Cognition 212 (2021) 104714

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Visual memorability in the absence of semantic content

Qi Lin *, Sami R. Yousif *, Marvin M. Chun, Brian J. Scholl

Department of Psychology, Yale University, New Haven, CT, 06520-8205, United States of America

ARTICLE INFO

Keywords:
Perception
Visual memory
Memorability
Long-term memory
Short-term memory

ABSTRACT

What determines how well people remember images? Most past research has explored properties of the people doing the remembering — such as their age, emotional state, or individual capacity. However, recent work has also characterized *memorability* — the likelihood of an image being remembered *across* observers. But what makes some images more memorable than others? Part of the answer must surely involve the meanings of the images, but here we ask whether this is the entire story: is there also purely *visual* memorability, driven not by semantic content but by perceptual features *per se*? We isolated visual memorability in an especially direct manner — by eliminating semantic content while retaining many visual properties. We did so by transforming a set of natural scene images using phase scrambling, and then testing memorability for both intact and scrambled images in independent samples. Across several experiments, observers saw sequences of images and responded anytime they saw a repeated image. We found reliable purely *visual* memorability at the temporal scales of both short-term memory (2–15 s) and longer-term memory (several minutes), and this could not be explained by the extent to which people could generate semantic labels for some scrambled images. Collectively, these results suggest that the memorability of images is a function not only of what they mean, but also of how they look in the first place.

1. Introduction

When you view an image, what determines whether you will remember it later on? Some of the relevant factors depend on you — e.g. on your working memory capacity (Kane & Engle, 2000), or how emotionally aroused you were when you saw the image (Cahill & McGaugh, 1998). But other relevant factors depend on the image itself. Some of these image-specific factors may be idiosyncratic: for example, you may remember one image because it reminds you of your childhood home. But other image-specific factors may hold across people. Indeed, recent work has characterized the *memorability* of images — the likelihood that they are remembered *across* observers (Bainbridge, Isola, & Oliva, 2013; Isola, Xiao, Parikh, Torralba, & Oliva, 2014; Khosla, Xiao, Torralba, & Oliva, 2012; for a review, see Bainbridge, 2019). Despite all sorts of individual differences, it turns out that people generally remember and forget the same images. This sort of image memorability seems relatively robust and reliable: it is observed across multiple experimental paradigms (Broers, Potter, & Nieuwenstein, 2017; Goetschalckx, Moors, & Wagemans, 2017); it is at least somewhat independent of which other images have been recently encountered (Bylinskii, Isola, Bainbridge, Torralba, & Oliva, 2015); and it does not seem to depend on how much time has elapsed since the image was first seen (Goetschalckx et al., 2017).

1.1. Semantic and perceptual factors

So what makes one image more memorable than others? Clearly, part of the answer will depend on the image's meaning. For example, people might reliably remember a particular image because it depicts a cute baby, or a threatening animal, or an unusual juxtaposition (such as a cute baby next to a threatening animal). Indeed, when images are coded for various properties, these sorts of semantic attributes and categories are the primary predictors of which images are more memorable than others (Isola et al., 2014; Khosla, Raju, Torralba, & Oliva, 2015). And this is perhaps to be expected from the wider study of memory, since conceptual factors play a substantial role in visual long-term memory, in both recognition (Konkle, Brady, & Alvarez, 2010) and recall (Bainbridge, Hall, & Baker, 2019).

However, images may also still be differentially memorable even when meanings seem to be held roughly constant: one image of a mountain, for example, might end up being almost 70% more memorable than another image of a mountain (e.g. see Fig. 2 from Bylinskii

* Corresponding authors.
E-mail addresses: qi.lin@yale.edu (Q. Lin), sami.yousif@yale.edu (S.R. Yousif).

https://doi.org/10.1016/j.cognition.2021.104714
Received 2 December 2018; Received in revised form 15 February 2021; Accepted 28 March 2021
Available online 7 May 2021

et al., 2015). What makes the difference in this case? Perhaps the answer just has to do with a finer grain of semantic resolution (such that one of the mountains just looks more dangerous than the other) or the fact (as in the examples from Bylinskii et al., 2015) that memorable images of mountains tend to contain humans. Here, in contrast, we ask whether this type of difference in memorability may also in part reflect purely visual factors that operate independently from semantics (e.g. if one mountain image has a markedly different spatial frequency profile or degree of visual segmentation).

1.2. The current study: Purely visual memorability?

In the current study, we ask if there is purely *visual* memorability — memorability driven not by semantic content but by perceptual features *per se*.

How can we empirically distinguish purely visual factors from semantic factors, especially given how correlated they may be (e.g. Kardan et al., 2015; Long, Konkle, Cohen, & Alvarez, 2016)? One strategy would be to measure visual properties in advance, and then to see how well those measurements predict memorability (e.g. Isola et al., 2014; Khosla et al., 2012). In fact, past work along these lines has indicated that simple visual features such as the mean or variance in hue and saturation are *not* reliable predictors (Isola et al., 2014; Lukavský & Děchtěrenko, 2017). The difficulty with this approach, however, is that the universe of possible visual properties is large, and so one must somehow divine which properties are most relevant in order to measure and test them. An alternative method could be to directly manipulate subtle visual details in the images (e.g. removing lamp posts or flower pots, as in the images used by Vogt & Magnussen, 2007), and then investigate the effects of such manipulations on memorability. But such manipulations are themselves still meaningful, even if they involve relatively small regions of the images. In the end, perhaps the biggest challenge with all such approaches is just that the images themselves still always have rich meaning (a 'conceptual hook'; Konkle et al., 2010) with highly recognizable objects, and so they are unable to completely unconfound meaning from perceptual features.

Accordingly, we take an especially direct but radically different approach in the current study: we isolated purely visual memorability by *eliminating* semantic content while retaining many lower-level visual properties. We first sampled images from a database that has been used to study memorability (Isola et al., 2014). To minimize the salience of semantic content, we limited our sample to natural scenes, many of which came from the same categories (e.g. forests, fields, lakes). We then disrupted the semantic contents of the scenes using phase scrambling (e.g. Oppenheim & Lim, 1981; Rossion & Caharel, 2011; Thomson, 1999), which involves randomizing an image's phase spectrum while maintaining its amplitude spectrum. As can be appreciated from Fig. 1, these manipulations severely constrain or even eliminate the meanings of the scenes that are so apparent in the intact images. But at the same time, they preserve many lower-level visual properties from the images — such as color distribution, spatial frequency profile, and degree of overall segmentation.[1] We then tested the memorability of intact and phase-scrambled versions of the same images. We also note that past work has repeatedly found that short-term memory is relatively more dependent on perceptual factors, whereas long-term memory is relatively more dependent on conceptual factors (e.g. Baddeley, 1966a, 1966b; Konkle et al., 2010). Accordingly, we tested for purely visual memorability at the temporal scales of both short-term memory (2–15 s) and longer-term memory (several minutes).

[1] Of course there are also several other image scrambling techniques, but we chose to use phase scrambling here because of how reliably it eliminates semantic content. (In contrast, some other manipulations — such as diffeomorphic scrambling [Stojanoski & Cusack, 2014] — may preserve meaningfulness to a greater degree.)

Since the scrambled scenes have been stripped of their previous semantic contents, we expect that they will be remembered far worse than the original intact images. However, as long as there is memory overall, we can assess memorability. If there is any substantial degree of purely visual memorability, then observers should still reliably remember certain scrambled images more than others. But if memorability does not ultimately involve a purely visual component, then the scrambling manipulations should effectively eliminate memorability.

2. Experiments 1a and 1b: Short term visual memorability

In an initial study, observers had to respond to repetitions of images that came after a short lag (of 2–15 s and 1–6 items, thus in the span of short-term memory). In Experiment 1a, we tested the memorability of images that were Intact or Phase-scrambled (in separate groups of observers). And in Experiment 1b, we tested just the Phase-scrambled images alone, now generated using a different random seed.

2.1. Method

2.1.1. Participants

156 and 78 observers recruited via Amazon Mechanical Turk participated in Experiment 1a and Experiment 1b, respectively. To participate, the workers must have: (a) completed at least 100 studies; (b) received an approval rate of at least 95% in past studies; and (c) had a US IP address. Additionally, we excluded any participants whose memory performance (measured by d') was lower than 0. (Especially for the Phase-scrambled images, the task may have been too hard, such that some observers were just randomly guessing. Such observers obviously couldn't contribute to questions about memorability, since there would be no memory signal to begin with.) There were thus 78 unique observers in each experimental group in each experiment, with this sample size chosen before data collection began to exactly match the average sample size from prior work (Isola et al., 2014). All observers gave informed consent, and were compensated for their time.

2.1.2. Stimuli

In Experiment 1a, we used 48 Intact (256-pixel square) images (all of natural scenes, without any humans or animals) from the target images released by Isola et al. (2014). The memorability scores of these 48 Short Term Target images spanned (and were uniformly distributed across) the full range of the original scores (as measured by Isola et al., 2014). In addition, 48 filler images were randomly sampled from the filler images of Isola et al. (2014) and were not analyzed. These Intact images were then each transformed using a phase scrambling algorithm (with MATLAB 2016b; MathWorks, Natick, MA). As depicted in Fig. 1, the phase scrambling algorithm (based on Prins, 2007) maintained the amplitude spectra from the Intact images, while randomizing their frequency spectra. For Experiment 1b, we re-scrambled the same set of Intact images using the phase scrambling algorithm, but with a different random seed. Each image was presented in the center of the observer's browser window, on a solid white background, surrounded by a 4-pixel solid black border.

2.1.3. Procedure

In Experiment 1a, each observer was assigned to one of the two image conditions (Intact or Phase-scrambled); in Experiment 1b, observers were all assigned to the Phase-scrambled condition. Each observer saw a sequence of images (each presented for 1.2 s, separated by a blank 1.2 s delay) and responded (by pressing 'r') whenever they saw a repeated image. Observers were given visual feedback (with the image's border turning from black to a lighter gray) whenever the response key was pressed, but they were given no feedback about whether their responses were correct or not (see Fig. 2). Unbeknownst to the observers, only the Short Term Targets repeated, each after a lag of 1–6 images, with 8 images repeated at each lag. There were thus 144

Fig. 1. Examples of Intact images (top) and the corresponding Phase-scrambled images (bottom).

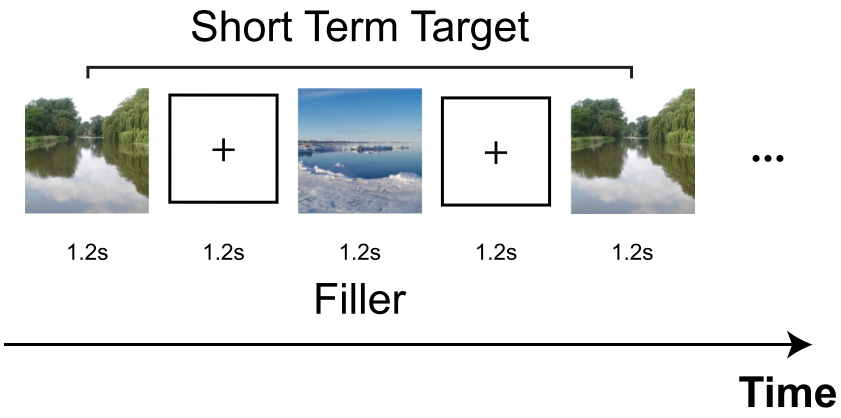

Fig. 2. The design of Experiments 1a and 1b. (A) Experimental groups; (B) Schematic representation of the memory task.

images in total (48 Short Term Targets, 48 Repeated Short Term Targets, and 48 Fillers), presented in a different randomized order (subject to the constraints stated above) for each observer.

2.2. Results

2.2.1. Memory performance

Before exploring memorability, we first assessed the overall fidelity of memory for the images. The d' values for Short Term Targets in all conditions in both experiments are depicted in Fig. 3A. Inspection of this figure suggests that there was reliable memory for the images in all conditions, but (unsurprisingly) that memory was better for the Intact images. These impressions were verified by the following analyses.

In Experiment 1a, Short Term Targets were remembered above chance for both image types (Intact: $t(77) = 24.96$, $p < .001$, $d = 4.02$, all tests in this experiment were compared against the Bonferroni corrected alpha level of 0.0125[0.05/4]; Phase-scrambled: $t(77) = 17.82$, $p < .001$, $d = 2.87$). As expected, memory for Short Term Targets was better for Intact images than for Phase-scrambled images ($t(154) = 12.24$, $p < .001$, $d = 1.97$). In Experiment 1b, Phase-scrambled Short Term Targets were remembered above chance ($t(77) = 18.38$, $p < .001$, $d = 2.96$).

2.2.2. Short term memorability

To assess memorability (i.e. the consistency with which images are remembered across people), we randomly split the observers into two halves, and then calculated two sets of 'memorability scores': corrected hit rates (hit rates – false alarm rates) for all images based on the two halves. We then correlated the two sets of memorability scores, and repeated this procedure 1000 times for each image condition. If there is memorability (such that some images are consistently remembered better than others across observers), then we would expect to see positive correlations between the memorability scores calculated based on

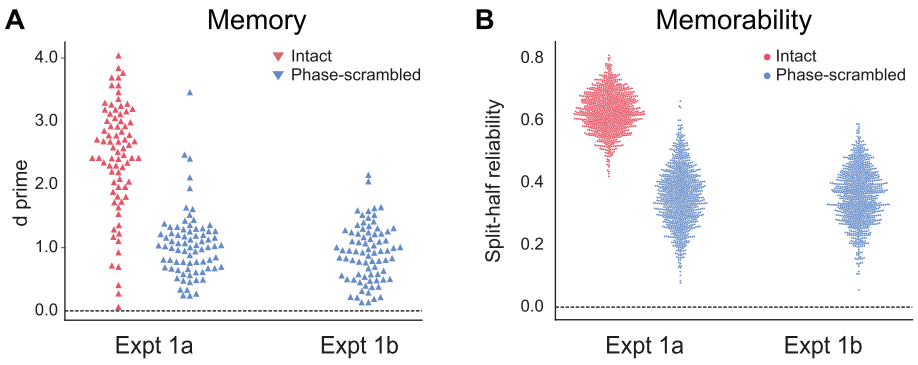

Fig. 3. Results of Experiments 1a and 1b. (A) Short-term memory performance, where each triangle represents a participant; (D) Split-half reliability of Short Term memorability scores, where each dot represents a Spearman rank correlation value of one split-half iteration.

the two halves, no matter how we split the observer sample. The Spearman rank correlations across the 1000 random splits are depicted in Fig. 3B. Inspection of this figure clearly suggests that there was Short Term memorability for both Intact and Phase-scrambled images (though to a greater degree for the Intact images). (We purposefully chose a conservative criterion to evaluate memorability; 'strong evidence' requires that 97.5% of the 1000 random splits resulted in a correlation larger than 0.) These impressions were verified by the following analyses.

In Experiment 1a, the averaged Spearman rank correlation was 0.63 for the Intact targets (95% CI: [0.50, 0.75]) — thus demonstrating for the first time that the Intact images are also differentially memorable at a timescale of only 2–15 s. (Although past memorability studies have employed different delays within the realm of long-term storage [e.g., Isola et al., 2014; Goetschalckx et al., 2017], the shortest lag from this past work is still 15 images, for a delay of approximately 36 s — which is still well outside the capacity and span of short-term memory [Miller, 1994; Cowan, 2001].) More remarkably, there was also reliable Short Term memorability for the Phase-scrambled targets — a result that was replicated in both Experiment 1a (0.36, 95% CI: [0.18, 0.54]) and Experiment 1b (0.36, 95% CI: [0.19, 0.52]). The distributions of memorability scores defined over the entire sample are depicted in Fig. 4.

2.2.3. Intact vs. phase-scrambled short term memorability

Having demonstrated Short Term memorability for both the Intact and Phase-scrambled images, it seems natural to wonder how these two forms of memorability are related. In particular, do the most (and least) memorable Intact images remain just as memorable when scrambled? To find out, we calculated the memorability scores for all images based on the full sample in each condition and then correlated them across different image conditions using Spearman rank correlation. The most forgettable and memorable images in each condition of Experiments 1a and 1b are presented in Figs. 5 (the top row in each panel) and 6 (top two rows in each panel). Inspection of these images suggests that the most memorable (and forgettable) Phase-scrambled images did not always correspond to the most memorable (and forgettable) Intact images. These observations were verified by the following analyses.

The memorability scores for Intact images from Experiment 1a did not correlate with those of the Phase-scrambled versions of the same images in either Experiment 1a ($\rho = 0.01$, $p = .943$) or Experiment 1b ($\rho = -0.07$, $p = .626$). And strikingly, there was also no correlation between the memorability scores for the Phase-scrambled images using different seeds across the two experiments ($\rho = 0.05$, $p = .718$).

2.3. Discussion

The results of this experiment clearly demonstrate that there is purely visual Short Term memorability — and that this effect is reliable and replicable: across short delays, people tend to retain the same set of images in their short-term memory, even when the meanings of the original images have been severely disrupted.

In addition, the lack of correlation between the memorability scores for the Phase-scrambled vs. Intact images suggests that the features driving the purely visual memorability in the Phase-scrambled images may operate independently of the semantic factors driving memorability in the Intact images. And these visual properties may also interact in nuanced ways. Recall that there was also no correlation between the memorability scores for the two sets of scrambled images generated using different random seeds. Consistent with the proposal that memorability reflects the statistical distinctiveness of a stimulus along a multidimensional set of axes (Bainbridge, 2019; Bainbridge, Dilks, &

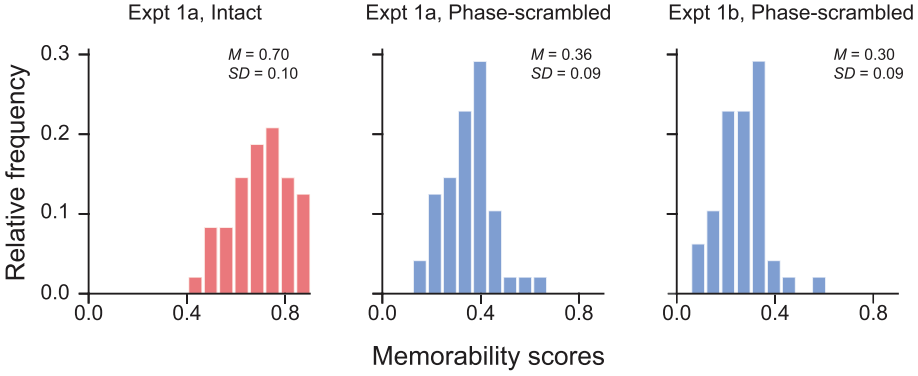

Fig. 4. Distributions of memorability scores for target images in Experiments 1a and 1b.

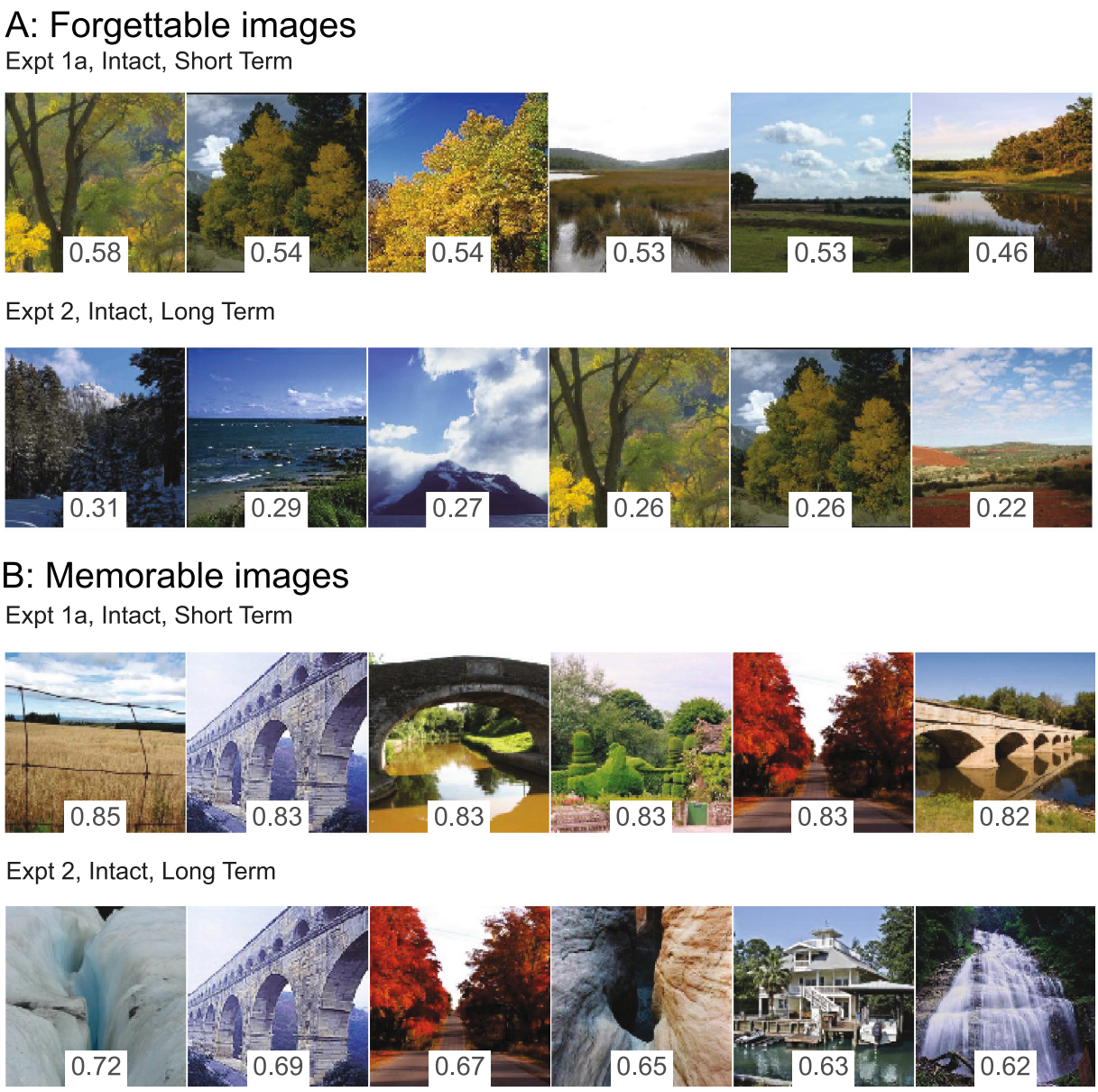

Fig. 5. (A) The 6 most forgettable and (B) the 6 most memorable target Intact images in Experiments 1a and 2. The numbers at the bottom of each image are the memorability scores.

Oliva, 2017), this suggests that the memorability scores of the scrambled images are not simply determined by the lower-level visual features preserved by phase scrambling, such as spatial frequency profile and degree of visual segmentation. Instead, they may be affected by the *interplay* between various dimensions of lower-level visual features (or the spatial configurations of the features) which may not be identical across different random seeds. (There has also been work showing that observers are more sensitive to spatial distortion and color changes in phase-scrambled images compared to natural images [Bex, 2010; Yoonessi & Kingdom, 2008] — which may also help to explain why the subtle changes to the phase-scrambled images led to different patterns of memorability scores.)

3. Experiment 2: Long term memorability

Purely visual memorability seems intuitively most likely to arise in a Short Term context (as in Experiments 1a and 1b), since as noted above, long-term memory is relatively more dependent on conceptual factors (e.g. Baddeley, 1966a; Konkle et al., 2010). At the same time, however, past work with intact images has not observed reliable effects on memorability of the time elapsed since the image was first seen (Goetschalckx et al., 2017). So would we still observe purely visual memorability even at a longer timescale? To find out, observers in Experiment 2 had to respond to repetitions of images that could come after either a short lag (of 2–15 s, as in Experiments 1a and 1b), or a longer lag (of several minutes, a new condition in this experiment). In separate groups of observers, we tested the memorability of images that were Intact or Phase-scrambled.

3.1. Method

This experiment was identical to Experiments 1a and 1b except as noted below.

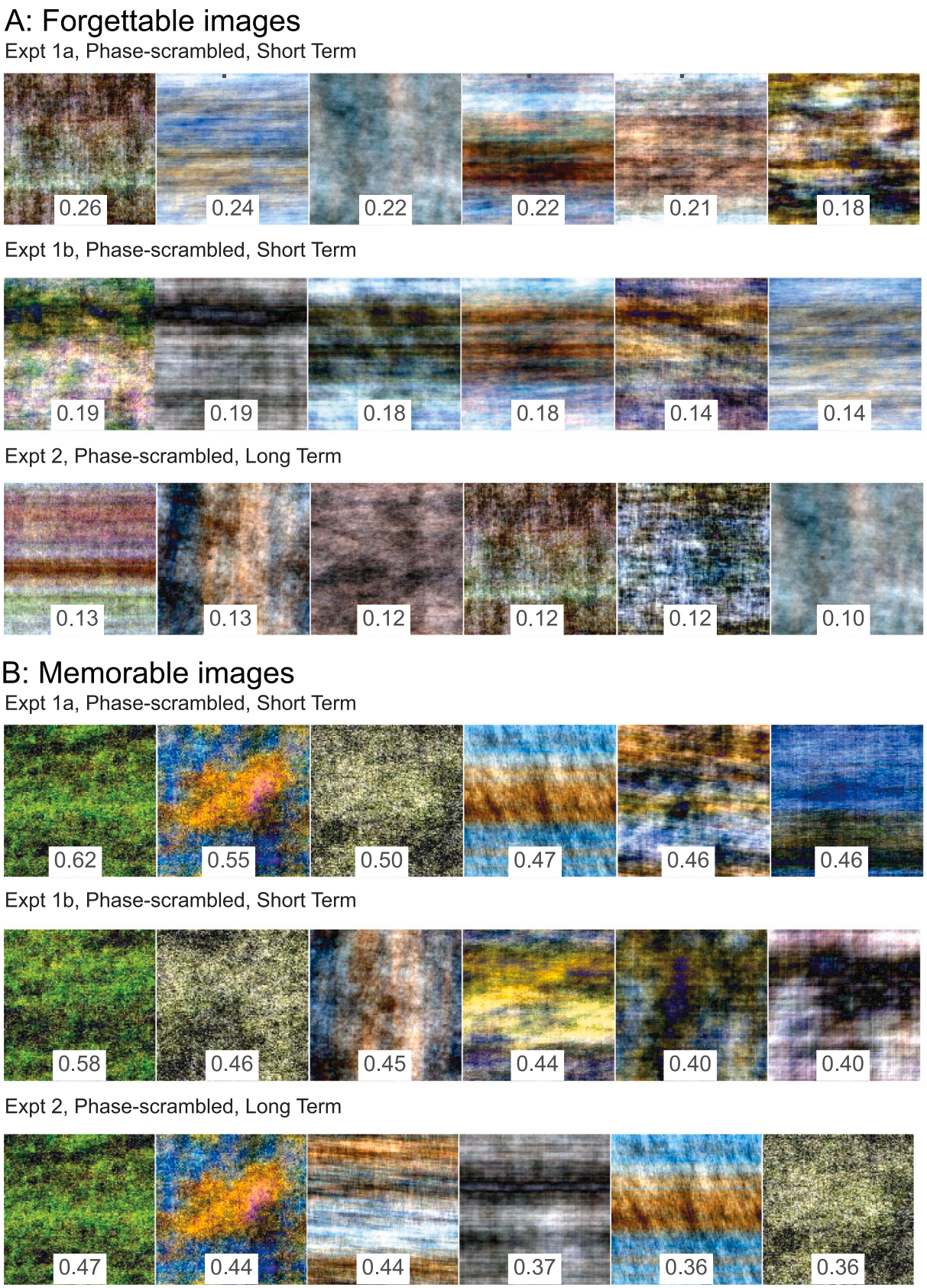

Fig. 6. (A) The 6 most forgettable and (B) the 6 most memorable target Phase-scrambled images in Experiments 1a, 1b and 2. The numbers at the bottom of each image are the memorability scores. Note that the images used for Experiments 1a and 2 were exactly the same whereas those used in Experiment 1b were generated from the same Intact images but with a different random seed.

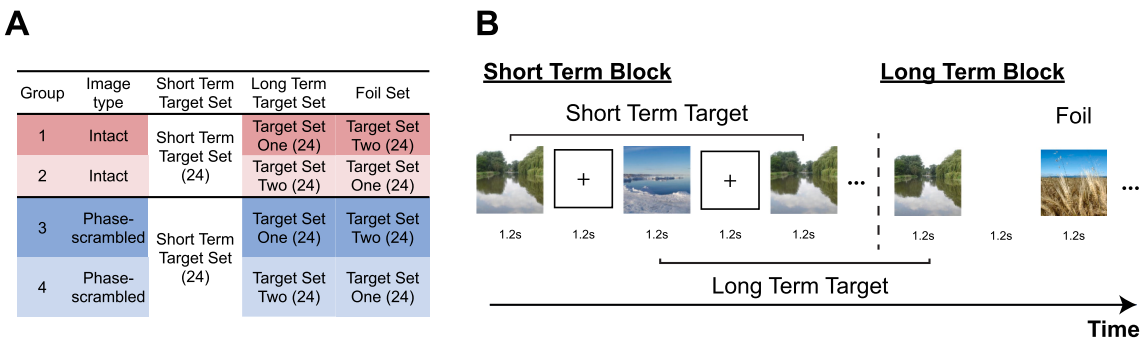

Fig. 7. The design of Experiment 2. (A) The four experimental groups; (B) Schematic representation of the Short Term and Long Term blocks.

3.1.1. Participants

312 observers were recruited, and each was assigned to one of four experimental groups (as described below and depicted in Fig. 7A): (a) Intact Target One; (b) Intact Target Two; (c) Phase-scrambled Target One; and (d) Phase-scrambled Target Two. Additionally, we again excluded any participants whose memory performance (measured by d') during the Short Term Block was lower than 0. There were thus 78 unique observers in each experimental group (with this sample size chosen to exactly match that of Experiments 1a and 1b).

3.1.2. Stimuli

We used the same 48 Intact and 48 Phase-scrambled images used as Short Term Targets in Experiment 1a as the Long Term Targets in Experiment 2. These 48 Long Term Target images were divided into two sets (Target Set One and Target Set Two) based on the memorability scores of the Intact images, with each set of 24 images having an equal distribution of memorability scores which spanned the full range of the original scores (as measured by Isola et al., 2014). For each observer, one set (determined randomly, and counterbalanced across observers) served as targets and the other served as foils. In addition to the 48 Long Term Target images, 24 Short Term Target images were randomly sampled from the 48 filler images used in Experiment 1a.

3.1.3. Procedure

Each observer completed both a Short Term block and a Long Term block (see Fig. 7B). In the Short Term block, observers saw a sequence of images (each presented for 1.2 s, separated by a blank 1.2 s delay) and responded (by pressing 'r') whenever they saw a repeated image. Observers thus viewed 72 images (with 24 Long Term Targets, 24 Short Term Targets, and 24 Repeated Short Term Targets), all in a different randomized order, with the constraint that the Short Term Targets repeated after 1–6 images, with 4 targets repeated at each delay. In the subsequent Long Term block, observers indicated (using the same response key) whether any of the images had been presented earlier in the Short Term block; all 24 Long Term Targets were presented, along with 24 new foil images — all presented in a randomized order (again each presented for 1.2 s, separated by a blank 1.2 s delay).

3.2. Results

3.2.1. Memory performance

Before exploring memorability, we again assessed the overall fidelity of memory for the images. The d' values for both Short Term and Long Term Targets in all conditions are depicted in Fig. 8A. Inspection of this figure suggests that there was reliable memory for the images in all conditions, but (unsurprisingly) that memory was again better for the Intact images. These impressions were verified by the following analyses.

Short Term Targets were remembered above chance for both image types (Intact: $t(155) = 28.85$, $p < .001$, $d = 3.28$, all tests in this experiment were compared against the Bonferroni corrected alpha level of 0.008[0.05/6]; Phase-scrambled: $t(155) = 25.95$, $p < .001$, $d = 2.95$). As expected, memory for Short Term Targets was better for Intact images than for Phase-scrambled images ($t(310) = 12.06$, $p < .001$, $d = 1.37$). These patterns were all replicated for the Long Term Targets, which were remembered above chance for both image types (Intact: $t(155) = 25.68$, $p < .001$, $d = 2.92$; Phase-scrambled: $t(155) = 12.44$, $p < .001$, $d = 1.41$). As expected, memory for Long Term Targets was better for Intact images than for Phase-scrambled images ($t(310) = 11.45$, $p < .001$, $d = 1.30$).

3.2.2. Memorability

To assess memorability, we again ran 1000 split-half iterations as described in Experiment 1. While the calculation of hit rates and false alarm rates for the Short Term Targets is straightforward, the calculation of these scores for Long Term Targets is a bit more nuanced. For each

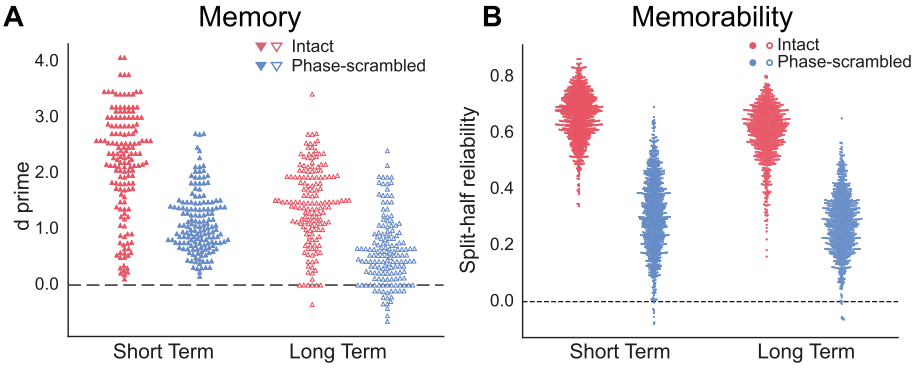

Fig. 8. Results of Experiment 2. (A) Short-term and long-term memory performance, where each triangle represents a participant; (B) Split-half reliability, both for Short Term and Long Term memorability scores, where each dot represents a Spearman rank correlation value of one split-half iteration.

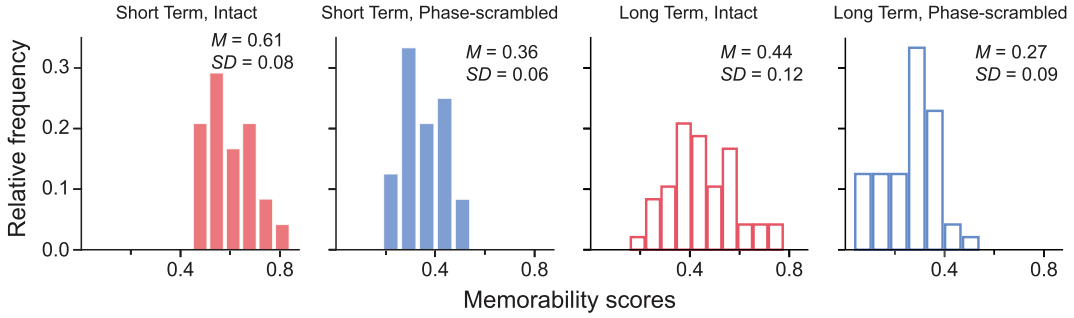

Fig. 9. Distributions of memorability scores for target images in Experiment 2.

Long Term Target, only half of the observers saw it as a target that appeared twice (once during the Short Term Block and once during the Long Term Block) while the other half saw it as a foil that appeared only once (during the Long Term Block). To determine the hit rate, we thus used the responses from observers who saw these Long Term Targets for the second time during the Long Term Block. However, there were two possible ways to calculate the false alarm rates: (a) a within-subject method, using responses from observers who saw the image as a target for the first time in the Short Term Block, and (b) a between-subjects method, using responses from observers who saw this image as a foil for the first time. Here we chose to use the within-subject method for the calculation of false alarm rates to better account for the response biases across individuals (e.g. participants who have a strong tendency to respond 'yes'). The Spearman rank correlations across the 1000 random splits are depicted in Fig. 8B. Inspection of this figure suggests that there was strong evidence for memorability across the board — in both Short Term and Long Term contexts, and for both Intact and Phase-scrambled images — with the key result here being the discovery of purely visual memorability (with the Phase-scrambled images) in a Long Term context. These impressions were verified by the following analyses.

For Short Term Targets, the averaged Spearman rank correlation was 0.66 for the Intact images (95% CI: [0.49, 0.82]). For the Phase-scrambled images, there was also reliable Short Term memorability (0.32, 95% CI: [0.04, 0.58]), thus replicating the demonstration of purely visual memorability in short-term memory for a third time (and despite a 50% reduction in the number of unique images). For Long Term Targets, the averaged Spearman rank correlation was 0.60 for the Intact images (95% CI: [0.37, 0.75]) — thus replicating the demonstration of Long Term memorability for the original images (Isola et al., 2014). Critically, there was also reliable Long Term memorability for the Phase-scrambled images (0.28, 95% CI: [0.08, 0.49]). The distributions of memorability scores defined over the entire sample are depicted in Fig. 9.

3.2.3. Intact vs. phase-scrambled long term memorability

The most forgettable and memorable images in each condition of Experiment 2 are presented in Figs. 5 and 6(the bottom row in each panel). Inspection of these images suggests again that the most memorable (and forgettable) Phase-scrambled images did not always correspond to the most memorable (and forgettable) Intact images. And this was statistically supported by the observation that the Long Term memorability scores for Intact images did not correlate with those of the Phase-scrambled versions of the same images ($\rho = -0.18$, $p = .217$).

3.2.4. The relationship between short term and long term memorability scores

There was a high correlation between the Long Term memorability scores (measured using the full sample in Experiment 2) and the Short Term memorability scores (measured using the full sample in Experiment 1a) for the Intact images ($\rho = 0.70$, $p < .001$). Similarly, there was a moderate correlation between Short Term and Long Term memorability scores with the Phase-scrambled images ($\rho = 0.36$, $p = .012$). The magnitudes of these correlations were similar across image types when we took into account the split-half consistencies as a measurement of the noise ceiling (relative degree = correlation between ST and LT memorability scores/mean of the average split-half correlations for ST and LT memorability scores; Intact: 1.13, Phase-scrambled 1.13).

To further investigate the relationship between the observed Short Term and Long Term memorability scores, we also re-ran the split-half analyses with one additional step: for each iteration, before correlating the two sets of target memorability scores based on the two halves, we first partialled out the *non-target* memorability scores (based on the full sample) from both sets. We then compared the resulting partial Spearman correlation with the Spearman correlation between the two sets of target memorability scores without controlling for the non-target memorability scores.

The results of this analysis are depicted in Fig. 10A and B. Inspection of these figures suggest two salient patterns. First, there were shared features that contribute to both the Short Term and Long Term

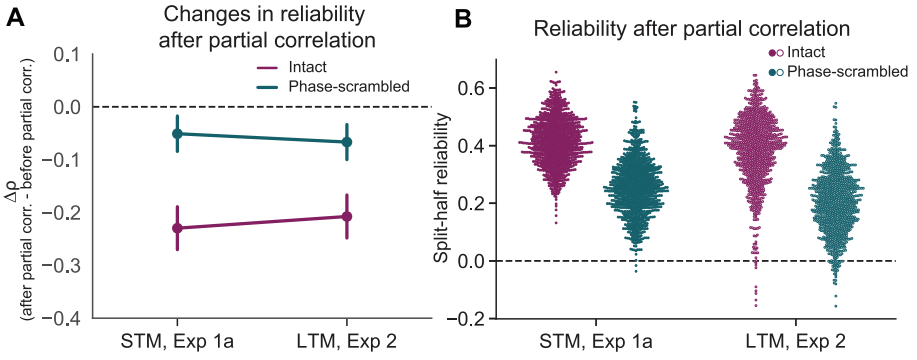

Fig. 10. Direct comparison between Short Term and Long Term memorability. (A) Changes in split-half reliability after controlling for the memorability scores from the non-target timescale (reliability after controlling for the memorability scores from the non-target timescale - reliability without controlling). Error bars represent the standard deviation of the corresponding change in correlation from the 1000 split-half iterations. (B) Split-half reliability after controlling for the memorability scores from the non-target timescale.

Table 1
Results from the split-half reliability analyses after controlling for the memorability scores from the non-target timescale.

		Changes in consistency compared to the original analyses without any control				Consistency after partial correlation		
Image Type	Timescale	Mean $\Delta\rho$	95 CI%	*P* (two-sided)	Prop. change	Mean ρ	95% CI	*P* (two-sided)
Intact	STM, Expt 1a	0.23	[0.16, 0.31]	.000***	36.63%	0.42	[0.26, 0.58]	.000***
Intact	LTM, Expt 2	0.21	[0.13, 0.29]	.000***	35.00%	0.39	[0.09, 0.58]	.018*
Phase-scrambled	STM, Expt 1a	0.05	[−0.02, 0.12]	.132	16.63%	0.26	[0.08, 0.45]	.004**
Phase-scrambled	LTM, Expt 2	0.07	[0.00, 0.13]	.044*	24.55%	0.21	[−0.01, 0.42]	.070^{-}

***: $p < .001$, **: $p < .01$, *: $p < .05$, ∼: $p < .1$.

memorability of Intact and Phase-scrambled images, respectively. Regressing out the non-target memorability scores led to a drop in the consistency of the memorability scores across all cases, although to a relatively smaller degree in the case of Short Term memorability for the Phase-scrambled images (Fig. 10A and Table 1; ST-Intact: $\Delta\rho = 0.23$, $p < .001$; LT-Intact: $\Delta\rho = 0.21$, $p < .001$; ST-Phase-scrambled $\Delta\rho = 0.05$, $p = .132$; LT-Phase-scrambled $\Delta\rho = 0.07$, $p = .044$). Second, there were also features uniquely associated with the memorability scores on each timescale for each image type — since despite the drop, the consistencies remained above zero across all cases after the partial correlation (Fig. 10B and Table 1; ST-Intact: mean $\rho = 0.42$, $p < .001$; LT-Intact: mean $\rho = 0.39$, $p = .018$; ST-Phase-scrambled mean $\rho = 0.26$, $p = .004$; LT-Phase-scrambled mean $\rho = 0.21$, $p = .070$).

3.3. Discussion

The results of this experiment replicated all of the essential patterns from Experiment 1, now in a Long Term context: we observed purely visual memorability in long-term memory, but (as with the Short Term memorability in Experiment 1) there was again no correlation between the memorability scores for the Phase-scrambled vs. Intact images — again suggesting that the features supporting purely visual memorability in Phase-scrambled images are distinct from those which drive memorability in the Intact images. In addition, the partial correlation analyses suggested that although there are shared features that were preserved across time and thus explained memorability at both Short Term and Long Term timescales, there may also be transformations that occur across time such that there is unique variance in the memorability scores of the same images at each timescale (which might happen, for example, if memories were more abstracted at a longer timescale).

4. Experiment 3: Does phase-scrambling truly remove semantic content?

We introduced the phase scrambling procedure by noting that "these manipulations severely constrain or even eliminate the meanings of the scenes that are so apparent in the intact images". But while this is indeed the entire reason why phase scrambling is typically used (e.g. Park, Brady, Greene, & Oliva, 2011), this was only supported in the current context by our informal intuitions (which perhaps readers who view our figures might share). But could the purely visual memorability for Phase-scrambled images observed in Experiments 1 and 2 somehow depend on subtle remaining traces of semantic content? We must admit that this is a possibility, especially given the recent demonstration of reliable correlations between certain low-level visual features and higher-level properties (e.g. Kardan et al., 2015; Long et al., 2016).

So to rule this out empirically, an independent group of observers were shown each scrambled image and were simply asked to guess what each original image had depicted. We were interested first in whether observers could indeed correctly guess the original image identities. But more importantly, we were also interested in whether performance on this task is statistically related to the Short Term memorability scores observed in Experiment 1 and the Long Term memorability scores observed in Experiment 2. (Even if people were not able to correctly identify what was in the image, for example, they still may have been 'wrong' in a systematic way — e.g. thinking incorrectly that a certain image in our set had originally been an angry face — which could relate to the memorability scores.)

4.1. Method

4.1.1. Participants

After exclusion (because of incomplete data, answers that were numbers rather than words, or identical responses to over 75% of the images), a final sample of 60 observers were recruited using Amazon Mechanical Turk. Each observer viewed either the Phase-scrambled images from Experiment 1a/2 or those from Experiment 1b. There were thus 30 unique observers in each group, with this sample size chosen before data collection began to exactly match the sample size from prior related work (Long et al., 2016). All observers gave informed consent, and were compensated for their time.

4.1.2. Stimuli

Images were drawn from the two sets of 48 Phase-scrambled images used in Experiments 1a/2 and 1b. Each image was presented in the center of the observer's browser window, on a solid white background, surrounded by a 4-pixel solid black border.

4.1.3. Procedure

Observers were told that they were going to see scrambled versions of some images and their task was to indicate, to the best of their ability, what the original image was — "limited to a word or two". We also gave some example answers (e.g. "bird", "river", "hat", "microwave", "basketball court", or "dog") to cover a wide range of possible categories. They completed 4 practice trials (with Phase-scrambled filler images) before labeling the 48 target images presented in randomized order. On each trial, observers saw a Phase-scrambled image, and then typed their answer in a text box below the image with no time limit.

4.1.4. Label coding

For each image set, two independent raters coded the responses as correct or incorrect. Specifically, the coders were shown the original Intact image, the ground truth labels from the SUN dataset, and the responses MTurk workers provided, and they indicated whether the responses could be used to describe the original image. The raters were quite consistent (agreeing on >90% of the answers for both image sets).

4.2. Results

4.2.1. Correct labeling and memorability

For each image, we calculated a correct labeling score as the percentage of responses rated as correct, with a generous threshold: as long as a response was rated as correct by either of the two raters, the response was deemed correct. Fig. 11A and C depict the distributions of the correct labeling scores for Phase-scrambled images from Experiments 1a/2 and 1b. For both image sets, people were unable to identify most of the images, but there was some variance (Experiments 1a/2 image set: $M = 22.85\%$, $SD = 16.41\%$, [Range = 0%, 77%]; Experiment

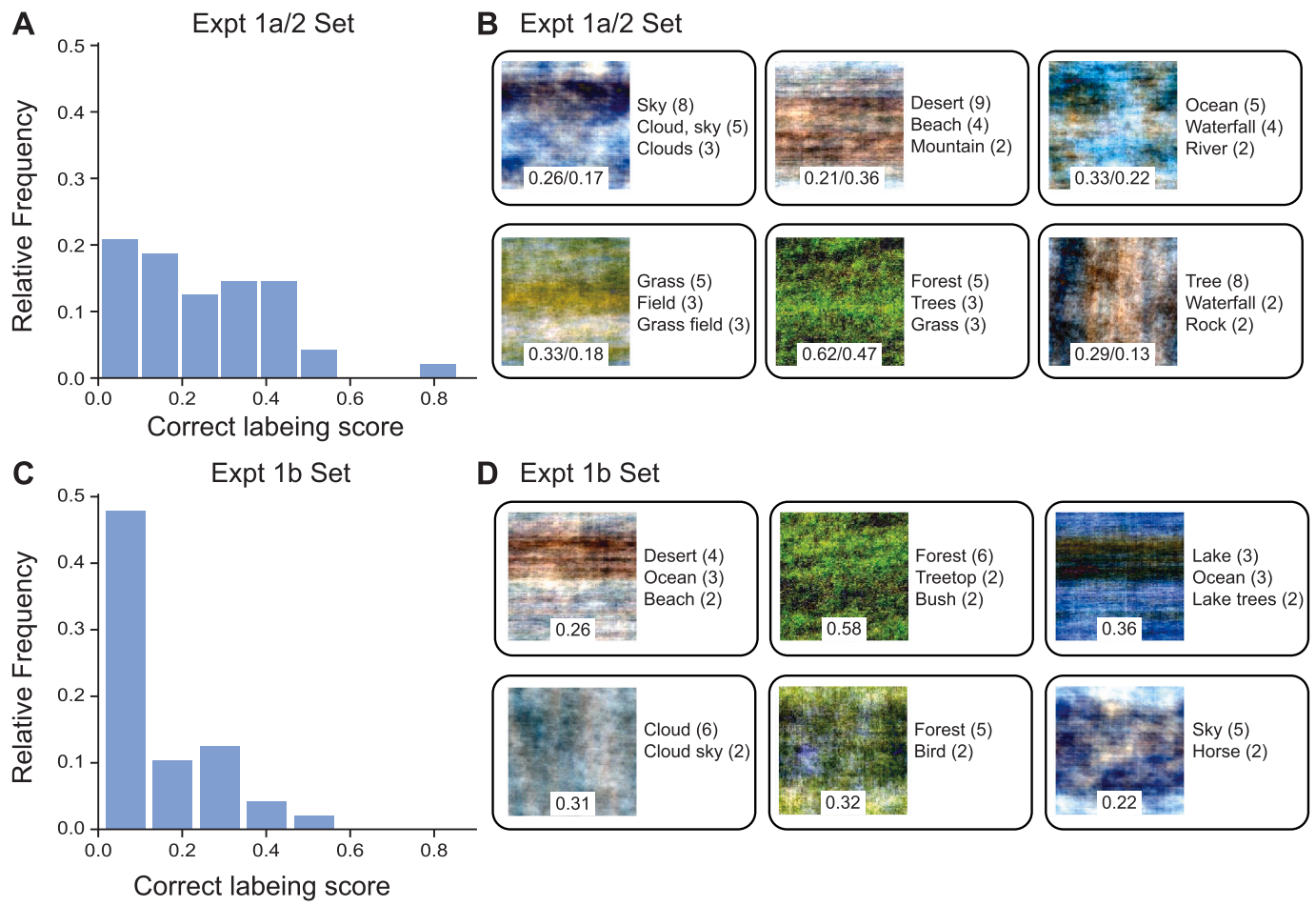

Fig. 11. Results from Experiment 3. (A) Distributions of correct labeling scores for Phase-scrambled target images from Experiments 1a/2. (B) Examples of the most consistently labelled images from Experiment 1a/2, with the top 3 most frequent labels (with counts >2) and the count of responses for each label in parentheses. The numbers at the bottom of each image are the Short Term/Long Term memorability scores. (C) Distributions of correct labeling scores for Phase-scrambled target images from Experiment 1b. (D) Examples of the most consistently labelled images from Experiment 2b, with the top 3 most frequent labels (with counts >2) and the count of responses for each label in parentheses. The numbers at the bottom of each image are the Short Term memorability scores.

1b image set: $M = 12.78\%$, $SD = 10.53\%$, [Range = 0%, 46.67%]).[2] Critically, however, the correct labeling scores were unrelated to either the Short Term memorability scores obtained in Experiments 1a and 1b ($\rho_{Expt1a} = -0.08$, $p_{Expt1a} = .606$; $\rho_{Expt1b} = 0.22$, $p_{Expt1b} = .136$) or the Long Term memorability scores obtained in Experiment 2 ($\rho_{Expt2} = 0.07$, $p_{Expt2} = .624$). However, there might be some component of the Long Term purely visual memorability scores that was driven by semantic features but was 'masked' by the components that were primarily driven by perceptual features. To rule out this possibility, we also measured the correlation between the correct labeling scores and the residuals in Long Term purely visual memorability after regressing out Short Term purely memorability scores for the same image set (as measured in Experiment 1a) and there was no relationship between them either ($\rho = 0.08$, $p = .586$). Collectively, these results suggested that even though people were able to correctly extract some semantic content for some images, the extent to which they were able to do so was not what fueled the purely visual memorability with such images.

4.2.2. *Consistent labeling and memorability*

We next asked whether the *consistency* in people's labeling (regardless of whether they were correct or not) predicted memorability. For each image, we first grouped similar responses together (via similarity scores >0.9 using the Ratcliff-Obershelp algorithm; Ratcliff & Metzener, 1988) — e.g. "snow mountain", "snowy mountains", and "snowy mountain") and we calculated a consistency score as the average counts of unique responses. Fig. 11B and D provide those images with the most consistent labeling across the two image sets. Although some images were given the same label by multiple people, consistency scores did not predict memorability scores ($\rho_{Expt1a} = 0.05$, $p_{Expt1a} = .716$; $\rho_{Expt1b} = 0.14$, $p_{Expt1b} = .337$; $\rho_{Expt2} = -0.11$, $p_{Expt2} = .459$; $\rho_{LTM\text{-}residuals} = -0.15$, $p_{LTM\text{-}residuals} = .312$). (We also explored other measures of consistency, including the single most frequent label count and the sum of the top 3 most frequent label counts. These measures all yielded very similar results when we correlated these consistency measures with memorability scores; all $ps > .276$.)

[2] In practice, it seemed unmysterious why the images with the highest correct labeling scores were identified relatively well. For example, the image with the highest score depicted a mountain, but the vast majority of the image depicted the sky; accordingly, the Phase-scrambled image also contained a large sky-colored region, and the most common response was "sky". In contrast, most scene categories do not have such a uniquely identifying color (here what we would commonly describe as "sky blue"). See Supplemental Figure 1 for examples of images with highest correct labeling scores.

4.3. Discussion

These results confirm our initial intuitions that the phase scrambling manipulation effectively eliminated any reliable, consistent semantic content from the original images that would be predictive of memorability.[3] And as such, these results thus suggest that semantic factors played very little if any role in the purely visual memorability observed in our study.

5. General discussion

The key result of this project is the novel demonstration of purely visual memorability across all three experiments (Experiments 1a, 1b, and 2) with different image sets, independent samples of observers, and different timescales. Whereas some previous work has explored the (relatively mild) contribution of various lower-level properties in intact images (e.g. Isola et al., 2014), the current data show for the first time that even in phase-scrambled images that have been largely stripped of their semantic content while preserving various perceptual features, some scrambled images are still intrinsically more likely to be maintained in both short-term and long-term memory across observers. Moreover, the data from Experiment 3 confirm that this effect is unrelated to any remaining semantic information (regardless of whether that information is correct or not).

5.1. Why might there be purely visual memorability in the first place?

It makes sense for us to remember some intact images better than others because of their differing ecological values: perhaps there is an advantage to remembering some locations more than others. (For example, it may be more important to remember the location where you saw a rattlesnake than to remember the location where you saw a rock; see Nairne & Pandeirada, 2016.) But why might people consistently remember some relatively meaningless blobs more than others, as with the phase-scrambled images?

Ultimately, memorability is a type of prioritization. And just like other forms of prioritization (e.g. regularities in what features grab our attention in crowded scenes), the purely visual memorability that we observed may reflect the fact that not all information is equally important, and that the prioritization process happens relatively automatically to any new information encountered as our minds are unable to remember it all. And if in fact memorable scenes are in general somehow more important, then it might be adaptive to prioritize this information as early as possible during online processing — including at early visual stages, when semantic information is not yet easy to detect.

This possibility is consistent with the recent observation that memorable images are categorized more accurately in an RSVP stream at durations as short as 13 ms (Broers et al., 2017; also see Goetschalckx, Moors, Vanmarcke, & Wagemans, 2019 who found memorability scores and categorizability at 33 ms were positively correlated after controlling for distinctiveness), suggesting that memorable and forgettable stimuli diverge extremely early during visual processing. However, it is difficult to draw any solid conclusions about purely visual vs. semantic information from such results — given that certain low-level visual features reliably signal higher-level properties (e.g. Kardan et al., 2015). For example, it turns out that large vs. small objects (e.g. cars vs. cups) have different perceptual features that can be extracted in early visual processing even when the sizes of their respective images are equated (Long et al., 2016). Similarly, maybe certain memorable images are categorized more accurately even from brief presentations not because of their visual features *per se*, but rather because of how those visual features reliably signal higher-level semantic properties. In contrast, the present study avoided such ambiguity by severely disrupting semantic information via image scrambling, leaving *only* certain lower-level visual properties intact. Taken together with the finding in Experiment 3 (that neither Short Term nor Long Term purely visual memorability can be explained by semantic factors), our study provides direct evidence that memorability can be driven by differences in perceptual features alone, and that the mind prioritizes some information in memory over others even when it can't assign semantic meaning to that information.

5.2. Conclusion

The current study contributes to a growing literature on the nature of memorability in two primary ways. Methodologically, it introduces image scrambling as a new way of revealing the types of information that underlie this form of prioritization. And theoretically, it suggests that memorability may not be driven only by one type of information (such as semantic associations); rather, the memorability of images may be a function not only of what they mean, but also of how they look.

Acknowledgements

For helpful conversation, we thank Wilma Bainbridge, the members of the Yale Perception and Cognition Laboratory, the Yale Visual Cognitive Neuroscience Lab, and the Turk-Browne Lab. We would also like to thank three anonymous reviewers for their constructive feedback. This project was funded by ONR MURI #N00014-16-1-2007 awarded to BJS and by an NSF Graduate Research Fellowship awarded to SRY.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2021.104714.

References

Baddeley, A. D. (1966a). The influence of acoustic and semantic similarity on long-term memory for word sequences. *The Quarterly Journal of Experimental Psychology, 18*, 302–309.

Baddeley, A. D. (1966b). Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *The Quarterly Journal of Experimental Psychology, 18*, 362–365.

Bainbridge, W., Hall, E., & Baker, C. (2019). Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory. *Nature Communications, 10*, 1–13.

Bainbridge, W. A. (2019). Memorability: How what we see influences what we remember. In D. M. Beck, & K. D. Federmeier (Eds.), *Knowledge and Vision* (pp. 1–27). Academic Press.

Bainbridge, W. A., Dilks, D. D., & Oliva, A. (2017). Memorability: A stimulus-driven perceptual neural signature distinctive from memory. *NeuroImage, 149*, 147–152.

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General, 142*, 1323–1334.

Bex, P. J. (2010). (In) Sensitivity to spatial distortion in natural scenes. *Journal of Vision, 10*, 23.

Broers, N., Potter, M. C., & Nieuwenstein, M. R. (2017). Enhanced recognition of memorable pictures in ultra-fast RSVP. *Psychonomic Bulletin & Review, 25*, 1080–1086.

Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., & Oliva, A. (2015). Intrinsic and extrinsic effects on image memorability. *Vision Research, 116*, 165–178.

Cahill, L., & McGaugh, J. L. (1998). Mechanisms of emotional arousal and lasting declarative memory. *Trends in Neurosciences, 21*, 294–299.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87–114.

Goetschalckx, L., Moors, P., Vanmarcke, S., & Wagemans, J. (2019). Get the picture? Goodness of image organization contributes to image memorability. *Journal of Cognition, 2*, 1–27.

Goetschalckx, L., Moors, P., & Wagemans, J. (2017). Image memorability across longer time intervals. *Memory, 26*, 581–588.

[3] Despite the fact that this result confirmed our intuitions vis-a-vis the Phase-scrambled images, we remain convinced that it was critical to verify this empirically — especially because this may be specific to phase scrambling (cf. footnote 1). Indeed, in other data not reported here, we have shown that observers' responses to a different type of 'texture-scrambled' image (e.g. Portilla & Simoncelli, 2000) *are* correlated with the memorability scores for those images — thus rendering that type of image scrambling unable to address our key questions about purely visual memorability in the first place.

Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2014). What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36*, 1469–1482.

Kane, M. J., & Engle, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 336–358.

Kardan, O., Demiralp, E., Hout, M. C., Hunter, M. R., Karimi, H., Hanayik, T., … Berman, M. G. (2015). Is the preference of natural versus man-made scenes driven by bottom–up processing of the visual features of nature? *Frontiers in Psychology, 6*, 471.

Khosla, A., Raju, A. S., Torralba, A., & Oliva, A. (2015). Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2390–2398).

Khosla, A., Xiao, J., Torralba, A., & Oliva, A. (2012). Memorability of image regions. In *Advances in Neural Information Processing Systems* (pp. 296–304).

Konkle, T., Brady, T. F., & Alvarez, G. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General, 139*, 558–578.

Long, B., Konkle, T., Cohen, M. A., & Alvarez, G. A. (2016). Mid-level perceptual features distinguish objects of different real-world sizes. *Journal of Experimental Psychology: General, 145*, 95–109.

Lukavský, J., & Děchtěrenko, F. (2017). Visual properties and memorising scenes: Effects of image-space sparseness and uniformity. *Attention, Perception, & Psychophysics, 79*, 2044–2054.

Miller, G. A. (1994). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 101*, 343.

Nairne, J. S., & Pandeirada, J. N. S. (2016). Adaptive memory: The evolutionary significance of survival processing. *Perspectives on Psychological Science, 11*, 496–511.

Oppenheim, A. V., & Lim, J. S. (1981). The importance of phase in signals. In *, 69. Proceedings of the IEEE* (pp. 529–541).

Park, S., Brady, T. F., Greene, M. R., & Oliva, A. (2011). Disentangling scene content from spatial boundary: Complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. *Journal of Neuroscience, 31*, 1333–1340.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision, 40*, 49–70.

Prins, N. (2007). Phase spectrum scrambling [electronic mailing list message]. Adapted from code retrieved from the visionscience.com, electronic mailing list: http://visionscience.com/pipermail/visionlist/2007/002181.html.

Ratcliff, J. W., & Metzener, D. E. (1988). Pattern matching: The gestalt approach. *Dr Dobbs Journal, 13*, 46.

Rossion, B., & Caharel, S. (2011). ERP evidence for the speed of face categorization in the human brain: Disentangling the contribution of low-level visual cues from face perception. *Vision Research, 51*, 1297–1311.

Stojanoski, B., & Cusack, R. (2014). Time to wave good-bye to phase scrambling: Creating controlled scrambled images using diffeomorphic transformations. *Journal of Vision, 14*, 1–16.

Thomson, M. G. (1999). Visual coding and the phase structure of natural scenes. *Network: Computation in Neural Systems, 10*, 123–132.

Vogt, S., & Magnussen, S. (2007). Long-term memory for 400 pictures on a common theme. *Experimental Psychology, 54*, 298–303.

Yoonessi, A., & Kingdom, F. A. (2008). Comparison of sensitivity to color changes in natural and phase-scrambled scenes. *Journal of the Optical Society of America A, 25*, 676–684.