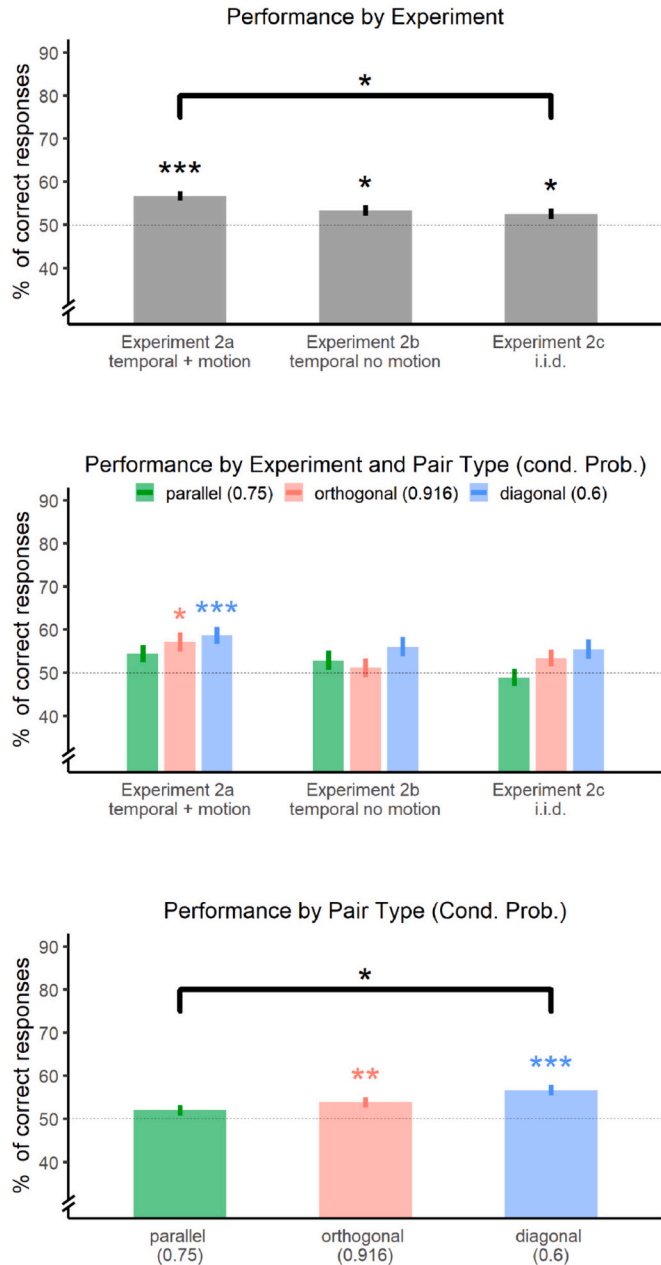


Fig. 4. – Results of Experiments 2a-c. The y-axis represents the performance on 2AFC trials. Error bars represent the standard error. The dotted line indicates the chance level. Stars represent the significance of statistical tests: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.005$. Top panel: results of pair type by experiment. Middle panel: main effects of experiment averaged over pair types. Lower panel: main effects of pair type averaged over experiments.

or column of the grid. This strip covered the central part of the cells in the middle row (or column) so that the shape present in these cells was not visible during the static sections of the presentation when the movement was paused (Fig. 5).

5.3. Procedure

The general procedure was built on Experiment 1a but -beyond the occluder- introduced one more change to the familiarization phase to achieve the desired temporal and spatial presentation modes of pairs. Every participant in Experiment 3a saw movement along only one orientation, i.e., only horizontal movement (left, right) or vertical movement (up, down). Participants first saw movement in one direction for 60 steps, then movement in the opposite direction for 60 steps, then 48 steps in the first movement direction again and finally, 48 steps in the second movement direction. The static occluder overlaid over the three central grid cells was always perpendicular to the movement direction. At each movement direction change, an attention check appeared, identical to the one used in Experiment 1b.

The combined result of this change and the superimposed occluder was a more extreme separation between the underlying pairs in terms of their defining type of statistics. Orthogonal pairs (pairs perpendicular to the movement direction) were only presented spatially and never temporally; that is, the two shapes of these pairs always appeared only together, never alone, and they never followed each other temporarily in

the same grid cell. In contrast, parallel pairs (pairs aligned with the movement direction) were never visible next to each other in the still images between movements (2 s) and, as mentioned in Exp 1, were essentially not visible next to each other during the short movement period (0.5 s). However, the shapes of the parallel pairs had perfect temporal coherence as they always followed each other in the same position of the grid. Diagonal pairs represented a noisier version of the parallel pairs as they had the same type of temporal coherence and essential lack of spatial co-occurrence as parallel pairs, but instead of following each other temporarily in the same grid cell, the two shapes of the pairs also had a spatial offset.

In the test phase following the nine-minute familiarization phase, participants completed three types of 2AFC tests.

The *Standard Learning Trials* test was identical to those in previous experiments, pitting against each other real vs. foil pair tests in each trial. In these trials, the orientations of the two alternatives within a trial were always the same.

In the *Spatial Learning Trials* test, the two options of a trial presented the same shapes of one given real pair but the shapes were presented once in their correct spatial arrangement (e.g., horizontal) and once in an opposite arrangement (vertical). Diagonal pairs were tested against themselves in a different diagonal arrangement. This test measured the participants' additional knowledge of the spatial structure of the pairs beyond the co-associations of the shapes of the pair.

In the *Bias Trials* test, the same logic was applied as in the *Spatial Learning Trials*, but a foil pair was used for each trial instead of real pairs. In these trials, there was no correct orientation and, thus, no correct answer since these pairs were not seen during the familiarization phase. Therefore, the test was suitable for assessing the participants' overall bias in choosing either the horizontal or the vertical orientation, independent of any knowledge of pair structures.

Participants first completed the *Spatial Learning Trials* and *Bias Trials* intermixed, followed by the *Standard Learning Trials*. This order was chosen to ensure minimal interference between the test trials.

5.4. Results

In the *Standard Learning Trials*, participants' performance with parallel pairs was significantly above chance: $M(SE) = 58.2(2.4)$, $t(109) = 3.4$, $p = 0.004$, $d = 0.33$, $BF = 23.5$. The performance with orthogonal ($M(SE) = 48.6(2.6)$, $t(109) = -0.52$, $p = 0.692$, $d = 0.05$, $BF = 0.12$) and diagonal ($M(SE) = 54.5(2.3)$, $t(109) = 1.99$, $p = 0.148$, $d = 0.19$, $BF = 0.70$) pairs was not different from chance. These results indicate that when the orientation of the true and foil pairs in the test trial was not different, participants showed evidence of learning only the "parallel" pairs -that is, the pairs oriented parallel to the direction of motion and perpendicular to the occluder, for which evidence was provided by temporal transitions. Meanwhile, participants failed to show learning the "orthogonal" and "diagonal" pair structures for which no evidence was provided by direct temporal transition but only by direct or more indirect spatial co-occurrence.

In the *Spatial Learning Trials*, performance with parallel pairs was significantly above chance: $M(SE) = 65.5(2.7)$, $t(109) = 5.68$, $p \leq 0.001$, $d = 0.54$, $BF = 1.0 \times 10^5$. The performance with orthogonal pairs was significantly below chance ($M(SE) = 36.8(2.6)$, $t(109) = -5.2$, $p \leq 0.001$, $d = 0.49$, $BF = 1.1 \times 10^4$), while the performance with diagonal pairs was not different from chance ($M(SE) = 52.5(2.6)$, $t(109) = 0.95$, $p = 0.692$, $d = 0.09$, $BF = 0.16$). These results confirm that when the shapes in the two test scenes were the same and thus the identity of the shapes in the pairs of the two alternative choices could not help the decision, participants' choices were highly influenced by the direction of the general motion of the patterns within the aperture. Importantly, this influence was a general motion bias rather than an integrated (shape identity + pair orientation) knowledge since participants erred significantly against the correct orientation, with pairs having their true orientation orthogonal to the motion direction.

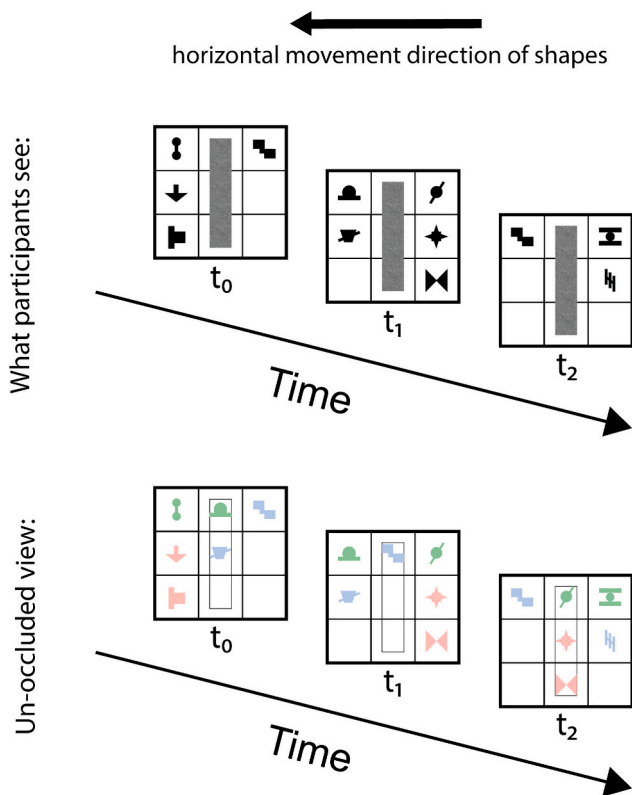


Fig. 5. – Familiarization phase of Experiment 3a + b. Experiment 3a + b introduce two key changes to the basic stVSL setup. (1) Participants see only either horizontal or vertical movement (counter-balanced between groups). (2) An occluder is overlaid over the three central grid cells perpendicular to the movement direction. The combined effect of these two changes is that pairs aligned with the movement direction are now presented only temporally, while pairs perpendicular to the movement direction are only presented spatially. The graphics visualize the condition using only horizontal motion. The top panel directly shows what participants see in the experiment. The bottom panel is only for illustration purposes as it highlights the underlying structure by color coding pairs and making the occluder transparent.

Since in the *Bias Trials*, there were no correct response options, the data for all parallel and orthogonal foil pairs was considered together and scored for choosing the parallel or orthogonal option. The trials for diagonal pairs were not considered in this analysis. For every participant, the proportion of parallel choices was expressed in percent and then subtracted by 50, to get a measure of bias away from a chance level of zero. Positive bias suggests more parallel choices and negative bias suggests more orthogonal choices. One-sample *t*-tests showed that participants chose the parallel options significantly more often: $M(SE) = 8.0(2.0)$, $t(109) = 3.99$, $p < 0.001$, $d = 0.38$, $BF = 149$.

For further analysis, the data of the *Standard Learning Trials* and *Spatial Learning Trials* were entered into one 2×3 mixed-ANOVA with test type (*Standard Learning Trials*, *Spatial Learning Trials*) and pair type (parallel, orthogonal, diagonal) as within-subject factors. The results showed a significant main effect of pair type ($F(2, 545) = 25.6$, $p < 0.001$, $\eta^2 = 0.19$) and a significant interaction ($F(4, 545) = 7.4$, $p < 0.001$, $\eta^2 = 0.06$). The main effect of test type was not significant, $F(1, 545) = 1.28$, $p = 0.26$, $\eta^2 = 0.01$. Tukey's SHD post-hoc tests showed significantly higher performance for the parallel pairs than the orthogonal pairs in the *Standard Learning Trials* ($p = 0.025$), and in the *Spatial Learning Trials* ($p < 0.001$). The diagonal pairs showed significantly higher performance than the orthogonal pairs only in the *Spatial Learning Trials* ($p < 0.001$). The parallel pair showed significantly higher performance than the diagonal pair only in the *Spatial Learning Trials* ($p = 0.001$).

To test whether the high deviation from chance in the *Spatial Learning Trials* was based solely on the bias also measured in the *Bias Trials* or if it additionally included knowledge about the orientation of the specific pairs, a direct comparison of both measures was conducted. For this purpose, the results for the parallel pairs and the orthogonal pairs in the *Spatial Learning Trials* were separately transformed into a measure of deviation from chance, as described for the *Bias Trials*. Paired *t*-tests showed that the deviation from chance for the parallel pairs ($t(109) = -2.60$, $p = 0.021$, $d = 0.30$, $BF = 2.6$) but not the orthogonal pairs ($t(109) = -1.92$, $p = 0.058$, $d = 0.22$, $BF = 0.62$) was significantly higher than the bias measured in the *Bias Trials*. This suggests that participants have knowledge about the actual orientation of pairs they have learned. The results for all test types are visualized at the top in Fig. 6.

5.5. Discussion

In this setup, the participants were able to learn only the temporally presented parallel pairs and not the spatially presented perpendicular ones. The results of the *Bias Trials* suggest that participants had a dominant overall bias in the test trials to choose the pair with an orientation that aligned with the movement direction perceived during the training, regardless of what the true orientation of the pair was originally. Nevertheless, participants did have knowledge about the actual orientation of pairs they had learned and the effect of this knowledge was detectable in the difference of their responses between parallel and orthogonal pairs. While the exact nature of this interaction between learning features of specific pairs, the perceived movement direction, and the observed overall bias remain unclear, these results highlight that statistical learning goes beyond simple counting of lower level co-occurrence statistics and also incorporates top down effects.

6. Experiment 3b: Dominance of perceived overall movement over learned statistics

Was the effect in Experiment 3a driven predominantly by the underlying spatio-temporal co-occurrence statistics, or was it also influenced by broader features of the stimulus presentation, such as the perceived overall motion of the scenes or the static structure of the occluder? Experiment 3b addressed this question by using the same spatio-temporal structure as Experiment 3a, retaining the occluder but

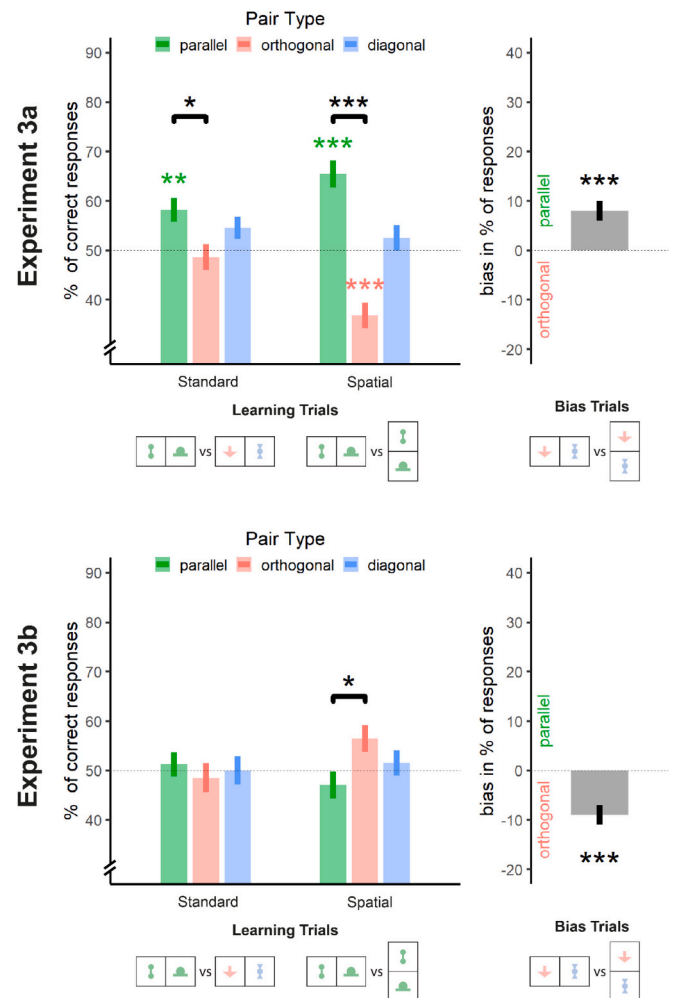


Fig. 6. Results of Experiments 3a and 3b. The y-axis represents the participants' mean performance on 2AFC trials. Error bars represent the standard error. The dotted line indicates the chance level of 50 % or 0 %. Stars represent the significance of the difference from chance. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.005$. The *Standard Learning Trials* was a standard learning test using one real pair from the training phase and one foil pair created by combining shapes of two real pairs. It measures learning of item co-occurrence. The *Spatial Learning Trials* test showed the same real pair twice. Once in its correct orientation and once rotated by 90° . It measures learning of the spatial arrangement of learned pairs. The *Bias Trials* test showed the same foil pair twice. Once horizontally and once vertically. There is no correct response, and it measures bias for one of the orientations.

removing the movement animation.

6.1. Participants

117 (52 female, mean age 30.4, $SD = 11.9$) participants gave informed consent prior to the experiments. Participants were recruited via prolific.co and received £ 2.3.

Prior to the analysis, two participants were removed for response bias (bias for one of the two response buttons >2 SD), six participants were excluded for failing the attention check (having the attention check message appear at least 10 times over all three instances of the attention check), and 7 participants were removed for acquiring explicit knowledge of the structure of the task (for data of explicit learners see Supplementary Materials).

6.2. Materials and procedure

The material was the same in Experiment 3b as in Experiment 3a. The procedure was also identical between the experiments in every way apart from the movement animation. While in Experiment 3a, the stimulus displays drifted, and hence the shapes moved in and out of the visible aperture with a 0.5-s animated movement between the 2-s steady sections, in Experiment 3b, the animation was replaced by a 0.5-s blank screen. Note that the spatial and temporal structure was not altered by this manipulation. Parallel pairs were still presented only temporally, and orthogonal pairs were still presented only spatially.

6.3. Results

6.3.1. Standard learning trials

One-sample *t*-tests showed that in the *Standard Learning Trials* the performance for all of the pairs was not different from chance: parallel ($M(SE) = 51.3(2.5)$, $t(99) = 0.51$, $p = 0.99$, $d = 0.05$, $BF = 0.125$), orthogonal ($M(SE) = 48.5(2.9)$, $t(99) = -0.51$, $p = 1.00$, $d = 0.05$, $BF = 0.126$), diagonal ($M(SE) = 50.0(2.9)$, $t(99) = 0.00$, $p = 1.00$, $d = 0.00$, $BF = 0.11$). Overall, these results suggest that participants did not reliably learn any pairs in this experiment.

6.3.2. Spatial learning trials

One-sample *t*-tests showed that in the spatial-real test, the performance for all of the pairs was not different from chance: parallel ($M(SE) = 47.0(2.7)$, $t(99) = -1.09$, $p = 1.00$, $d = 0.11$, $BF = 0.198$), orthogonal ($M(SE) = 56.5(2.7)$, $t(99) = 2.41$, $p = 0.107$, $d = 0.24$, $BF = 1.7$), diagonal ($M(SE) = 51.5(2.5)$, $t(99) = 0.59$, $p = 1.00$, $d = 0.06$, $BF = 0.13$).

6.3.3. Bias trials

Data for this test was converted to a measure of bias away from chance, as in Experiment 3a. One-sample *t*-tests showed that participants chose the orthogonal options significantly more often: $M(SE) = -9.0(2.0)$, $t(99) = -4.55$, $p < 0.001$, $d = 0.46$, $BF = 1051$.

To test whether the significant difference between parallel and orthogonal pairs in the *Spatial Learning Trials*, found in Experiment 3a along the direction of the overall bias (results of *Bias Trials*), is also present here, a paired *t*-test was performed. The frequentist analysis of the results showed significantly higher performance for the orthogonal pair trials ($t(99) = -2.16$, $p = 0.033$, $d = 0.35$), but Bayesian evidence did not provide strong support for this conclusion ($BF = 1.03$). To test whether the results of the *Spatial Learning Trials* were in line with the bias measured in the *Bias Trials*, a direct comparison of both measures was conducted. For this purpose, the results for the parallel pairs and the orthogonal pairs in the *Spatial Learning Trials* were separately transformed into a measure of deviation from chance, as described for the *Bias Trials*. Paired *t*-tests showed no deviation from the bias measured in the *Bias Trials* for either the parallel pairs ($t(99) = -2.20$, $p = 0.061$, $d = 0.25$) or for the orthogonal pairs ($t(99) = -0.87$, $p = 0.389$, $d = 0.11$), but there was strong evidence for ruling out learning only in the case of orthogonal pairs ($BF = 0.16$) not for the parallel ones ($BF = 1.1$). The results for all test types are visualized in Fig. 6.

6.4. Discussion

Overall, we found no learning of specific pairs in this setup despite the fact that not only the static spatial statistics but also the instructions for Experiments 3a and 3b were identical, clearly stating that shapes would be moving in and out horizontally. Thus, participants in Experiment 3b were aware of the movement of the underlying scene even without directly observing it through animation. In addition, the spatial structures in Experiment 3b were more isolated and fewer in any given scene compared to setups without an occluder; they were fully predictable when reappearing from behind the occluder and were not overshadowed by global motion. Still, participants were equally unable

to learn the orthogonal pairs in both experiments. Meanwhile, the parallel pairs were also not learned in Experiment 3b, showing a significant drop in performance compared to Experiment 3a, despite the two experiments having identical temporal statistics aside from the presence of animation. These results suggest that the learning of temporally presented parallel pairs in Experiment 3a was not driven primarily by temporal regularities, but rather resulted from a synergistic interaction between the statistical structure and the visually observed global movement.

Importantly, while in Experiment 3a participants showed an overall bias to choose options aligned with the movement direction, participants in Experiment 3b preferred options perpendicular to that direction. This shift in biases suggests that while in Experiment 3a the overall bias was induced by the observed general movement, the opposite bias observed in Experiment 3b was due to the static spatial layout defined by the orientation of the occluder and the shapes of the perpendicular pairs seen next to each other. These static-spatial-layout-based statistics were also available in Experiment 3a, but they were apparently overshadowed by the bias generated by the perceived global movement. Thus, our combined results from the two experiments highlight the flexible combination of local statistics and general biases according to their relative salience during the statistical learning process.

7. General discussion

7.1. Synthesis of results

Utilizing a novel VSL paradigm, the present study provides a systematic exploration of three major factors influencing the formation of internal representations in humans through implicit statistical learning: (a) the noisiness of the statistical contingencies available to the observer, (b) the interplay between purely spatial and purely temporal statistics during learning, and (c) the extent and nature of modulatory effects from pre-existing biases.

Assessing the first factor, the effect of noise is important because, by definition, lowering the transitional or co-occurrence probabilities of elements in sensory input (i.e., increasing noise) should hinder statistical learning and some informal results support this reasoning. However, the context and nature of the noise matter, as they can also create situations where reduced local transitional probabilities actually enhance the overall learning of structure (Gómez, 2002). Our results confirm this two-sided effect of noise. In our paradigm, observers were able to implicitly learn purely spatial visual structures, and this learning was not significantly affected by the presence of added spatial statistical noise, as evidenced by the lack of performance difference between Experiments 1a and 1b.

In contrast, this learning was strongly influenced by factors beyond purely spatial statistical probabilities -namely, the presence of temporal statistics derived from the coherent movement of the stimulus (Experiment 2c). This suggests that even in the most restricted version of coexisting spatial and temporal statistics -where the underlying structure is purely spatial and the temporal statistics can be clearly factorized away from the spatial structure by attributing them to global motion- the process of learning spatial structures in the world does not rely solely, or even predominantly, on momentary spatial co-occurrences, but also on temporal aspects of the input. This nontrivial effect is clearly demonstrated by the surprising superiority in learning diagonal pairs in Experiment 2, despite these having the highest level of spatial noise. By adding an occluder to the scene, we assessed the limits of this interaction -using temporal statistics to learn spatial structures- and found a complete transferability: purely spatial structures were learned from temporally presented statistical input without any direct presentation of the true spatial layout (Experiments 3a and 3b).

Regarding the third factor, the extent and nature of modulatory effects from pre-existing biases, the findings of Experiment 3 point to a more complex influence than the simple transfer of temporally

presented statistics to the formation of spatial structure representations. The Bias test results from Experiments 3a and 3b indicate that high-level knowledge of the same biasing information alone is insufficient to initiate an interaction between temporal statistics and the implicit learning of spatial structure. Although participants in Experiment 3b were fully aware of the direction of global motion during familiarization, this knowledge itself did not manifest as an implicit bias: in both the Bias test trials (using random pairs) and the Spatial test trials (using shapes from true pairs), participants tended to select pairs oriented consistently with the occluder's orientation during familiarization. Only the presence of a directly observable animation cue in Experiment 3a reversed this pattern, resulting in a bias toward the direction of motion in both the explicit Bias test and the more implicit Spatial test, effectively overriding the occluder's influence. The necessity of presenting global motion information in a sensory (rather than purely conceptual) format in order to influence implicit learning and generalization underscores a key distinction between the mechanisms underlying implicit sensory inference and explicit causal reasoning.

Interestingly, although the motion-based bias in Experiment 3a also influenced the Standard test, the deviation from chance was smaller for both parallel and orthogonal pairs compared to the Spatial test in the same experiment—even though the foils in the Standard test could be rejected based on both orientation and shape identity, whereas in the Spatial test only orientation could be used. This suggests an interplay more complex than simple additivity among shape familiarity, pair familiarity, motion-biased familiarity of pair orientation, and occlusion-based bias.

In sum, the general message of our present study is twofold. First, we found that even the most restricted combination of spatial and temporal aspects in the input—alongside a purely spatial underlying structure—produces results that cannot be predicted by simply extrapolating from prior studies of spatial and temporal statistical learning conducted in isolation. This suggests that perceiving and learning spatial and temporal regularities through VSL are not two separate, independent processes, but rather two intimately interacting components of an integrated mechanism. While this conclusion may not be entirely surprising, the extent of the interaction between spatial and temporal aspects of the input has not been clearly demonstrated before. Moreover, the vast majority of statistical learning studies do not investigate situations in which spatial and temporal aspects are combined. Therefore, our results should motivate the development of more sophisticated paradigms to systematically explore this integrated learning mechanism.

Second, even when combined, contingency-based low-level spatio-temporal statistical learning is not an independent process operating separately from more abstract aspects of the observer's internal knowledge. In parallel with learning specific chunks of the input scenes based on low-level spatio-temporal co-occurrence statistics, people also automatically develop various higher-level representations of general features of the input (e.g., variability of the input, overall motion direction, presence of an occluder, etc.). This knowledge generates biases that interact with—and strongly influence—implicit spatial structure learning on equal footing with low-level statistical contingencies. These higher-level biases are typically more general and can influence the learning process in a non-local manner—for example, by enhancing the learning of all spatial co-occurrence statistics that share orientation with the direction of global motion.

The complex and unexpectedly strong interactions among global motion-based and occlusion-based biases, along with the spatial and temporal statistics of the input—including the complete transfer of temporally presented correlational information to spatial knowledge—pose a serious challenge to traditional computational explanations of VSL based solely on co-occurrence or transitional probability counting. Instead, they support a shift away from this currently dominant interpretation of visual statistical learning toward a view of VSL as a flexible inference-making mechanism—one that continuously integrates various types of sensory and knowledge-based evidence to produce a fuller

interpretation of scenes during perception and learning, as proposed in Fiser and Lengyel (2022).

7.2. Relation to previous research

A few earlier statistical learning studies have already investigated the joint learning of temporal and spatial regularities. These studies found that infants could learn spatio-temporal sequences defined by the order of global positions (Kirkham, Slemmer, Richardson, & Johnson, 2007), that adults could transfer visually learned spatial associations to detect some of the same associations when presented temporally and vice versa (Turk-Browne & Scholl, 2009), and that spatio-temporal regularities could guide attention (Xu, Theeuwes, & Los, 2023). However, these studies primarily focused either on demonstrating the existence of such learning (e.g., *Can we use spatial and temporal statistics concurrently at all for learning?*) or on specific applications of the learned spatial and/or temporal statistics (e.g., *Can we use a subset of the learned temporal statistics in spatial tasks and vice versa? Can we use learned spatial or temporal statistics to guide attention?*).

The current work extends these previous studies in multiple ways. For example, while it has been shown that VSL is flexible enough to support successful performance on a temporal test after learning a spatial structure, and vice versa (Turk-Browne & Scholl, 2009), these findings can be explained by assuming that learning either type of statistics results in general associations between shapes. However, this does not demonstrate a higher level of complexity—namely, that participants retained any meaningful spatial structure after purely temporal learning (or vice versa) beyond simple co-association. In contrast, the current study connects the two domains in an ecologically relevant way by directly investigating the extent to which temporal coherence during learning can establish spatially defined structures, such as an oriented pair. Our results show that this learning goes beyond simple co-association of visible elements in the spatial structure and can operate through unconscious inference based purely on temporal structural information.

Similarly, Tummeltshammer and colleagues presented infants with a spatial structure setup in their spatial context condition that resembled the structure used in the current stVSL paradigm (Tummeltshammer, Amso, French, & Kirkham, 2017). However, in their trials, shapes entered from one end of the screen, moved across it in a single direction, and exited at the other end—thus introducing a temporal order to stimuli that were intended to be spatially defined. Furthermore, the same shape pairs appeared multiple times simultaneously on the screen, potentially introducing an uncontrolled pop-out effect of spatial regularity. In contrast, such effects can either be avoided or systematically studied—both in isolation and in interaction—using our stVSL setup, enabling more comprehensive conclusions about these interactions.

The most relevant prior work is a study demonstrating that, when possible, observers implicitly form temporal sequences based on spatial configurations rather than on single objects in a multi-element display (Yan, Ehinger, Pérez-Bellido, Peelen, & de Lange, 2023). This work can be viewed as complementary to ours: their main focus is the role of spatial regularities in the acquisition and perception of temporal patterns, whereas we focus on the role of temporal regularities in the acquisition of spatial patterns. Once these two questions are sufficiently understood in isolation, our setup could be extended to combine them in a single paradigm to investigate the numerous interconnected levels of spatial and temporal organization formed in real-world visual input, as our temporally presented spatial patterns could be arranged arbitrarily to predict each other.

The current work is also related to studies on learning spatial and temporal regularities outside the field of VSL. A special case of learning spatial representations from temporal statistics is investigated by studies of the trace learning rule, which explore how invariant object representations can emerge in an unsupervised manner by temporally associating different observed spatial patterns (Wallis & Rolls, 1997).

However, these studies address a somewhat distinct and complementary problem: how perceptually very different spatial patterns can be progressively co-associated with the same “object category” by relying on the temporal proximity of these patterns in time. While the trace rule focuses on building hierarchical representational structures of abstract categories, our study focuses on how relevant spatial statistical structure can be learned even at the lowest level of representation, based on multiple types of structural information.

Studies on *amodal completion* have investigated visual perception under partial occlusion (Kanizsa, 1985; van Lier & Gerbino, 2015). Other studies have investigated the top-down influences of prior object knowledge (Hazenbergh, Jongasma, Koning, & van Lier, 2014; Hazenbergh & van Lier, 2016; Yun, Hazenbergh, & van Lier, 2018). These studies are complementary to the current work in that they focus on partial presentations or occlusions during perception or inference, whereas we investigate these effects during learning. The crucial link between these studies and ours is their reliance on the unconscious inference mechanism to explain results and demonstrate how inference and learning are strongly intertwined at the computational level (Fiser, Berkes, Orbán, & Lengyel, 2010).

Another related line of research, labeled *aperture viewing* (Morgan, Findlay, & Watt, 1981) and *minimal videos* (Ben-Yosef, Kreiman, & Ullman, 2020) focuses on the minimal spatial and temporal information necessary to recognize objects. Similar to our work, both studies consider the integration of spatial and temporal information, with one crucial difference. These studies demonstrate how, in a rich spatio-temporal input space, the two types of information can be used interchangeably to achieve recognition. In contrast, we use a controlled situation to show the limits of how much one type of information is sufficient to create a representation of the other. Another study in the related area of spatial navigation reported that perceived spatiotemporal continuity helps with spatial long-term memory, independent of explicit memory performance (Liverence & Scholl, 2015). These results align with our findings, showing that removing overt cues to spatio-temporal coherence (i.e., movement animation) hinders the implicit formation of memories of spatial patterns.

In sum, our study differs from earlier reports in two general ways. First, it provides a more comprehensive and controlled examination of the possible interactions between spatial and temporal sensory structures and internally generated biases. Second, it aims to establish general constraints and gain insights into the computational frameworks that can explain the emerging behavioral patterns in statistical learning.

7.3. Potential extensions

Our experimental design was deliberately kept simple: spatially fixed shape-pair structures, a single global motion pattern for temporal statistics, and one occlusion structure. This minimal setup was sufficient to address the study’s questions and demonstrated that, even under such simplified conditions, complex interplays between statistical input and learning can emerge. While we highlighted the interactions between various effects shaping statistical learning, we did not provide a direct explanation for one of the intricate patterns emerging in Experiments 2a–c—namely, that the performance rankings for different pair types did not align with the strength of traditional conditional probabilities in spatial statistics. Specifically, we found that diagonal pairs were learned best, suggesting that temporal transitional probabilities to neighboring (but not identical) cells may have exerted a stronger influence during the unconscious inference process.

Although further investigation of this effect was beyond the scope of the current study, future research is needed to clarify which of several potential underlying mechanisms is at play. First, the implicit integration of spatial information across time in our experiment may be strongest not at the exact same spatial location, but equally strong at neighboring locations. This effect could be driven by afterimages resulting in masking and/or by an extension of the mechanism that

supports conventional spatial VSL over time. In this case, participants may have associated the content of a grid cell not only with the content of neighboring cells at the same moment, but also with previously seen content stored in working memory. Second, the effect could be influenced by the relative uniqueness of the diagonal pairs compared to other types. While there are two structurally identical parallel and two orthogonal pairs (i.e., pairs of two shapes next to each other horizontally or vertically), the two diagonal pairs are more distinct—one arranged high-left to low-right, the other low-left to high-right. As shown in recent work, such structural features can influence VSL through potential interference (Garber & Fiser, 2024).

The two properties jointly inducing the learning biases in the current experiments—perceived movement direction and the perceived arrangement of shapes—are likely just two examples from a larger set of factors that can influence statistical learning. One such factor is the modality in which learning occurs. An intriguing question is whether the same paradigm, implemented in both the visual and auditory modalities, would yield similar specific results beyond any overall differences in learning efficiency. Recent evidence suggests that, in the case of simple temporal chunking of a long sound or visual scene sequence, human observers exhibit similar biases in both audition and vision (Garami & Fiser, 2024). If similar homology is found in learning structural information across the two modalities, it would provide strong support for the domain-general learning of statistical learning (Frost, Armstrong, Siegelman, & Christiansen, 2015). The controlled setup of the present experimental design offers a natural testbed to explore this issue.

There are two important extensions of the basic design that should be addressed by future research to better link statistical learning to human representational learning in its full complexity. The first extension relates to research under the titles of transfer learning and curriculum learning. Studies in these domains investigate how further abstractions of biases—such as global motion and occlusion in our paradigm—can emerge during the training phase and influence the future learning of representational hierarchies (Dekker, Otto, & Summerfield, 2022; Whittington et al., 2020). These abstract biases—for example, attributing motion to a visual region without observed motion in that area, or expecting the appearance of a particular shape from behind an occluder—can emerge only after lower-level structures have been learned and therefore do not belong to the original explicit feature space. A recent study using a new sVSL paradigm showed evidence of such transfer learning in a VSL context (Garber & Fiser, 2024). This paradigm can be naturally combined with the design of the present study to investigate the emergence of higher-level biases during transfer learning in a spatio-temporal context.

The second direction is “active learning,” which can be explored within the current paradigm by giving participants control over the direction of movement. In this setup, the nature and dynamics of participants’ exploratory behavior—and its interaction with already acquired knowledge—could be examined, similar to a recent gaze-contingent approach (Arató et al., 2024), using novel measures of VSL such as predicting upcoming stimuli based on currently presented partial pairs. Furthermore, our spatio-temporal VSL paradigm enables the exploration of neural correlates through methods such as neural frequency tagging, which could be extended from its current application in purely temporal statistical learning (Batterink & Paller, 2017; Moser et al., 2021) to the learning of spatially defined structures.

The extensions described above outline a progression from traditional VSL methods, which investigate the basic steps of implicit learning, to paradigms better suited to examining more directly how implicit learning in humans can support the more intricate integration of spatial and temporal statistics with knowledge-based concepts. The ultimate result of such integration is the emergence of a complex, abstract structural description, or “world model” (Bramley, Zhao, Quillien, & Lucas, 2023). Accordingly, the relevant research questions in these paradigms will shift from those asked in the present study to questions about the nature of such abstract representations of our complex,

continuously changing dynamic environment -for example, the nature of “object representation.” Our stVSL paradigm and the results presented in this study thus provide a first small step toward a fuller understanding of how humans learn a comprehensive and coherent, yet parsimonious, representation of the world.

CRedit authorship contribution statement

Dominik Garber: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **József Fiser:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Acknowledgments

This research was funded in part by the Austrian Science Fund (FWF) [DOI:10.55776/16793].

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.cognition.2025.106324>.

Data availability

The experimental data that support the findings of this study are available on OSF: https://osf.io/v2r85/?view_only=b0dc196385b54afea3aa768908b0306f

References

- Arató, J., Rothkopf, C. A., & Fiser, J. (2024). Eyemovements reflect active statistical learning. *Journal of Vision*, 24(5). <https://doi.org/10.1167/jov.24.5.17>, 17, 1–18.
- Aslin, R. N. (2017). Statistical learning: A powerful mechanism that operates by mere exposure. *Wiley Interdisciplinary Reviews. Cognitive Science*, 8(1–2), Article e1373. <https://doi.org/10.1002/wcs.1373>
- Baillargeon, R. (2008). Innate ideas revisited: For a principle of persistence in infants' physical reasoning. *Perspectives on Psychological Science*, 3(1), 2–13. <https://doi.org/10.1111/j.1745-6916.2008.00056.x>
- Batterink, L. J., & Paller, K. A. (2017). Online neural monitoring of statistical learning. *Cortex*, 90, 31–45. <https://doi.org/10.1016/j.cortex.2017.02.004>
- Ben-Yosef, G., Kreiman, G., & Ullman, S. (2020). Minimal videos: Trade-off between spatial and temporal information in human and machine vision. *Cognition*, 201, Article 104263. <https://doi.org/10.1016/j.cognition.2020.104263>
- Bramley, N. R., Zhao, B., Quillien, T., & Lucas, C. G. (2023). Local search and the evolution of world models. *Topics in Cognitive Science. Advanced online publication*.
- Carlson, V. R. (1962). Size-constancy judgments and perceptual compromise. *Journal of Experimental Psychology*, 63(1), 68–73. <https://doi.org/10.1037/h0045909>
- Dekker, R. B., Otto, F., & Summerfield, C. (2022). Curriculum learning for human compositional generalization. *Proceedings of the National Academy of Sciences of the United States of America*, 119(41), Article e2205582119. <https://doi.org/10.1073/pnas.2205582119>
- Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, 12(6), 499–504. <https://doi.org/10.1111/1467-9280.00392>
- Fiser, J., & Aslin, R. N. (2002a). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 28(3), 458–467. <https://doi.org/10.1037/0278-7393.28.3.458>
- Fiser, J., & Aslin, R. N. (2002b). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences*, 99(24), 15822–15826. <https://doi.org/10.1073/pnas.232472899>
- Fiser, J., & Aslin, R. N. (2005). Encoding multielement scenes: Statistical learning of visual feature hierarchies. *Journal of Experimental Psychology. General*, 134(4), 521–537. <https://doi.org/10.1037/0096-3445.134.4.521>
- Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Sciences*, 14(3), 119–130. <https://doi.org/10.1016/j.tics.2010.01.003>
- Fiser, J., & Lengyel, G. (2022). Statistical learning in vision. *Annual Review of Vision Science*, 8(17), 1–26. <https://doi.org/10.1146/annurev-vision-100720-03343>
- Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences*, 19(3), 117–125.
- Garami, L., & Fiser, J. (2024). Temporal segmentation principles in vision and audition. *Journal of Vision*, 24(10), 1117.
- Garber, D., & Fiser, J. (2024). Structure transfer and consolidation in visual implicit learning. *eLife*. <https://doi.org/10.7554/eLife.100785.1>
- Gepshtein, S., & Kubovy, M. (2000). The emergence of visual objects in space-time. *Proceedings of the National Academy of Sciences of the United States of America*, 97(14), 8186–8191. <https://doi.org/10.1073/pnas.97.14.8186>
- Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, 13(5), 431–436. <https://doi.org/10.1111/1467-9280.00476>
- Hazenberg, S. J., Jongsma, M. L. A., Koning, A., & van Lier, R. (2014). Differential familiarity effects in Amodal completion: Support from behavioral and electrophysiological measurements. *Journal of Experimental Psychology. Human Perception and Performance*, 40(2), 669–684. <https://doi.org/10.1037/a0034689>
- Hazenberg, S. J., & van Lier, R. (2016). Disentangling effects of structure and knowledge in perceiving partly occluded shapes: An ERP study. *Vision Research*, 126, 109–119. <https://doi.org/10.1016/j.visres.2015.10.004>
- Hochberg, J. (1968). In the mind's eye. In R. N. Haber (Ed.), *Contemporary theory and research in visual perception*. Holt, Rinehart & Winston.
- Isbilen, E. S., & Christiansen, M. H. (2022). Statistical learning of language: A Meta-analysis into 25 years of research. *Cognitive Science*, 46(9), Article e13198. <https://doi.org/10.1111/cogs.13198>
- Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2), 201–211. <https://doi.org/10.3758/bf03212378>
- Kanizsa, G. (1985). Seeing and thinking. *Acta Psychologica*, 59(1), 23–33.
- Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, 83(2), B35–B42. [https://doi.org/10.1016/S0010-0277\(02\)00004-5](https://doi.org/10.1016/S0010-0277(02)00004-5)
- Kirkham, N. Z., Slemmer, J. A., Richardson, D. C., & Johnson, S. P. (2007). Location, location, location: Development of spatiotemporal sequence learning in infancy. *Child Development*, 78(5), 1559–1571. <https://doi.org/10.1111/j.1467-8624.2007.01083.x>
- Lee, A. L. F., Liu, Z., & Lu, H. (2021). Parts beget parts: Bootstrapping hierarchical object representations through visual statistical learning. *Cognition*, 209, Article 104515. <https://doi.org/10.1016/j.cognition.2020.104515>
- van Lier, R., & Gerbino, W. (2015). In J. Wagemans (Ed.), *Perceptual completions*. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199686858.013.040>
- Liu, Y., Dolan, R. J., Kurth-Nelson, Z., & Behrens, T. E. J. (2019). Human replay spontaneously reorganizes experience. *Cell*, 178(3), 640–652.e14. <https://doi.org/10.1016/j.cell.2019.06.012>
- Liverence, B. M., & Scholl, B. J. (2015). Object persistence enhances spatial navigation: A case study in smartphone vision science. *Psychological Science*, 26(7), 955–963. <https://doi.org/10.1177/0956797614547705>
- Morgan, M. J., Findlay, J. M., & Watt, R. J. (1981). Aperture viewing: A review and a synthesis. *The Quarterly Journal of Experimental Psychology*, 34(2), 211–233. <https://doi.org/10.1080/14640748208400837>
- Moser, J., Batterink, L., Hegner, Y. L., Schlegler, F., Braun, C., Paller, K. A., & Preissl, H. (2021). Dynamics of nonlinguistic statistical learning: From neural entrainment to the emergence of explicit knowledge. *NeuroImage*, 240, Article 118378. <https://doi.org/10.1016/j.neuroimage.2021.118378>
- Piaget, J. (1954). *The construction of reality in the child*. Basic Books.
- Rolls, E. T. (2012). Invariant visual object and face recognition: Neural and computational bases, and a model. *VisNet. Frontiers in Computational Neuroscience*, 6, 35. <https://doi.org/10.3389/fncom.2012.00035>
- Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, 56(5), 356–374. <https://doi.org/10.1016/j.jmp.2012.08.001>
- Santolin, C., & Saffran, J. R. (2018). Constraints on statistical learning across species. *Trends in Cognitive Sciences*, 22(1), 52–63. <https://doi.org/10.1016/j.tics.2017.10.003>
- Stone, J. V. (1998). Object recognition using spatiotemporal signatures. *Vision Research*, 38(7), 947–951. [https://doi.org/10.1016/S0042-6989\(97\)00301-5](https://doi.org/10.1016/S0042-6989(97)00301-5)
- Sun, J., & Perona, P. (1998). Where is the sun? *Nature Neuroscience*, 1(3), 183–184. <https://doi.org/10.1038/630>
- Tummelshammer, K., Amso, D., French, R. M., & Kirkham, N. Z. (2017). Across space and time: Infants learn from backward and forward visual statistics. *Developmental Science*, 20(5), Article e12474. <https://doi.org/10.1111/desc.12474>
- Turk-Browne, N. B. (2012). Statistical learning and its consequences. In , 59. *Nebraska symposium on motivation. Nebraska Symposium on Motivation* (pp. 117–146). https://doi.org/10.1007/978-1-4614-4794-8_6
- Turk-Browne, N. B., & Scholl, B. J. (2009). Flexible visual statistical learning: Transfer across space and time. *Journal of Experimental Psychology. Human Perception and Performance*, 35(1), 195–202. <https://doi.org/10.1037/a0096-1523.35.1.195>
- Wade, N. J., Spillmann, L., & Swanston, M. T. (1996). Visual motion aftereffects: Critical adaptation and test conditions. *Vision Research*, 36(14), 2167–2175. [https://doi.org/10.1016/0042-6989\(95\)00266-9](https://doi.org/10.1016/0042-6989(95)00266-9)
- Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology*, 51(2), 167–194. [https://doi.org/10.1016/S0301-0082\(96\)00054-8](https://doi.org/10.1016/S0301-0082(96)00054-8)
- Whittington, J. C. R., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. J. (2020). The Tolman-Eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5), 1249–1263.e23. <https://doi.org/10.1016/j.cell.2020.10.024>

Xu, Z., Theeuwes, J., & Los, S. A. (2023). Statistical learning of spatiotemporal regularities dynamically guides visual attention across space. *Attention, Perception & Psychophysics*, 85(4), 1054–1072. <https://doi.org/10.3758/s13414-022-02573-5>

Yan, C., Ehinger, B. V., Pérez-Bellido, A., Peelen, M. V., & de Lange, F. P. (2023). Humans predict the forest, not the trees: Statistical learning of spatiotemporal structure in

visual scenes. *Cerebral Cortex*, 33(13), 8300–8311. <https://doi.org/10.1093/cercor/bhad115>

Yun, X., Hazenberg, S. J., & van Lier, R. (2018). Temporal properties of amodal completion: Influences of knowledge. *Vision Research*, 145, 21–30. <https://doi.org/10.1016/j.visres.2018.02.011>