Predictability in natural images determines V1 firing rates and synchronization: A deep neural network approach

Cem Uran[a,*], Alina Peter[a,*], Andreea Lazar[a], William Barnes[a,b], Johanna Klon-Lipok[a,b], Katharine A Shapcott[a,c], Rasmus Roese[a], Pascal Fries[a,d], Wolf Singer[a,b,c], Martin Vinck[a,e,]

[a]*Ernst Strüngmann Institute (ESI) for Neuroscience in Cooperation with Max Planck Society, Frankfurt, Germany*
[b]*Max Planck Institute for Brain Research, Frankfurt, Germany*
[c]*Frankfurt Institute for Advanced Studies, Frankfurt, Germany*
[d]*Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, Netherlands*
[e]*martin.vinck@esi-frankfurt.de*

Abstract

Feedforward deep neural networks for object recognition are a promising model of visual processing and can accurately predict firing-rate responses along the ventral stream. Yet, these networks have limitations as models of various aspects of cortical processing related to recurrent connectivity, including neuronal synchronization and the integration of sensory inputs with spatio-temporal context. We trained self-supervised, generative neural networks to predict small regions of natural images based on the spatial context (i.e. inpainting). Using these network predictions, we determined the spatial predictability of visual inputs into (macaque) V1 receptive fields (RFs), and distinguished low- from high-level predictability. Spatial predictability strongly modulated V1 activity, with distinct effects on firing rates and synchronization in gamma- (30-80Hz) and beta-bands (18-30Hz). Furthermore, firing rates, but not synchronization, were accurately predicted by a deep neural network for object recognition. Neural networks trained to specifically predict V1 gamma-band synchronization developed large, grating-like RFs in the deepest layer. These findings suggest complementary roles for firing rates and synchronization in self-supervised learning of natural-image statistics.

Keywords: Natural Images; Deep Learning; Machine Learning; Inpainting; Sparse Coding; Efficient Coding; Predictive Coding; Gamma; Beta; Oscillations; V1; Synchronization; Macaque; Awake; Electrophysiology.

Introduction

The ability to make accurate predictions is central to adaptive behavior in a complex world, and relies on the existence of strong regularities in both time and space. For example, when we look at a face, we expect to see a mouth below the nose. But, if we were to see a mouth above the eyebrows, we would be highly surprised and explore whether we had been tricked or discovered a new type of face. The brain can generate these kinds of prediction errors by comparing sensory inputs with the spatiotemporal context. Sensory prediction errors could play an important role in sensory learning (Rao and Ballard, 1999), analogous to the dopamine reward system (Schultz et al., 2017). In particular, they could be important for self-supervised learning, i.e. learning based on the data itself, without requiring human labels (Bengio et al., 2012; DiCarlo et al., 2012). Labeled data may be sparsely available for humans and other species, who are however excellent statistical learners throughout development and beyond (Santolin and Saffran, 2018; Krogh et al., 2013). Self-supervised learning is not yet accounted for by state-of-the-art models of the primate ventral stream like deep neural networks (DNNs) for object recognition (e.g. VGG-16), which conventionally rely on human labels for training. In addition, these DNNs do not account for the human ability to interpret sensory evidence in light of the spatiotemporal context: In contrast to humans, these networks effectively classify objects based on evidence accumulated from local features ("texture bias"), without an understanding of spatial relationships among elements (Geirhos et al., 2018)

In canonical feedforward models of visual processing, including DNNs, response properties of neurons are described based on their classical receptive field (RF). In area V1, neuronal responses are commonly characterized by small RFs encoding features like orientation and spatial frequency (Hubel and Wiesel, 1965). However, in natural images, stimuli are never presented in isolation, but are rather embedded in a spatiotemporal context with certain statistical regularities. Stimuli presented outside the classical RF can strongly modulate V1 responses (Kapadia et al., 1999; Spillmann et al., 2015; Gilbert, 1992; Coen-Cagli et al., 2015; Vinje and Gallant, 2000). In particular, V1 responses typically tend to be sparser for larger stimuli, in a way that depends on the homogeneity/heterogeneity of the image patch (Coen-Cagli et al., 2015; Vinje and Gallant, 2000). This sparsification has been linked to efficient and predictive coding principles (Rao and Ballard, 1999; Schwartz and Simoncelli, 2001; Coen-Cagli et al., 2015). The spatiotemporal context can also strongly influence the synchronization of neuronal activity across distributed columns: Vi-

*Equally contributing

sual stimuli can induce characteristic γ-oscillations (30-80Hz) that are synchronized between distant V1 populations with non-overlapping RFs (Gray et al., 1989). V1 γ-synchronization is strongly modulated by the size of stimuli (Gieselmann and Thiele, 2008) and we have proposed that it reflects the predictability of visual inputs given the spatiotemporal context (Vinck and Bosman, 2016; Peter et al., 2019). Synchronization is a common feature of cortical circuits with recurrent interactions between inhibitory and excitatory cells, and exerts powerful effects on synaptic plasticity and neural transmission (Sejnowski and Paulsen, 2006; Salinas and Sejnowski, 2001; Buzsáki, 2006; Buzsáki and Draguhn, 2004; Wang, 2010; Bressler et al., 1993). In sum, although feedforward models of the visual system may explain classical RF properties of V1 neurons, they poorly account for the contextual modulation and correlation structure of V1 activity, which likely result from both horizontal and top-down feedback (Adesnik et al., 2012; Angelucci et al., 2017; Vinck and Bosman, 2016).

Quantitative models based on DNNs for object recognition have been highly successful in modelling the RFs of neurons in visual cortex (DiCarlo et al., 2012; Cadena et al., 2019). Yet, similar quantitative approaches for modelling the contextual modulation of firing rates and neural dynamics are lacking. Importantly, such a network model should not be limited to specific elementary features like orientation, texture or color, but explicitly learn image statistics in natural scenes. Previous approaches for quantifying contextual properties in natural images were restricted to assessing homogeneity/heterogeneity in a binary way and limited to detecting specific features using Gabor-like RFs (Coen-Cagli et al., 2015, 2012). Here, we directly started out from the notion of spatial prediction and developed a predictive, deep neural network to generate predictions of image content into a neuron's RF based on the spatial context, i.e. the rest of the image. The neural network was trained in a self-supervised manner, without needing categorical labels or feature specification. Based on these image predictions, we derived measures of spatial predictability and distinguished low- and high-level predictability. We then presented natural images to awake monkeys while recording V1 activity, and analyzed how predictability modulates firing rates and synchronization patterns. We compared these analyses with two previously used approaches: (1) Predicting neural activity from the activation of artificial neurons (ANs) in a deep neural network for object recognition; (2) training a deep neural network *de novo* to predict neural activity from the image and characterizing the resulting computations by generating maximally driving images. Preliminary results of this work have been presented previously at the FENS 2018 and Bernstein 2018 meeting.

Results

We recorded multi-unit (MU) spiking activity and local field potentials (LFP) from 32-64 channel semi-chronic microelectrode arrays in V1 of three macaque monkeys performing a passive fixation task (Figure 1A). After a pre-stimulus period, we presented a natural image for 1.2s (in a subset of sessions, for 0.6s). Natural images subtended 11.5 to 15 degrees of visual angle (dva). In addition, RFs were mapped with moving bar stimuli. RF eccentricities ranged between 2.5 and 10.6dva (average 5.2dva). The average RF diameter was 1.44dva (note that this was likely overestimated because of small eye movements).

Spatial predictability in natural images

We developed a new method to quantify spatial predictability of visual inputs to the RF in an unsupervised way, and without specification of features like orientation or spatial frequency (Figure 1B, see Methods): We define spatial predictability as the extent to which the RF inputs can be predicted from the rest of the image. We refer to the part of the image covering a neuron's RF as the "RF image-patch". To quantify spatial predictability, we trained a deep neural network with a U-net architecture. The input to this neural network was the embedding context, i.e. the image region surrounding the RF image-patch (224x224 pixels; approximately 5-6dva wide). The network was then trained to generate a predicted RF image-patch based on this contextual input, a form of "inpainting" (see Methods). The size of the generated output was set to 64x64 pixels, roughly corresponding to the size of the neuron's RFs (see Methods). As a loss function for training, we used a measure of perceptual similarity between the input (ground-truth) RF image-patch and the predicted RF image-patch (see Methods). The neural network was trained using a large number of natural images that were *not* presented during the recordings. Training was self-supervised, in the sense that only data from the image itself was used for training, without human labels.

Once the neural network was trained, we used it to quantify spatial predictability for the images presented during the recordings. For this, we used various measures of perceptual similarity to compare the network prediction to the ground-truth RF image-patch. In other words, we compared two quantities using perceptual similarity measures: (1) The *actual* visual input to the neurons; and (2) the visual input that was *predicted* based on the rest of the image. To this end, we computed the pixel-by-pixel correlations between the ground-truth and the predicted RF image-patch. We call this the "structural predictability", a measure comparable to the well-known perceptual-similarity measure SSIM (see Methods). Further below, we will consider other measures of spatial predictability based on DNNs for object recognition (Figure 4).

How does spatial predictability determine distinct aspects of neural activity?

Neural activity can be separated into two parts, namely output (spikes) and input signals (synaptic potentials; both excitatory and inhibitory). In extracellular recordings, these can be distinguished with specific band-pass filters on the recorded electric signals (see Methods). This yields measures of spiking activity and the LFP, a measure of population synaptic activity (Pesaran et al. (2018), Figure 2A-B).

We first examined the relationship between spatial predictability and neuronal synchronization. To detect synchronized activity in the neuronal population, we computed LFP

power spectra. These showed a typical $1/f$ trend, which reflects, at least in part, low-pass filtering of synaptic currents (Figure 2A), Pesaran et al. (2018)). In addition, LFP spectra contained two visually-modulated spectral peaks in so-called (high) β (18-30Hz) and γ (30-80Hz) frequency bands, indicating synchronization of synaptic activity (Buzsáki and Draguhn (2004); Pesaran et al. (2018), Figure 2A). We estimated the magnitude of these peaks by removing the $1/f$ trend and fitting polynomials (see Methods). Spatial predictability was computed for each image and recording site as detailed above (Figure 1, see Methods). We then computed LFP spectra for 32 quantiles of spatial predictability (Figure 2A). We found that LFP spectra differed markedly as a function of spatial predictability (Figure 2A): For images with higher spatial predictability, LFP power spectra showed a narrow-band peak in the γ-frequency range, which was several times larger than the pre-stimulus power (Figure 2A). However, for images with low spatial predictability, LFP spectra did not show γ-peaks (Figure 2A). From the lowest to highest levels of spatial predictability, γ-synchronization increased monotonically and by approximately 300% (Figure 2A,C). This finding was consistent across the three monkeys (monkey H, I and A: Pearson's r=0.94, 0.85, 0.87). Furthermore, the occurrence of γ required that there was a RF "surround", i.e. that the natural image extended beyond 1dva (Figure Extended Data 1), consistent with previous work (Gieselmann and Thiele, 2008; Peter et al., 2019). Finally, in congruence with previous work, we found that γ-synchronization was reflected in rhythmic spiking activity (Figure Extended Data 2). In an additional analysis, we directly correlated LFP γ-power with predictability across all images, without using quantiles (see Methods). This analysis showed a positive correlation between LFP γ-synchronization and spatial predictability (Figure 2D).

Unexpectedly, we observed that at the lowest levels of spatial predictability, LFP spectra showed a prominent peak in the (high) β-frequency band (18-30Hz) range (Figure 2A). Activity in this frequency band has previously been linked to cortical feedback and tends to be more prominent in higher brain regions like parietal cortex (Bressler et al., 2007; Buschman and Miller, 2007; Bastos et al., 2015). We found that β-band activity showed a monotonic decrease as a function of spatial predictability (Figure 2C,D). LFP β-peaks were accompanied by beta-rhythmic firing in V1 multi-unit (MU) activity (Figure Extended Data 2). Note that LFP β-peaks were detected in two animals and were negatively related to predictability in both (Monkey H: Pearson's r=-0.79; Monkey I: r=-0.54). In a third monkey, β-peaks were not detected.

Next, we examined the correlation between firing-rate intensity (Figure 2B-D) and spatial predictability. Firing-rate intensity was computed both for early (50-150ms) and late stimulus periods (200-600ms), by determining the instantaneous energy of the MU activity (see Methods). Early MU firing-rate intensity showed a weak, negative correlation, but late firing-rates were not significantly correlated with spatial predictability (Figure 2B-D). The relationship between firing intensity and spatial predictability was non-monotonic, such that MU firing rates peaked at intermediate levels of spatial predictability (Figure 2C). These findings on rates were surprising, given that predictive coding models postulate that V1 firing rates encode spatial prediction errors (Rao and Ballard, 1999). Further below, we will show that V1 rates correlate with different aspects of predictability (Figure 4).

Our findings indicate that V1 synchronization is determined by spatial predictability. Consequently, synchronization may depend on the image category: Images that contain man-made objects typically contain more predictable structure (e.g. buildings) than pictures of nature (e.g. bushes; see Figure 2E for example images; Figure Extended Data 3). We found that β-synchronization was relatively strong for pictures of nature, whereas γ-synchronization was relatively strong for pictures with man-made contents in the RFs (Figure 2F). Likewise, we predicted that γ and β should distinguish between image patches with and without an object boundary, because object boundaries will typically have a predictable, edge-like structure. Indeed, γ and β were respectively strong and weak when RFs covered an object boundary (Figure 2G; see Methods). By contrast, firing rates did not differ between nature and man-made pictures, and did not depend on the presence of an object boundary (Figure 2F-G).

There has been substantial discussion about the prevalence of γ-synchronization for natural images (Hermes et al., 2015; Brunet et al., 2015; Brunet and Fries, 2019; Brunet et al., 2014; Bartoli et al., 2019); differences between studies could easily be caused by the selection of images (e.g. hand-picked or random, image category). For randomly-selected natural images, which only fulfilled some global criteria on luminance and contrast (see Methods), we observed that spatial predictability had a probability distribution close to uniform (Figure 2H; see also Figure Extended Data 3). This predicts substantial variation in γ and β across images. Indeed, γ-peaks were detected only for about 50% of sites, both for nature and man-made categories, wheras β-peaks were primarily found for the nature category (Figure 2H).

The analyses above were performed for grayscale images. Yet, it has been shown that color can modify both γ-synchronization and neuronal firing rates (Shirhatti and Ray, 2018; Peter et al., 2019; Wachtler et al., 2003; Rols et al., 2001). In a separate set of sessions, we therefore presented a set of images both in grayscale and in their original color. We found that across images, γ and β-synchronization were strongly correlated between grayscale and color images (Pearson's r=0.81 and 0.7, respectively, $P < 0.001$ for both). Thus, the relationships between β/γ and spatial predictability at the grayscale level explains variance in β/γ for natural color images.

For the rest of the paper we focus primarily on firing rates and gamma, because 18-30Hz β-peaks were detected in only 2 out of 3 monkeys and occurred only for a small subset of images. For β, we report the analyses that are shown below in Figure Extended Data 4.

Interaction with stimulus drive

Next, we wondered how other image parameters like luminance contrast influence synchronization and firing rates,

and how these factors interact with spatial predictability (Figure 3A). In natural scenes, luminance contrast varies across image regions and with lighting conditions (see Figure Extended Data 3). Previous work has shown that under low-contrast conditions, the classical RF (cRF) shows a relative expansion, indicating that surround modulation depends on luminance contrast (Sceniak et al., 1999; Kapadia et al., 1999; Cavanaugh et al., 2002).

For each recording site, we computed the luminance contrast of its corresponding RF image-patch (as RMS contrast; see Methods). Firing rates showed a characteristic increase as a function of luminance contrast with saturation around 20% (Figure 3B-C). Consistent with previous work (Henrie and Shapley, 2005; Ray and Maunsell, 2010; Hadjipapas et al., 2015), we observed that luminance contrast positively correlated with γ-synchronization (Figure 3B-E; for β see Figure Extended Data 4). The γ peak frequency showed a moderate increase with contrast (Figure 3C).

Because γ-synchronization was positively correlated both with spatial predictability and luminance contrast, we wondered how these variables interacted. We have previously postulated that spatial predictability and luminance contrast should interact in a multiplicative way, based on the idea that inputs to V1 neurons become more predictable under conditions of high signal-to-noise (Peter et al., 2019). To investigate this, we fitted a multiple linear regression model. In this model, the predictors were spatial predictability (Figure 2), luminance contrast and their interaction. This full regression model explained substantially more variance in γ-synchronization as compared to luminance contrast or predictability alone (Figure 3D). We found a significant interaction between luminance contrast and predictability, with a higher T-statistic than for the individual variables (mean T-statistic contrast -1.42± 0.89, $p > 0.05$, T-test; predictability -0.84±0.78, $p > 0.05$; predictability × contrast 4.2 ± 1.2, $p < 0.001$). The interaction term of luminance contrast and predictability explained almost as much variance as the full regression model (Figure 3D). This multiplicative effect is consistent with the absence of γ-synchronization for stimuli with low contrast, small stimuli, or stimuli with low spatial predictability (Figure 3E, Figure Extended Data 1; see also Figure 4D). In contrast to γ-synchronization, we found that firing rates were mainly predicted by luminance contrast (Figure 3D).

Next, we examined whether other stimulus factors, namely spatial frequency, stimulus orientation and luminance explained further variance (see Methods). These variables were computed both for the 64x64 center patch and the 224x224 image region. Including these additional predictors only slightly increased the explained variance for γ-synchronization (Figure 3D). However, including these factors had a very strong effect on explained variance for firing rates, consistent with the Gabor-like RF properties of V1 neurons (Figure 3D).

High and low-level predictability

In the above analyses (Figure 1-2), we quantified predictability by computing the pixel-by-pixel correlations between the ground-truth and the predicted RF image-patch. This definition of predictability primarily reflects the low-level structure in an image. However, neurons might in principle encode sensory predictions or prediction errors of varying levels of complexity, ranging from low-level content (e.g. pixel values) to high-level semantic information.

To distinguish different levels of predictability, we used an established model of firing-rate representations in the visual system, namely the VGG-16, a deep neural network for object recognition (Cadena et al. (2019), Figure 4A). The VGG-16 consists of multiple layers. Each layer contains artificial neurons (ANs) at different locations, and at each location there are ANs with multiple feature-selectivities (in 3D: layers × features × locations) (Figure 4A). In such a network, the activity of ANs in early and deep layers reflects low and high-level content, respectively (Gatys et al., 2015). For each AN in the VGG-16, we computed the activation for the original image and the predicted image (see Methods; Figure 4A)). Based on the comparison between these activations, we computed a predictability measure for each layer of the VGG-16, using two kinds of functions: Structural and content predictability (Figure 4A). Layer-specific content-predictability was defined as the negative of content loss (Gatys et al., 2015), which is the average Euclidean distance between the AN's activation for the ground-truth and predicted image-patch. Layer-specific structural-predictability was defined in two steps: (1) For each VGG-16 AN with a given feature selectivity, we computed the Pearson correlation across locations; (2) we averaged these correlations across ANs with different feature selectivities (see Methods). The main difference between these two kinds of predictability is that structural predictability is determined by the spatial pattern of activation, whereas content predictability compares the activation strength of each VGG-16 AN between predicted and actual input.

We found that for both measures, the correlation between predictability and V1 firing rates became increasingly negative towards the deeper layers of the VGG-16 (Figure 4B,C). Furthermore, firing rates were better predicted from content predictability than structural predictability (Figure 4B). Thus, unexpectedly, spatial predictability in terms of high-level content predictability was the most reliable predictor of V1 firing rates (Figure 4B-C; see Discussion).

We observed the opposite pattern for γ-synchronization (for β see Figure Extended Data 4): γ was strongly and positively correlated with the structural predictability in the input layer and the early layers of the VGG-16 (Figure 4B). However, it was weakly correlated with content and structural predictability in the deep layers of the VGG-16. Accordingly, the correlation with spatial predictability decreased across layers (Figure 4B). Furthermore, correlations with content predictability were substantially weaker than with structural predictability (Figure 4B). Thus, including structural and content predictability from deeper layers did not improve the prediction of γ-synchronization (Figure 4C).

In Figure 2, we did not find a significant correlation between late firing rates and structural predictability, a measure comparable to the perceptual similarity measure SSIM. Recently, a perceptual similarity measure has been developed based on VGG-16 activations across all layers, LPIPS (Learned Perceptual Image Patch Similarity) (Zhang et al., 2018). LPIPS has

been reported to outperform other perceptual similarity measures like structural similarity (SSIM) (Zhang et al., 2018). Because firing rates showed strong correlations with predictability in deep layers of the VGG-16, we expected that they would also be correlated with predictability defined as LPIPS. Indeed, late rates were negatively correlated and monotonically related to LPIPS-predictability (Figure Extended Data 5). Furthermore, correlations between γ and LPIPS were comparable to correlations between γ and structural predictability, most likely reflecting the contribution of early VGG-16 layers to the LPIPS measure (Figure Extended Data 5).

We designed a further experiment to explicitly dissociate firing rates from γ-synchronization. We presented noise textures (brown, white, pink) in two conditions: Center-surround match (same noise in RF and surround) and mismatch (different noise type in RF and surround). Because all of these stimuli had low structural predictability, we did not expect to find any γ-synchronization. Indeed the spectra did not show γ peaks for any of the conditions (Figure 4D). However, these stimuli differed in terms of higher-level content predictability, considering that ANs in deeper layers of the VGG-16 have texture selectivity. Accordingly, we found that firing rates were substantially higher for texture mismatches than homogeneous textures (Figure 4D). Together, these findings demonstrate that firing rates and synchronization properties are modulated by distinct aspects of spatial predictability.

Predicting neural activity from a deep, feedforward neural network for object recognition

We found that neural activity showed consistent correlations with spatial predictability, which we quantified by training a self-supervised deep neural network to predict/inpaint missing image patches. We also used the VGG-16 to distinguish low from high-level spatial predictability. Next, we wished to contrast our approach with a previously used approach, in which the similarity of neural responses with the activity of ANs in a feedforward neural network for object recognition is directly assessed (in our case VGG-16). For example, because γ reflects predictable feature relationships across space, it might correlate well with the activity of VGG-16 ANs that have larger RFs and reside in deeper layers of the VGG-16. In that case, there might be nothing "special" about γ synchronization in comparison to activity of neurons in higher visual areas.

To investigate this, we performed additional analyses in which we linearly predicted neural activity from ANs in the VGG-16, similar to previous work for V1 firing rates (Cadena et al. (2019), Figure 5A). For all of the convolutional VGG-16 ANs, we computed the responses (i.e. activations) to each image that we presented during the recordings (Figure 5A). We then used sparse L1-regression (with cross-validation; see Methods) to predict neural activity for each recording site (Figure 5A). This was done separately for different VGG-16 layers. In a first regression analysis, we only used VGG-16 ANs with convolutional RFs that overlapped with the RFs of the to-be-predicted recording site (see Methods).

We found that late firing rates were relatively well predicted from the activation of VGG-16 ANs (Figure 5B); maximum R values around 0.45). Explained variance was highest for the middle layers of the VGG-16 (3conv1 to 4conv2), consistent with previous findings (Cadena et al. (2019), Figure 5B). This result makes sense, given that the convolution RF-size in 3conv3 is effectively 40x40 pixels, which corresponds to approximately 1dva in our experiments (see Table 1). Firing-rate intensity also showed significant correlations with the activity of VGG-16 ANs in the lower and the higher layers. For the lower layers, this likely reflects the influence of low-level features like luminance, contrast, etc. (Figure 5B).

For γ-synchronization, correlations were generally much lower than for firing rates (maximum around 0.09) (Figure 5B), and substantially lower in comparison to regression with spatial predictability and luminance contrast as predictors (see Figure 3). We observed that γ-synchronization was best predicted from the activation of ANs in the middle layers, peaking in VGG-16 layer 4conv1 (Figure 5B). Correlations were very low for the early VGG-16 layers (Figure 5B), consistent with the idea that gamma is predicted from regularities over larger regions of space.

In a second L1 regression analysis, we predicted neural activity by also including ANs with convolutional RFs that were *not* overlapping with the RFs of the recording site (i.e. a 3x3 degrees input patch including the surround). To determine by which locations in the image the MU firing rates were driven, we computed the absolute prediction weights as a function of the RF position of the VGG-16 ANs in the image. The resulting "heatmaps" were then aligned to the estimated center of the recorded MU. We found that, as expected, firing rates had small, circular RFs. These RF were visible even for the VGG-16 input layer (Figure 5C). As expected, spatial structure disappeared in the deeper layers of the neural network, because of the large convolutional RFs in those layers (Figure 5C). We performed the same L1 regression analysis for γ-synchronization. In contrast to firing rates, RFs for γ-synchronization showed a clear RF structure only towards the middle layers of the VGG-16 (Figure 5C; for β see Figure Extended Data 4).

A deep neural network that predicts γ-synchronization

In Figure 5, we found that γ-synchronization is poorly predicted by the activity of ANs in the VGG-16 network for object recognition. This suggests that γ-synchronization results from a different kind of computation that is not captured by the computations in the VGG-16 network for object recognition. We therefore trained a deep neural network *de novo*, gamma-net, in which we directly predicted γ-synchronization from the image without computation of any specific, elementary features (Figure 6A). Subsequently, we used "image synthesis" techniques to generate maximally driving images. This follows previous approaches used for firing rates in several species (Cadena et al., 2019; Walker et al., 2019; Lindsey et al., 2019; McIntosh et al., 2016). For the first layers of gamma-net, we used the pre-trained VGG-16 layers. This was done in order to avoid fitting a very large number of parameters. The network then consisted of three additional layers (two convolutional ones and one read-out layer) that were trained *de novo*. The objective loss function was the γ peak-amplitude.

Using this approach, we found that γ-synchronization could be well predicted from the input image (peak correlation around 0.64; Figure 6B). Highest explained variance was observed when using pre-trained layers up to the middle VGG-16 layers (Figure Extended Data 6). The correlation obtained with gamma-net was slightly higher than the correlation that was observed in the analysis above, where we predicted γ-synchronization from spatial predictability and luminance-contrast (Figure 4 and 6B). Note that some of the unexplained variance may have been due e.g. to factors like microsaccades, volume conduction and arousal, or estimation errors.

Next, we studied the RFs of the ANs in gamma-net. To this end, we generated images that were maximally driving the output of the network, or maximally driving the activity of specific ANs in the network (see Methods) (Figure 6C). We found that the RFs of ANs in the deepest layer of the network consistently had a grating-like structure, even though its training set did not include grating images. Furthermore, the maximally driving image of the network had a grating-like structure (Figure 6C). This is consistent with the finding that the most widely used stimulus in V1 neuroscience - a grating - tends to generate very strong γ-synchronization (Spyropoulos et al., 2020; Onorato et al., 2020).

To investigate whether gamma-net recapitulated the basic feature-selectivity of γ-oscillations, we presented stationary gratings to the network. We presented gratings of 16 different orientations and varied the size, spatial frequency and luminance contrast. We found that γ-synchronization emerged only for stimuli that extended beyond the cRF (Figure 6D), consistent with Figure Extended Data 1. Similar to what we found for predictability and luminance contrast (Figure 3E), there was a multiplicative dependence of γ-synchronization on size and luminance contrast. Finally, we observed that γ-synchronization was particularly strong for intermediate spatial frequencies (Figure 6D), consistent with previous work in macaques and humans (Jia et al., 2011; Adjamian et al., 2004a). Together, these findings indicate that γ-synchronization results from computations integrating over both cRF and surround inputs, and further support the idea that γ reflects center-surround predictability (Vinck and Bosman, 2016).

Discussion

Feedforward, deep neural networks for object recognition are a promising model of visual processing and can accurately predict firing-rate responses along the primate ventral stream. Yet, these networks currently have limitations as models of various aspects of cortical processing related to recurrent connectivity, including neuronal synchronization and the integration of sensory inputs with spatiotemporal context. Here, we trained self-supervised, generative neural networks to predict small regions of natural images based on the spatial context (i.e. inpainting). Using these network predictions, we determined the spatial predictability of visual inputs into (macaque) V1 receptive fields (RFs), and distinguished low- from high-level predictability. We compared these analyses with two other approaches previously used: (1) Predicting neural activity from the activation of artificial neurons (ANs) in a deep neural network for object recognition. (2) Training a deep neural network to predict neural activity from the image and characterizing the resulting computations by generating maximally driving images.

Our first main finding is that predictable contextual relationships in natural images are a key determinant of V1 synchronization (Vinck and Bosman, 2016). V1 γ-synchronization was originally discovered by (Gray et al., 1989), who expressed the key insight that it is modulated by feature relationships across space. Subsequently, this phenomenon was conceived of as a mechanism for perceptual binding based on psychological Gestalt principles (Singer and Gray, 1995; Von Der Malsburg, 1994), which was criticized on theoretical and experimental grounds by several groups (Roelfsema et al., 2004; Shadlen and Movshon, 1999; Palanca and DeAngelis, 2005; Lima et al., 2010). Recently, Vinck and Bosman (2016) proposed that γ-synchronization results from center-surround predictability and the mutual predictability of visual inputs at different RF locations ("synchronization through mutual predictability"; STMP). One of the motivations of this study was to test this hypothesis directly. Previous studies on contextual modulation of γ-synchronization were restricted to specific features like orientation, motion direction or color (Biederlack et al., 2006; Peter et al., 2019; Gray et al., 1989; Veit et al., 2017; Vinck and Bosman, 2016; Gieselmann and Thiele, 2008; Gail, 2000). The current study of contextual modulation has two principal strengths: (1) We used a large set of automatically selected natural images. (2) We used high-resolution RF mapping combined with a predictive neural network analysis. This approach was not restricted to any particular "human-defined" features and distinguished low- from high-level predictability. Using these methods, we found that strong, low-level predictability of visual inputs induces γ-synchronization, whereas weak predictability induces β-synchronization in V1.

Our second main finding is that firing rates and neural synchrony encode distinct aspects of spatial predictability. A previous study in anesthetized monkeys has shown that V1 firing rates are reduced for natural image regions that are spatially homogeneous, compared to heterogeneous image regions (Coen-Cagli et al., 2015). This homogeneity was quantified in a binary manner by examining responses of simple-cell Gabor functions in the RF center and the surround. Using a neural network approach, we defined a continuous measure of predictability, and distinguished low-level and high-level forms of spatial predictability. To our surprise, V1 firing rates were enhanced when RF inputs were poorly predicted at the level of *high-level content* representations (in VGG-16), but were very weakly correlated with spatial predictability at the level of pixel-by-pixel correlations in the input image. This suggest that V1 firing rates are modulated by the extent to which they can be predicted by higher-order visual areas. This notion is consistent with the observed selectivity of prediction errors for face identitythe inferior temporal lobe, which in lower areas reflects the feature selectivity of higher areas (Schwiedrzik and Freiwald, 2017; Issa et al., 2018). Together with the present study, these findings challenge the idea that sensory prediction-errors propagate in the *feedforward* direction, and are rather consistent with hierar-

chical error signaling (Issa et al., 2018; Heeger, 2017).

A third main finding is that γ-synchronization and firing rates carry unique and complementary information for natural scenes (see also (Peter et al., 2019)). Previous studies have found diverse correlations of γ-synchronization with firing rates (Jia et al., 2013; Vinck and Bosman, 2016; Gieselmann and Thiele, 2008; Ray et al., 2008; Ray and Maunsell, 2011). For instance, some studies have suggested that γ-synchronization reflects surround inhibition or sparse firing (Gieselmann and Thiele, 2008; Jadi and Sejnowski, 2014; Vinck and Bosman, 2016; Vinck et al., 2013). Yet, we found that γ and firing rate encoded distinct aspects of predictability and could be experimentally dissociated with noise stimuli, which argues against this interpretation. Whereas firing rates were negatively modulated by high-level predictability, γ-synchronization was positively correlated with predictability of *low-level structure*. Because these low-level image representations should correspond to inputs into V1, we conceptualize γ as an "input-predictability" signal. Thus, the more γ, the lower the dimensionality of the image region around the RF, and the more redundant the visual inputs into the neurons' RF (i.e. the less unique information provided by thalamic inputs). High-contrast grating stimuli are one of the most widely used visual stimuli in neuroscience, because they have a low dimensionality and are easy to parametrize. Yet, because of their low dimensionality and high contrast, these grating stimuli also induce unusually strong network oscillations that are not representative of "normal" V1 activity. In fact, grating stimuli are associated with discomfort and photoepilepsy (Adjamian et al., 2004b), and can induce γ-oscillations with dynamical system parameters close to criticality/network instability (Spyropoulos et al., 2020).

Previous work has associated γ and β-synchronization to feedforward and feedback processing (Bressler et al., 2007; van Kerkoerle et al., 2014; Richter et al., 2018; Bastos et al., 2015) (but see Schneider et al. (2020)). Based on predictive coding models, these frequency-bands were therefore postulated to reflect prediction errors and predictions, respectively (Bastos et al., 2012; Arnal et al., 2011; Bressler et al., 2007). For γ, we found the opposite result, at least in the domain of spatial predictions. Moreover, the inverse relationship of β-synchronization to spatial predictability seems *prima facie*, difficult to reconcile with β as a "carrier wave" for top-down predictions. In fact, one would expect strong β-synchronization if bottom-up activity can be explained away by top-down predictions. However, the opposite held true.

The function of cortical γ-synchronization has been a subject of debate (Ray and Maunsell, 2015; Hermes et al., 2015; Burns et al., 2011; Brunet et al., 2014), and the present findings have various implications for this debate. First, our findings support the hypothesis that γ-synchronization mediates contextual interactions over horizontal, patchy connections (Vinck and Bosman, 2016; Gray et al., 1989). In this scenario, the impact of asynchronous bottom-up inputs from Layer IV into Layer II/III would be diminished when they provide redundant information, whereas the impact of recurrent, horizontal and feedback inputs would be boosted by gamma-resonance (Vinck and Bosman, 2016). Thus, our findings suggest that γ plays a role in top-down and horizontal processing, rather than feedforward processing.

Second, because of their relationship to spatial predictability, γ-synchronization (and β) could be important for the unsupervised learning of spatiotemporal natural-image statistics. In predictive coding models, error signals are construed as *high-dimensional*, resulting from subtracting top-down predictions from bottom-up inputs (Rao and Ballard, 1999). Alternatively, error signals in recurrent neural networks can be encoded by a *low-dimensional* global loss, which can modify synaptic plasticity in interaction with high-dimensional Hebbian "eligibility" traces defined by ongoing activation patterns of neurons (Bellec et al., 2020). Oscillatory wave patterns are interesting candidates for global loss functions in recurrent networks because: (1) They are low-dimensional and can be distributed across neuronal populations; (2) they directly modulate synaptic plasticity (Wespatat et al., 2004; Sjöström et al., 2001; Sejnowski and Paulsen, 2006; Vinck et al., 2010); and (3) they can be superimposed onto a population rate vector without information loss, forming a dual coding scheme.

A function of γ-synchronization in synaptic plasticity is compatible with two properties that are difficult to reconcile with the putative roles of γ-oscillations in attention and feedforward inter-areal communication (Hermes et al., 2015; Akam and Kullmann, 2012; Ray and Maunsell, 2015): (1) γ- and β-activity is induced only by a subset of images and highly stimulus-dependent (Ray and Maunsell (2015); Hermes et al. (2015), Figure 2). Notably, we found that γ was poorly predicted by activity of ANs in the VGG-16 network for object recognition (Figure 5B). Thus, one does not expect enhanced feedforward processing for image regions that induce strong γ. In fact, RF regions with strong γ may typically contain little unique information (see also (Peter et al., 2019)). (2) γ-dynamics are well reproduced by a noise-driven dampened harmonic oscillator, forming bursts that are separated by prolonged periods of asynchronous firing (Burns et al., 2011; Spyropoulos et al., 2020). This stochastic variability requires very long integration windows for reliable selective communication (Akam and Kullmann, 2012). Yet, for synaptic plasticity, it is the gradient descent across an *ensemble* of images that matters. A sustained presence of synchronization for each individual image or time window is therefore not required. Instead, a short-lived burst of gamma oscillations may be sufficient to modify synaptic connections, analogous to hippocampal sharp-wave ripples and their purported function in plasticity and learning (Buzsáki, 1989). Furthermore, γ-synchronization may be an effective mechanism to induce plasticity formation due to spike-timing-dependent-plasticity mechanisms (Sjöström et al., 2001; Wespatat et al., 2004; Galuske et al., 2019). Interestingly, V1 γ-synchronization is associated with high-frequency burst-firing in a specialized cell-class (Gray and McCormick, 1996; Onorato et al., 2020), which can further boost plasticity formation (Lisman, 1997). Gamma-synchronization might modify synaptic plasticity not only within V1, but could be a mechanism to shape the projections patterns from V1 to extrastriate areas according to natural image statistics. In particular, it might spur the unsupervised formation of extrastriate RFs that are sensitive

to specific feature constellations, like extended object boundaries which rely on convergent projections from distant V1 populations. In this way, γ-synchronization could be a mechanism to shape connectivity patterns according to statistical regularities in natural scenes.

Acknowledgements and Authorship Contributions

Conceptualization and design: CU, AP, MV. Data analysis: CU. Design of neural networks: CU and MV. Implementation of neural networks: CU. Recordings: AP, CU, RR, JKL. Software: CU, AP, AL, KS. Surgery: JKL, WB, WS. Materials and reagents: WS, PF, MV. Supervision: MV. Writing: CU, AP, MV. We thank Michael Schmid and Richard Saunders for implantation of monkey H. This projected was supported by ERC Starting Grant to MV (SPATEMP) and a BMF Grant to MV (Bundesministerium fuer Bildung und Forschung, Computational Life Sciences, project BINDA, 031L0167).

Methods

All procedures complied with the German and European regulations for the protection of animals and were approved by the regional authority (Regierungspräsidium Darmstadt).

Surgical procedures

Three adult macaque monkeys (Macaca mulatta, two male, one female) were used in this study. Implantations were made in the left hemisphere of V1. In Monkey H, we implanted a Utah array with 64 microelectrodes (inter-electrode distance 400 μm, half of them with a length of 1 mm and half with a length of 0.6 mm, Blackrock Microsystems), and inserted a reference wire under the dura towards parietal cortex. In monkeys A and I, we implanted a semi-chronic microelectrode array Microdrive (SC32-1, Gray Matter Research), containing 32 independently movable tungsten electrodes (inter-electrode distance of 1.5 mm); here, the microdrive chamber was used as the reference. Note that no histological verification of layers/depths was performed, because the animals are still alive. We estimate that our recordings mainly sample activity from layers 2-4, because the vast majority of recording sites do not show the typical inversion of the first deflection of the event-related potential as is found in the deep layers (Li et al., 2015; Schroeder et al., 1998), this is further supported by the electrode lengths in monkey H. Further details on surgical procedures can be found in (Peter et al., 2019).

Behavioral task

All monkeys were trained on a fixation task. The animals were positioned 83 (monkey H) or 64 cm (monkeys A, I) in front of a 22 inch 120 Hz LCD monitor (Samsung 2233RZ, Ghodrati et al. (2015); Wang (2011)). Monkeys self-initiated trials by fixating on a small fixation spot, which was presented at the center of the screen, and had to maintain fixation during the entire trial. Trials during which the eye position deviated from the fixation spot by more than 0.8 (monkey H) or 1.44 visual deg (monkey I and A) radius were aborted. Correct trials were rewarded with diluted fruit juice. For further details on the task see Peter et al. (2019).

Recordings

Recordings were made with a Tucker Davis Technologies (TDT) system. Data were filtered between 0.35 and 7500 Hz (3 dB filter cutoffs) and digitized at 24.4 kHz (TDT PZ2 preamplifier). Stimulus onsets were recorded with a custom-made photodiode. Eye movements and pupil size were recorded at 1000 Hz using an Eyelink 1000 system with infrared illumination. For further details see Peter et al. (2019).

Visual stimulation

Image selection. Natural images were acquired from the Yahoo Flickr Creative Commons 100 Million (YFCC100M) Dataset (Thomee et al., 2015). The resolution of these images was high enough to match the resolution of the LCD monitor given the stimulus size (see below). Images were included if user tags included any of the following: Animal, building, closeup, flower, house, indoor, landscape, natural, object, texture, tool, toy, tree. Any of the following tags led to exclusion: Blur, blurry, bokeh, digital, art, artwork, artist, text, writing, drawing, painting, cartoon, graphic, graphic+design, illustration, logo, desktop. A set of 340 images was selected as stimulus set for the monkeys. Note that from the same Yahoo Flickr database, a training and validation set was selected for the deep neural network training (see further below); this training set was different from the images presented to the monkey. All images were converted to grayscale, except for a subset of sessions in which we presented the images also in color. The 340 selected images had to fulfill the following criteria, which are comparable to the ones used in Coen-Cagli et al. (2015): (1) A mean RGB value between 40 and 200. (2) An average luminance-contrast, measured as root-mean-squared (RMS) contrast, above 0.2. (3) A centroid of spatial frequency (defined further below in the Image Statistics section) greater than 0.5 dva (degrees of visual angle). These criteria excluded images that were excessively bright, dark, or spatially uniform. However, note that we studied V1 responses to uniform grayscale surfaces in (Peter et al., 2019)).

Image standardization. We cropped each image to 600x600 pixels, and applied two further transformations to the images, similar to (Coen-Cagli et al., 2015): (1) We set the global luminance-contrast for each image to 0.6, by using a sigmoid projection of pixel values. (2) We rescaled and centered the images to have a mean RGB value of 128, equal to the background. For stimulus presentation, we approximately centered the images on the cluster of recorded V1 RFs. Stimuli had a width of 11.5 dva (monkey H, stimulus centered horizontally at x=+2 and vertically at y=-3 dva from fixation) or 15 dva (monkey A x=+3.55 and y=-0.12, monkey I x=+3.45 and y=-0.02).

Further image selection criteria. Stimulus sets for a given recording session consisted of 20 images, each presented for

a total of 10-20 trials. Each session contained a subset of ten images where at least part of the image had a high spatial predictability (>0.85, see below). This was done to ensure that a sufficient number of image patches with high predictability were sampled, given the low probability of finding image patches with very high spatial predictability (Figure Extended Data 3). These images were found as follows: (1) Around the clusters of RFs, a region of 3 x 3 degrees visual angle was selected (ROI). (2) The ROI was divided into 1 degree patches and the spatial predictability of 9 non-overlapping patch locations were quantified (see section Image Statistics). (3) At least 2 out of 9 patch locations were required to have a predictability value (defined with SSIM) above 0.85. The other 10 images were, in terms of predictability, selected randomly. Note that correlations between spatial predictability and synchronization were also found below 0.85 (Figure 2), and that for all 20 images, there were typically some recordings sites with RFs corresponding to low predictability. Additionally, all 20 images were required to have, inside the ROI, an average luminance-contrast above 0.15 and a centroid of the spatial frequency below 8 cycles per degree; this prevents aliasing, given the visual acuity of macaques at the recorded eccentricities.

Data analysis

Preprocessing. We analyzed only correctly performed trials. We downsampled the LFPs to ≈1.02kHz using Matlab's decimate.m function. Line noise was removed using a two-pass 4th order Butterworth bandstop filters between 49.9-50.1, 99.7-100.3 and 149.5-150.5 Hz. Similar to previous studies (Schmid et al., 2013; Self et al., 2013; Xing et al., 2012; Legatt et al., 1980; Peter et al., 2019), we computed MU (Multi Unit) signals from the broadband signal, by bandpass filtering between 300 and 6000 Hz (4th order butterworth), rectification, and applying low-pass filtering and downsampling in the same way as for the LFPs. For the calculation of rate modulations, this MU signal was smoothed with a Gaussian kernel with an SD of 20 ms. For further details see Peter et al. (2019).

RF estimation. RF were mapped with moving bar stimuli (spanning the entire monitor). Moving bars (width 0.1 deg, speed 10-17 deg/s) were presented with 8-16 different orientations. MU responses were computed as a function of stimulus position, after correcting for response latency. MU responses for each movement direction were then fitted by a Gaussian function. We used this fit to extract the 10th percentile and the 90th percentile. Across the 16 directions, this yielded 32 data points, which were fit with an ellipse. This ellipse was defined as that MU's RF. The RF size was defined as the diameter based on 2*sqrt(a*b), where a and b are the major and minor radius. The preferred stimulus orientation was also computed using these bar stimuli (and was highly consistent with orientation tuning based on static gratings).

Electrode selection. We included all electrodes for analysis that met the following criteria: (1) The MU showed a response to RF stimulation that was at least two SD above stimulation outside the RF. (2) The MU response during the response period (0.05-0.15 s post stimulus onset) of at least one stimulus condition in the session was at least 2 SD above the corresponding baseline (-0.4-0 s before stimulus onset).

Estimation of LFP power spectra. For monkey H, we made use of the constant spacing of neighboring electrodes in the array to improve power spectral estimation: Power spectra were approximated by the complex mean of the cross-spectra of a channel with its two same-depth nearest neighbors. This reduced uncorrelated high-frequency fluctuations due to spiking, which can affect the $1/f$ slope of LFP spectra; qualitatively similar results were obtained using the unipolar power spectra. LFP epochs were multi-tapered (±5 Hz smoothing) (Pesaran et al., 2018), Fourier transformed and squared to estimate LFP power spectral densities.

Quantification of LFP gamma-band and beta-band amplitude. In Peter et al. (2019), we developed an algorithm to extract the amplitude of narrow-band gamma-band (or beta-band) oscillations, with several advantages compared to previous methods. Note that qualitatively similar correlations with spatial predictability were obtained by simply using baseline-corrected power in β and γ bands (see Figure 2 and 3). The algorithm had the following structure:

1. Power spectra were log-transformed and the frequency axis was also sampled in log-spaced units to avoid overweighing the contribution of high-frequency datapoints. All subsequent polynomial fits were performed on the 5-200 Hz range.
2. $1/F^n$ correction was performed by fitting an exponential to the LFP power spectrum, excluding data points in the range of 10-85 Hz in order to avoid any influence on peak detection. (For a subset of figures, namely 2 and 3, we corrected the power by dividing by the pre-stimulus baseline).
3. To determine the polynomial order, we used a cross validation procedure to prevent overfitting. Polynomials of order 1-20 were fit to ΔP as a function of frequency for a "training set". We then evaluated the mean squared error using the same polynomial fit on a "test set" for each of the 20 orders. This procedure was then repeated for multiple iterations and we chose the order with the best median performance.
4. On the polynomial fit, local maxima and minima in the β (18-30 Hz) or γ (30-80 Hz) range were identified. The peak frequency was the location of the maximum. The amplitude was then assessed from the difference between the value of the polynomial fit at the maximum and the average of the polynomial fit at F_{min} and $2F_{max} - F_{min}$, where F_{min} is the frequency of the nearest local minimum to the left of the maximum (we used the left one, to avoid any influence of spike-bleed-in at higher frequencies).
5. The amplitude was quantified as a fold-change.

Rate modulation. Rate modulation was computed as

$$R = (M_{stim}/M_{base}) \ , \tag{1}$$

where M_{stim} and M_{base} represent the MU firing activity in the stimulus and baseline period, respectively. Spike-density functions were normalized in the same way.

Statistics. Error bars or shaded error regions correspond to ± one standard errors of the mean (SEM) across recording sites. Violin plots show the median together with the 25-75 percentiles and the data distribution estimated using Matlab's ksdensity.m function.

The statistics used in the rest of the manuscript were as follows:

- In Figure 2C, 3B, and Extended Data 5, we show average neural activity as a function of spatial predictability or luminance-contrast. We pooled all channels and images together and then formed non-overlapping bins of 250 images. We then performed a Spearman-rank correlation across these 250 images.
- In Figure 2D and 3D, we correlated γ-peak amplitude with predictability or luminance-contrast across sessions for each channel separately. We tested whether the average correlation was significantly different from zero across channels by using a two-sided T-test.
- In Figure 3D and 4C, we tested whether different models were significantly different from each other by using 10-fold cross-validation, and a paired T-test. The additional regression variables used in 3D were: Center luminance, center spatial frequency and absolute deviation from preferred MU orientation, as well as the luminance and spatial frequency of the 224x224 pixel image patch around the RF center.
- In Figure 2F and G, we performed a two-sided T-test between nature and man-made or object-boundary vs. no boundary.
- In Figure 4B, we used a Pearson correlation to quantify the trend across layers.
- In Figure 5B and 6B, we compared the strength in correlations between blocks of VGG-16 layers using a pairwise T-test (early: layer 1-4; middle: layer 5-9; deep: 10-13).

Identification of object boundaries and man-made vs. nature pictures

To automatically identify object boundaries, we used the Mask-R convolutional neural network. Using this network, we determined if there is a detected object contour that intersects with the MU's RFs. We distinguished man-made vs. nature picture as follows: Pictures were categorized as man-made if man-made structure (e.g. buildings) was within the 4×4 degree image region centered on the RFs of the recording array (for examples see Figure 2). Images with both nature and man-made content were not considered for this comparison.

Deep Neural Networks methods

We used DNNs for several purposes. In Figures 1-4 we trained a network to predict missing parts of an image to obtain several measures of spatial predictability. We will first describe the stimuli used to train this network, its architecture and training procedures. Following this, we describe other networks in which we directly predicted neural activity from the image.

Preprocessing of training stimuli

We resized the images from the Yahoo Flickr dataset (see Section "Visual Stimulation") to 300×300 pixels. To improve robustness and generalization, we applied standard data augmentation of the images for the training, consisting of several operations from the Tensorflow image module (Abadi et al., 2015). Each operation had a 50% chance of being applied for a given image: brightness (max_delta=0.1), contrast ([low, high] = [0, 1]), hue (max_delta = 0.1), saturation ([low, high] = [0, 1]), convert to black & white, horizontal flip. Resulting images were then randomly cropped to 224×224 pixels. The mean RGB value [123.68, 116.779, 103.939] of ImageNet was subtracted from each image, in order to use the VGG-16 network for the initial layers of the network (Simonyan and Zisserman, 2014).

Mask generation. In order to make a network that can robustly predict missing inputs, we trained the network with binary masks that either had occluders or missing pixels (low signal-to-noise ratio). Binary masks were randomly selected from 3 types: elliptic, rectangular, or salt and pepper noise. Rectangular and elliptic mask types consisted of 2-3 missing regions that were randomly selected to be 20×20 pixels to 80×80 pixels (0.5-2 degrees) in area. One side of the rectangle (or axis for the ellipse) was randomly picked between 20 to 80 pixels. Finally random rotations were applied to each missing region. Salt & Pepper noise had a sparsity ratio of at least 20%. Images including the mask were randomly cropped to size 224×224, and horizontally flipped with a 50% chance.

Architecture of deep neural network for inpainting

The neural network architecture was based on the U-Net architecture (Falk et al., 2019; Ronneberger et al., 2015), with the following modifications: For initialization, the encoder part of U-Net was replaced by all the convolutional and pooling layers of the VGG-16 network, using the Keras implementation (Simonyan and Zisserman, 2014; Chollet et al., 2015). Transfer learning using VGG-16 has been previously used in image segmentation (Iglovikov and Shvets, 2018), image reconstruction (Uhrig et al., 2017), style transfer (Gatys et al., 2015), and image inpainting (Liu et al., 2018). The resulting network architecture consisted of five blocks, each of which had two or three convolutional layers (3×3) with ReLU (rectified linear) activation functions, followed by a max-pooling (2×2) operation. The decoder consisted of five blocks, each with a nearest-neighbor upsampling layer (2×2), followed by two convolutional layers. The output layer was a convolutional layer with, as is conventional, a linear activation function.

Partial Convolution. All convolution operations in the network, including the VGG-16 network, were implemented as partial convolutions. Partial convolution has been introduced with the sparsity-invariant convolutional network where the input to each convolution is paired with a binary mask indicating which pixels are observable or missing, respectively (Uhrig et al., 2017). Partial observability of the inputs during the training makes the network robust to input sparsity, regardless of the task of the network. We implemented a modified version with mask updates per network operation, as described in (Liu et al.,

2018). The idea of partial convolution is that the missing region is gradually filled, and that the filled-in information is used for filling in the rest of the missing pixels in an iterative way.

Loss function for training

Convolutional neural network (CNN) activations have been previously used as a basis for perceptual similarity metrics instead of traditional measures such as SSIM (Wang et al., 2004) or L2-norm (Dosovitskiy and Brox, 2016; Gatys et al., 2015; Zhang et al., 2018). Even though these perceptual loss functions based on CNNs match human perception better, image generation based on those loss functions can suffer from high frequency artifacts (Olah et al., 2017; Johnson et al., 2016). To minimize high-frequency artifacts, we therefore implemented the reconstruction and content loss functions in the Fourier domain, similar to decorrelated image parameterization described in (Mordvintsev et al., 2018; Odena et al., 2016).

Total loss consisted of reconstruction, content, and style losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reconstruction}} + p_{\text{content}} \cdot \mathcal{L}_{\text{content}} + p_{\text{style}} \cdot \mathcal{L}_{\text{style}}$$

Here, p_{style} and p_{content} are hyperparameters (see further below). Reconstruction loss consisted of three terms: (i) The difference between the amplitudes of Fourier transforms; (ii) the log difference between the Fourier amplitudes; (iii) the phase similarity between the predicted and the original image, which has been previously applied to auditory signal synthesis based on the Fourier space (Arık et al., 2018):

$$\begin{aligned}
\mathcal{L}_{\text{reconstruction}} &= \mathrm{F}(\boldsymbol{y}, \hat{\boldsymbol{y}})\,; \\
\mathrm{F}(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= \ell_1\Big(\|\mathfrak{F}(\boldsymbol{y})\| - \|\mathfrak{F}(\hat{\boldsymbol{y}})\|\Big) \\
&\quad + p_{\log} \cdot \ell_1\Big(\ln\|\mathfrak{F}(\boldsymbol{y})\| - \ln\|\mathfrak{F}(\hat{\boldsymbol{y}})\|\Big) \\
&\quad + p_{\text{phase}} \cdot \ell_1\Big(\|\mathfrak{F}(\boldsymbol{y})\| \cdot \|\mathfrak{F}(\hat{\boldsymbol{y}})\| \\
&\qquad - \|\mathrm{Re}(\mathfrak{F}(\boldsymbol{y}))\| \cdot \|\mathrm{Re}(\mathfrak{F}(\hat{\boldsymbol{y}}))\| \\
&\qquad - \|\mathrm{Im}(\mathfrak{F}(\boldsymbol{y}))\| \cdot \|\mathrm{Im}(\mathfrak{F}(\hat{\boldsymbol{y}}))\|\Big)
\end{aligned}$$

Here, $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ are the original and predicted image. The operator $\mathfrak{F}$ denotes the Discrete Fourier Transform, $\|\mathfrak{F}(\boldsymbol{y})\|$ denotes the Fourier magnitude, and ℓ_1 denotes the L1-norm across all frequencies.

Content loss was also defined in the Fourier domain, taking as input the cth feature in the λth layer activations of the VGG-16 network:

$$\mathcal{L}_{\text{content}} = \sum_{c} \sum_{\lambda} \mathrm{F}(\phi_{\lambda,c}(\boldsymbol{y}), \phi_{\lambda,c}(\hat{\boldsymbol{y}}))$$

Finally, style loss is the difference between the Gramian matrices of the Fourier amplitude of the predicted and the original images, where the Gramian matrices contain the covariance matrix across features:

$$\mathcal{L}_{\text{style}} = \sum_{\lambda} \ell_1\Big(\mathrm{G}_{cc'}(\|\mathfrak{F}(\phi_{\lambda,c}(\mathrm{y}))\|) - \mathrm{G}_{cc'}(\|\mathfrak{F}(\phi_{\lambda,c}(\hat{\mathrm{y}}))\|)\Big)$$

Four layers of VGG-16 network with fixed weights were used for content and style loss.

Training and Hyperparameter Optimization

We initialized the weights of the encoder as VGG-16 model weights that were pre-trained on ImageNet. The remaining weights were initialized using He-initialization (He et al., 2015) and bias terms as 0. The network weights were optimized using the Adam optimizer with a learning rate of $5e^{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e^{-7}$, (Kingma and Ba, 2014). All hyperparameters were defined as continuous variables where the search space was [1e-5, 1e3] sampled in logarithmic scale (Lindauer et al., 2019). We optimized the hyperparameters using a method that combines Bayesian optimization with Hyperband as described in (Falkner et al., 2018). We ran the optimization for 50 iterations with a minimum budget of 8 and a max budget of 16. For hyperparameter optimization, we used a subset of the image dataset for training (16384 images) and validation (1024 images) of the network. We used SSIM based reconstruction loss as the evaluation metric for the hyperparameter optimization. To analyze the hyperparameter space and the importance of individual hyperparameters, we usd fANOVA (Hutter et al., 2014). The importance of a hyperparameter is the fraction of explained variance (mean across 100 repetitions ± SEM) of the validation SSIM-loss across the entire hyperparameter space. The resulting hyperparameters and their importances were:

Name	Value	Importance
p_{style}	0.18	0.13 ± 0.002
$p_{content}$	36.90	0.22 ± 0.002
p_{log}	5.54	0.14 ± 0.002
p_{phase}	3.02e-05	0.29 ± 0.003
p_{lr}	4.48e-05	0.14 ± 0.002

Image Statistics

Spatial Predictability, for a given image, predictability was computed by masking out the central 1 degree patch (similar in size to previous studies studying contextual modulation, Coen-Cagli et al. (2015)). The RF image-patch was then predicted by the inpainting algorithm described above. The to-be-predicted image patch is denoted $\mathbf{I}$, a matrix of $N \times N$ pixels.

Structural predictability. Structural predictability was defined as the Pearson correlation of two images or VGG-16 activations.

$$\rho = \frac{\mathrm{cov}(X, Y)}{\sigma_x \sigma_y}$$

Content predictability. Content loss was defined as as the L2-norm of the difference between the VGG-16 λ layer activations of two images. This was defined for each layer of the VGG-16 separately:

$$L_\lambda = \frac{1}{I_\lambda J_\lambda C_\lambda} \sum_{i,j,c} \ell_2\Big(\phi_{\lambda,c}(\boldsymbol{y}_{i,j}) - \phi_{\lambda,c}(\hat{\boldsymbol{y}}_{i,j})\Big)$$

Luminance-contrast Luminance-contrast was measured as the Root-mean-square (RMS) contrast:

$$\text{RMS} = \sqrt{\frac{1}{N^2}\sum_{ij}\left(I_{ij} - \bar{I}\right)^2},$$

where $\bar{I}$ is the mean value of I. RMS was defined to range between 0 (minimum) and 1 (maximum) luminance-contrast.

Spectral Centroid. The center of mass of the power spectrum is the mean spatial frequency in the image patch, f_k, weighted by the total power, r_k in the frequency bands.

$$R = \frac{\sum_k r_k f_k}{\sum_k r_k}.$$

Orientation. Mean orientation in the RF was computed weighted by the total power in the orientation band [0:5:180] degrees, across spatial frequencies 1.0, 2.8284, 8.0 cycles per degree, similar to the orientation selectivity of the receptive field. We estimated the orientation selectivity of the RF as the weighted circular mean across different orientations as described in (Ringach et al., 2002). The population resultant vector, R, is the orientation selectivity magnitude, $|R|$, and the mean angle $\arg(R)$, where

$$R = \frac{\sum_k r_k e^{i2\theta_k}}{\sum_k r_k}.$$

Here r_k is the peak MU response for the orientation θ_k.

Predicting neural activity from VGG-16 activations

In Figure 5, we predicted neural activity from the VGG-16 activations. If VGG-16 units had a smaller RF than 1 dva, then only VGG-16 units were used with a RF center within the 1 degree region around the multi-unit's RF center. If VGG-16 units had a larger RF than 1 dva, then only VGG-16 units were used that fully covered the central 1 degree. For each location in the VGG-16, we predicted the neural responses from a vector of VGG-16 units with different feature selectivities. For this, we used 10% of the stimuli as a test set and used the rest as the training set. We used L1-constrained linear regression and 10-fold cross validation to select the L1-constraint parameter λ. Regression coefficients that best explain the training set were then used to predict the neural signals for the test set. The correlation values (r) were averaged across VGG-16 locations.

Receptive field in the VGG-16 are shown in Table 1. We calculated the receptive fields for VGG-16 units as described in (Araujo et al., 2019) using their toolbox `https://github.com/google-research/receptive_field/`. The receptive field sizes of the neural network in the middle layers were comparable to the receptive field sizes in V1, which has been previously described in Cadena et al. (2019).

Prediction from VGG-16 activations: gamma-net

In Figure 6, we used a convolutional neural network to predict γ-peaks based on the input image. In total, training data consisted of 31988 image patches, by pooling across monkeys, channels and sessions. We used all image patches from a unique %10 of the stimuli as a test set and used the rest for training. We resized the 224x224 image patches to 84x84 in order to reduce the number parameters in the network. We used the VGG-16 model from keras applications, with frozen weights pre-trained on ImageNet for transfer learning. The VGG-16 activations were the input to a network that consisted of 2 convolutional layers and a readout layer. We compared predictions from different VGG-16 input layers, and for the main examples of RFs used *conv3_3* as input (Figure 6). Convolutional layers consisted of (3x3) filters with bias, stride (2), valid padding, L1-norm kernel regularization (0.001), leaky ReLU activation (0.1), and dropout (0.5). The final two convolutional layers had 32 and 16 units. The readout layer consisted of (4x4) filters with bias and leaky ReLU activation (0.1).

We visualized the feature activations of gamma-net by maximizing the mean activation of a unit or the network output. Feature activations were maximized with the Adam optimizer (2048 iterations and learning rate of 0.05). In order to have fewer visualization artifacts on the activations, we implemented several techniques as described in Olah et al. (2017, 2018): (1) Single color channel decorrelated Fourier domain parametrization of the activation image. (2) Standard transformations (padding, jitter, random scaling and rotation). All visualizations were made using the toolbox Lucid `https://github.com/tensorflow/lucid`. To further study the selectivity of this network, we presented grating stimuli of different spatial frequencies, orientations, luminance-contrasts and sizes to this network (Figure 6D).

Tools

Neural network visualizations were made by LaTeXpackage PlotNeuralNet `https://github.com/HarisIqbal88/PlotNeuralNet`. Software and packages used for the analysis include MATLAB, matplotlib (Hunter, 2007), numpy (Van Der Walt et al., 2011), docker (Merkel, 2014), tensorflow v1.14 (Abadi et al., 2015), and (Chollet et al., 2015). The Titan X Pascal used for this research was donated by the NVIDIA Corporation.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. software available from tensorflow.org.

Adesnik, H., Bruns, W., Taniguchi, H., Huang, Z.J., Scanziani, M., 2012. A neural circuit for spatial summation in visual cortex. Nature

Adjamian, P., Holliday, I.E., Barnes, G.R., Hillebrand, A., Hadjipapas, A., Singh, K.D., 2004a. Induced visual illusions and gamma oscillations in human primary visual cortex. European Journal of Neuroscience

Adjamian, P., Holliday, I.E., Barnes, G.R., Hillebrand, A., Hadjipapas, A., Singh, K.D., 2004b. Induced visual illusions and gamma oscillations in human primary visual cortex. European Journal of Neuroscience

Akam, T.E., Kullmann, D.M., 2012. Efficient "Communication through Coherence" requires oscillations structured to minimize interference between signals. PLOS Comp. Biol.

Angelucci, A., Bijanzadeh, M., Nurminen, L., Federer, F., Merlin, S., Bressloff, P.C., 2017. Circuits and mechanisms for surround modulation in visual cortex. Ann. Rev. Neurosci.

Araujo, A., Norris, W., Sim, J., 2019. Computing receptive fields of convolutional neural networks. Https://distill.pub/2019/computing-receptive-fields.

Arık, S.Ö., Jun, H., Diamos, G., 2018. Fast spectrogram inversion using multi-head convolutional neural networks. IEEE Signal Processing Letters

Arnal, L.H., Wyart, V., Giraud, A.L., 2011. Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. Nat. Neurosci.

Bartoli, E., Bosking, W., Chen, Y., Li, Y., Sheth, S.A., Beauchamp, M.S., Yoshor, D., Foster, B.L., 2019. Functionally distinct gamma range activity revealed by stimulus tuning in human visual cortex. Current Biology

Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., Friston, K.J., 2012. Canonical microcircuits for predictive coding. Neuron

Bastos, A.M., Vezoli, J., Bosman, C.A., Schoffelen, J.M., Oostenveld, R., Dowdall, J.R., De Weerd, P., Kennedy, H., Fries, P., 2015. Visual areas exert feedforward and feedback influences through distinct frequency channels. Neuron

Bellec, G., Scherr, F., Subramoney, A., Hajek, E., Salaj, D., Legenstein, R., Maass, W., 2020. A solution to the learning dilemma for recurrent networks of spiking neurons.

Bengio, Y., Courville, A.C., Vincent, P., 2012. Unsupervised feature learning and deep learning: A review and new perspectives. CoRR arXiv:1206.5538.

Biederlack, J., Castelo-Branco, M., Neuenschwander, S., Wheeler, D.W., Singer, W., Nikolić, D., 2006. Brightness Induction: Rate Enhancement and Neuronal Synchronization as Complementary Codes. Neuron

Bressler, S.L., Coppola, R., Nakamura, R., 1993. Episodic multiregional cortical coherence at multiple frequencies during visual task performance. Nature

Bressler, S.L., Richter, C.G., Chen, Y., Ding, M., 2006. Top-down cortical influences in visual expectation, in: IJCNN'06.,

Bressler, S.L., Richter, C.G., Chen, Y., Ding, M., 2007. Cortical functional network organization from autoregressive modeling of local field potential oscillations. Statistics in medicine

Brunet, N., Bosman, C.A., Roberts, M., Oostenveld, R., Womelsdorf, T., De Weerd, P., Fries, P., 2015. Visual Cortical Gamma-Band Activity During Free Viewing of Natural Images. Cereb. Cortex.

Brunet, N., Vinck, M., Bosman, C.A., Singer, W., Fries, P., 2014. Gamma or no gamma, that is the question. Trends Cogn. Sci.

Brunet, N.M., Fries, P., 2019. Human visual cortical gamma reflects natural image structure. NeuroImage

Burns, S.P., Xing, D., Shapley, R.M., 2011. Is gamma-band activity in the local field potential of V1 cortex a "clock" or filtered noise? J. Neurosci.

Buschman, T.J., Miller, E.K., 2007. Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices. Science

Buzsáki, G., 2006. Rhythms of the Brain.

Buzsáki, G., Draguhn, A., 2004. Neuronal oscillations in cortical networks. Science

Buzsáki, G., 1989. Two-stage model of memory trace formation: A role for "noisy" brain states. Neuroscience

Cadena, S.A., Denfield, G.H., Walker, E.Y., Gatys, L.A., Tolias, A.S., Bethge, M., Ecker, A.S., 2019. Deep convolutional models improve predictions of macaque V1 responses to natural images. PLoS computational biology

Cavanaugh, J.R., Bair, W., Movshon, J.A., 2002. Nature and interaction of signals from the receptive field center and surround in macaque v1 neurons. Journal of neurophysiology

Chollet, F., et al., 2015. Keras.

Coen-Cagli, R., Dayan, P., Schwartz, O., 2012. Cortical surround interactions and perceptual salience via natural scene statistics. PLoS Comput Biol

Coen-Cagli, R., Kohn, A., Schwartz, O., 2015. Flexible gating of contextual influences in natural vision. Nat Neurosci

DiCarlo, J.J., Zoccolan, D., Rust, N.C., 2012. How does the brain solve visual object recognition? Neuron

Dosovitskiy, A., Brox, T., 2016. Generating images with perceptual similarity metrics based on deep networks, in: Advances in neural information processing systems,

Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marrakchi, Y., Böhm, A., Deubner, J., Jäckel, Z., Seiwald, K., et al., 2019. U-net: deep learning for cell counting, detection, and morphometry. Nature methods

Falkner, S., Klein, A., Hutter, F., 2018. BOHB: Robust and efficient hyperparameter optimization at scale, in: Dy, J., Krause, A. (Eds.), Proceedings of the 35th International Conference on Machine Learning, PMLR, Stockholmsmässan, Stockholm Sweden. pp. 1437–1446.

Gail, A., 2000. Contour Decouples Gamma Activity Across Texture Representation in Monkey Striate Cortex. Cerebral Cortex

Galuske, R.A., Munk, M.H., Singer, W., 2019. Relation between gamma oscillations and neuronal plasticity in the visual cortex. Proceedings of the National Academy of Sciences

Gatys, L.A., Ecker, A.S., Bethge, M., 2015. A neural algorithm of artistic style.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W., 2018. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.

Ghodrati, M., Morris, A.P., Price, N.S.C., 2015. The (un)suitability of modern liquid crystal displays (LCDs) for vision research. Front. Psychol.

Gieselmann, M.A., Thiele, A., 2008. Comparison of spatial integration and surround suppression characteristics in spiking activity and the local field potential in macaque V1. Eur. J. Neurosci.

Gilbert, C.D., 1992. Horizontal integration and cortical dynamics. Neuron

Gray, C., McCormick, D., 1996. Chattering cells: superficial pyramidal neurons contributing to the generation of synchronous oscillations in the visual cortex. Science

Gray, C.M., König, P., Engel, A.K., Singer, W., 1989. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. Nature

Hadjipapas, A., Lowet, E., Roberts, M.J., Peter, A., De Weerd, P., 2015. Parametric variation of gamma frequency and power with luminance contrast: A comparative study of human MEG and monkey LFP and spike responses. NeuroImage

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,

Heeger, D.J., 2017. Theory of cortical function. Proceedings of the National Academy of Sciences

Henrie, J.A., Shapley, R., 2005. LFP power spectra in V1 cortex: The graded effect of stimulus contrast. J. Neurophysiol.

Hermes, D., Miller, K., Wandell, B., Winawer, J., 2015. Stimulus dependence of gamma oscillations in human visual cortex. Cereb. Cortex.

Hubel, D.H., Wiesel, T.N., 1965. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. Journal of neurophysiology

Hunter, J.D., 2007. Matplotlib: A 2d graphics environment. Computing in science & engineering

Hutter, F., Hoos, H., Leyton-Brown, K., 2014. An efficient approach for assessing hyperparameter importance, in: Proceedings of International Conference on Machine Learning 2014 (ICML 2014),

Iglovikov, V., Shvets, A., 2018. Ternausnet: U-net with VGG11 encoder pre-trained on imagenet for image segmentation.

Issa, E.B., Cadieu, C.F., DiCarlo, J.J., 2018. Neural dynamics at successive stages of the ventral visual stream are consistent with hierarchical error signals. Elife

Jadi, M.P., Sejnowski, T.J., 2014. Cortical oscillations arise from contextual interactions that regulate sparse coding. Proc. Natl. Acad. Sci. U.S.A.
Jia, X., Smith, M.A., Kohn, A., 2011. Stimulus selectivity and spatial coherence of gamma components of the local field potential. Journal of Neuroscience
Jia, X., Tanabe, S., Kohn, A., 2013. Gamma and the coordination of spiking activity in early visual cortex. Neuron
Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution, in: European conference on computer vision, Springer.
Kapadia, M., Westheimer, G., Gilbert, C., 1999. Dynamics of spatial summation in primary visual cortex of alert monkeys. PNAS
van Kerkoerle, T., Self, M.W., Dagnino, B., Gariel-Mathis, M.A., Poort, J., van der Togt, C., Roelfsema, P.R., 2014. Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex.
Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization.
Krogh, L., Vlach, H., Johnson, S., 2013. Statistical learning across development: Flexible yet constrained. Frontiers in Psychology 3, 598.
Legatt, A.D., Arezzo, J., Vaughan, H.G., 1980. Averaged multiple unit activity as an estimate of phasic changes in local neuronal activity: effects of volume-conducted potentials. J. Neurosci. Methods
Li, X., Chen, Y., Lashgari, R., Bereshpolova, Y., Swadlow, H.A., Lee, B.B., Alonso, J.M., 2015. Mixing of chromatic and luminance retinal signals in primate area V1. Cerebral Cortex
Lima, B., Singer, W., Chen, N.H., Neuenschwander, S., 2010. Synchronization dynamics in response to plaid stimuli in monkey V1. Cereb. Cortex
Lindauer, M., Eggensperger, K., Feurer, M., Biedenkapp, A., Marben, J., Müller, P., Hutter, F., 2019. Boah: A tool suite for multi-fidelity bayesian optimization and analysis of hyperparameters.
Lindsey, J., Ocko, S.A., Ganguli, S., Deny, S., 2019. A unified theory of early visual representations from retina to cortex through anatomically constrained deep cnns.
Lisman, J.E., 1997. Bursts as a unit of neural information: making unreliable synapses reliable. Trends in neurosciences
Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B., 2018. Image inpainting for irregular holes using partial convolutions, in: Proceedings of the European Conference on Computer Vision (ECCV),
McIntosh, L., Maheswaranathan, N., Nayebi, A., Ganguli, S., Baccus, S., 2016. Deep learning models of the retinal response to natural scenes, in: Advances in neural information processing systems,
Merkel, D., 2014. Docker: lightweight linux containers for consistent development and deployment. Linux journal
Mordvintsev, A., Pezzotti, N., Schubert, L., Olah, C., 2018. Differentiable image parameterizations. Https://distill.pub/2018/differentiable-parameterizations.
Odena, A., Dumoulin, V., Olah, C., 2016. Deconvolution and checkerboard artifacts. Distill
Olah, C., Mordvintsev, A., Schubert, L., 2017. Feature visualization. Https://distill.pub/2017/feature-visualization.
Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., Mordvintsev, A., 2018. The building blocks of interpretability. Https://distill.pub/2018/building-blocks.
Onorato, I., Neuenschwander, S., Hoy, J., Lima, B., Rocha, K.S., Broggini, A.C., Uran, C., Spyropoulos, G., Klon-Lipok, J., Womelsdorf, T., Fries, P., Niell, C., Singer, W., Vinck, M., 2020. A distinct class of bursting neurons with strong gamma synchronization and stimulus selectivity in monkey V1. Neuron
Palanca, B.J., DeAngelis, G.C., 2005. Does neuronal synchrony underlie visual feature grouping? Neuron
Pesaran, B., Vinck, M., Einevoll, G., Sirota, A., Fries, P., Siegel, M., Truccolo, W., Schroeder, C., Srinivasan, R., 2018. Investigating large-scale brain dynamics using field potential recordings: analysis and interpretation.
Peter, A., Uran, C., Klon-Lipok, J., Roese, R., Van Stijn, S., Barnes, W., Dowdall, J.R., Singer, W., Fries, P., Vinck, M., 2019. Surface color and predictability determine contextual modulation of V1 firing and gamma oscillations. eLife
Rao, R.P., Ballard, D.H., 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci.
Ray, S., Hsiao, S.S., Crone, N.E., Franaszczuk, P.J., Niebur, E., 2008. Effect of stimulus intensity on the spike-local field potential relationship in the secondary somatosensory cortex. J. Neurosci.
Ray, S., Maunsell, J.H., 2011. Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. PLOS Biol.
Ray, S., Maunsell, J.H., 2015. Do gamma oscillations play a role in cerebral cortex? Trends in cognitive sciences
Ray, S., Maunsell, J.H.R., 2010. Differences in gamma frequencies across visual cortex restrict their possible use in computation. Neuron
Richter, C.G., Coppola, R., Bressler, S.L., 2018. Top-down beta oscillatory signaling conveys behavioral context in early visual cortex. Sci. Rep.
Ringach, D.L., Shapley, R.M., Hawken, M.J., 2002. Orientation selectivity in macaque v1: diversity and laminar dependence. Journal of Neuroscience
Roelfsema, P.R., Lamme, V.A., Spekreijse, H., 2004. Synchrony and covariation of firing rates in the primary visual cortex during contour grouping. Nat. Neurosci.
Rols, G., Tallon-Baudry, C., Girard, P., Bertrand, O., Bullier, J., 2001. Cortical mapping of gamma oscillations in areas V1 and V4 of the macaque monkey. Visual Neuroscience
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer.
Salinas, E., Sejnowski, T.J., 2001. Correlated neuronal activity and the flow of neural information. Nat Rev Neurosci
Santolin, C., Saffran, J.R., 2018. Constraints on statistical learning across species. Trends in Cognitive Sciences
Sceniak, M.P., Ringach, D.L., Hawken, M.J., Shapley, R., 1999. Contrast's effect on spatial summation by macaque v1 neurons. Nature neuroscience
Schmid, M.C., Schmiedt, J.T., Peters, A.J., Saunders, R.C., Maier, A., Leopold, D.A., 2013. Motion-sensitive responses in visual area V4 in the absence of primary visual cortex. J Neurosci
Schneider, M., Dann, B., Sheshadri, S., Scherberger, H.G., Vinck, M., 2020. A general theory of coherence between brain areas.
Schroeder, C.E., Mehta, A.D., Givre, S.J., 1998. A spatiotemporal profile of visual system activation revealed by current source density analysis in the awake macaque. Cerebral cortex (New York, NY: 1991)
Schultz, W., Stauffer, W.R., Lak, A., 2017. The phasic dopamine signal maturing: from reward via behavioural activation to formal economic utility. Current opinion in neurobiology
Schwartz, O., Simoncelli, E.P., 2001. Natural signal statistics and sensory gain control. Nat Neurosci
Schwiedrzik, C.M., Freiwald, W.A., 2017. High-level prediction signals in a low-level area of the macaque face-processing hierarchy. Neuron
Sejnowski, T.J., Paulsen, O., 2006. Network oscillations: emerging computational principles. J. Neurosci.
Self, M.W., van Kerkoerle, T., Supèr, H., Roelfsema, P.R., 2013. Distinct Roles of the Cortical Layers of Area V1 in Figure-Ground Segregation. Curr. Biol.
Shadlen, M.N., Movshon, J.A., 1999. Synchrony unbound: a critical evaluation of the temporal binding hypothesis. Neuron
Shirhatti, V., Ray, S., 2018. Long-wavelength (reddish) hues induce unusually large gamma oscillations in the primate primary visual cortex.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition.
Singer, W., Gray, C.M., 1995. Visual feature integration and the temporal correlation hypothesis. Annu. Rev. Neurosci.
Sjöström, P.J., Turrigiano, G.G., Nelson, S.B., 2001. Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. Neuron
Spillmann, L., Dresp-Langley, B., Tseng, C., 2015.
Spyropoulos, G., Dowdall, J.R., Scholvinck, M.L., Bosman, C.A., Lima, B., Peter, A., Onorato, I., Klon-Lipok, J., Roese, R., Neuenschwander, S., et al., 2020. Spontaneous variability in gamma dynamics described by a linear harmonic oscillator driven by noise.
Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J., 2015. Yfcc100m: The new data in multimedia research.
Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A., 2017. Sparsity invariant cnns, in: 2017 International Conference on 3D Vision (3DV), IEEE.
Van Der Walt, S., Colbert, S.C., Varoquaux, G., 2011. The numpy array: a structure for efficient numerical computation. Computing in Science & Engineering
Veit, J., Hakim, R., Jadi, M.P., Sejnowski, T.J., Adesnik, H., 2017. Cortical gamma band synchronization through somatostatin interneurons. Nat. Neurosci
Vinck, M., Bosman, C.A., 2016. More gamma more predictions: Gamma-

synchronization as a key mechanism for efficient integration of classical receptive field inputs with surround predictions. Front Syst Neurosci

Vinck, M., Lima, B., Womelsdorf, T., Oostenveld, R., Singer, W., Neuenschwander, S., Fries, P., 2010. Gamma-Phase Shifting in Awake Monkey Visual Cortex. J. Neurosci.

Vinck, M., Womelsdorf, T., Buffalo, E.A., Desimone, R., Fries, P., 2013. Attentional modulation of cell-class-specific gamma-band synchronization in awake monkey area V4. Neuron

Vinje, W.E., Gallant, J.L., 2000. Sparse coding and decorrelation in primary visual cortex during natural vision. Science

Von Der Malsburg, C., 1994. The correlation theory of brain function.

Wachtler, T., Sejnowski, T.J., Albright, T.D., 2003. Representation of color stimuli in awake macaque primary visual cortex. Neuron

Walker, E.Y., Sinz, F.H., Cobos, E., Muhammad, T., Froudarakis, E., Fahey, P.G., Ecker, A.S., Reimer, J., Pitkow, X., Tolias, A.S., 2019. Inception loops discover what excites neurons most using deep predictive models. Nature neuroscience

Wang, P., 2011. An LCD monitor with sufficiently precise timing for research in vision. Front. Hum. Neurosci

Wang, X.J., 2010. Neurophysiological and computational principles of cortical rhythms in cognition. Physiol. Rev.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., et al., 2004. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing

Wespatat, V., Tennigkeit, F., Singer, W., 2004. Phase sensitivity of synaptic modifications in oscillating cells of rat visual cortex. Journal of Neuroscience

Xing, D., Yeh, C.I., Burns, S., Shapley, R.M., 2012. Laminar analysis of visually evoked activity in the primary visual cortex. Proc. Natl. Acad. Sci. U.S.A.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

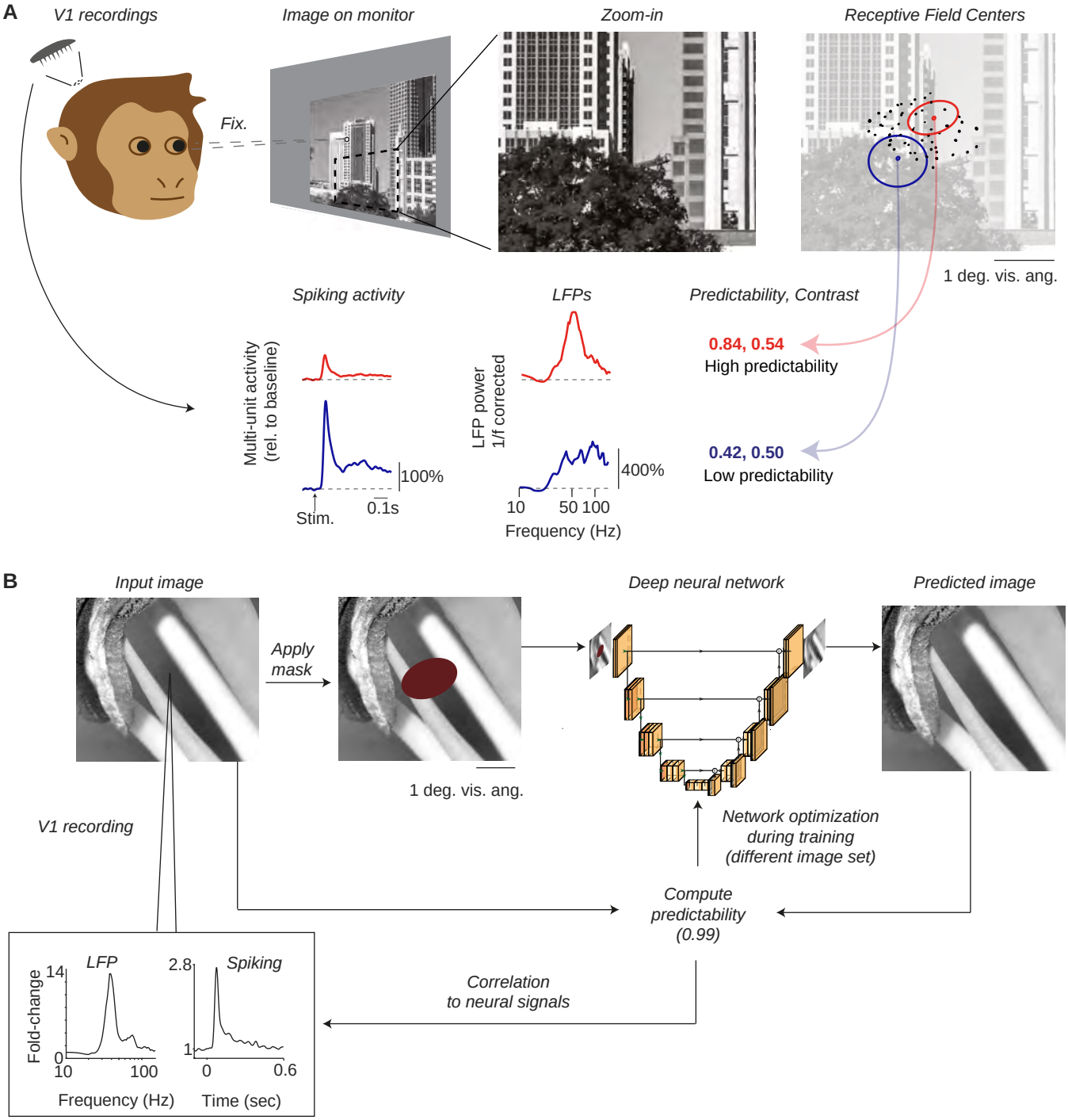

Figure 1: Illustration of experiments and image prediction network. **(A)** We recorded Local Field Potentials (LFPs) and spiking activity from area V1 of awake monkeys. Monkeys performed a passive fixation task. We presented natural images of > 11 degree visual angle in size. Shown are two example sites with their receptive fields (RFs) overlaid on a zoomed-in part of the image. Bottom: left to right: Firing intensity, LFP power spectra, and values of luminance contrast and spatial predictability based on RF locations. The red recording site shows clear narrow-band LFP gamma oscillations around 50 Hz, whereas the blue site shows a broad-band increase in power due to increased spiking activity (Ray and Maunsell, 2011). **(B)** Illustration of a deep neural network that predicts visual inputs into the RFs. A mask with approximately the same size as the recording site's RF is applied to an image. The image with the mask is then entered as an input to a deep neural network with a U-net architecture. This network generates/predicts the full image with the image content behind the mask filled-in. Spatial predictability is computed by comparing the ground-truth input image and the predicted image. The predictability signal is then used for network optimization during the training stage. After network training, a novel set of images is presented to the network and the monkeys. We then analyze how spatial predictability determines neural activity. Shown are firing intensity and LFP power for the recording site with a RF corresponding to the mask. LFP spectrum shows clear gamma oscillations around 40 Hz.

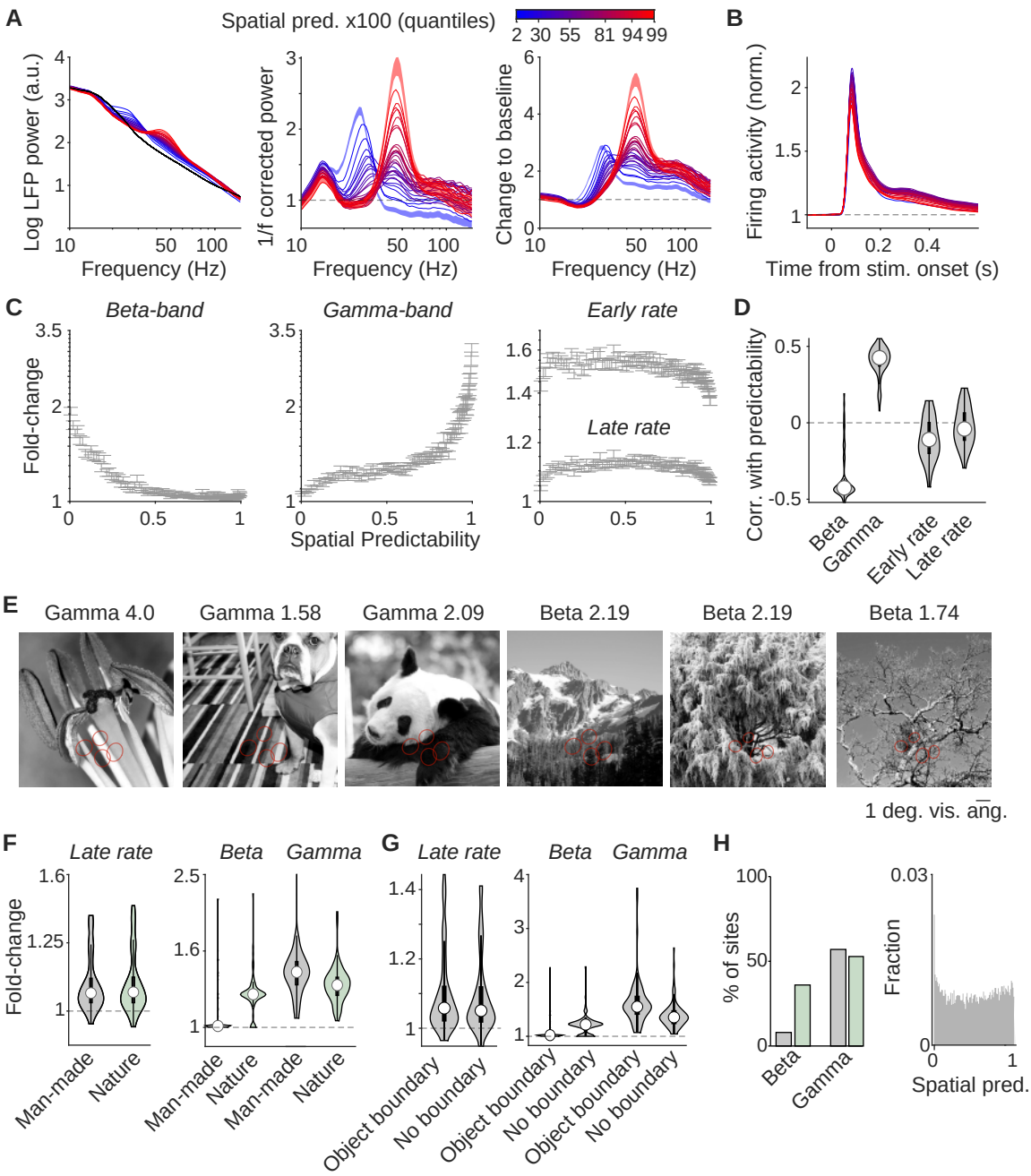

Figure 2: Distinct relationships of firing rates and neural synchronization with spatial predictability. **(A)** Left: Average LFP power spectra for monkey H. Spectra show distinct peaks in the γ and β frequency range. Blue-to-red colors indicate levels of spatial predictability. Black line indicates the pre-stimulus period. Middle: Average 1/f corrected LFP power spectra. SEMs are shown for the lowest and highest quantile of predictability. Right: Average baseline-corrected LFP power spectra. **(B)** Multi-unit firing rate for different levels of spatial predictability. **(C)** Left: Average (±SEM) of 1/f corrected β-peak amplitude. Average was computed across all recording sites in the three animals (N = 72 sites). Middle: Same for γ. Right: Same for early (50-150 ms) and late firing rates (200-600 ms). **(D)** Average Pearson-r correlation across recording sites between spatial predictability and γ, β and rate. Correlation were computed for each channel separately, across all images presented across sessions. Correlations were significant for β, γ and early rates (all $P < 0.001$) but not for late rates ($P = 0.11$) (T-test). Absolute correlations were higher for γ and β than early and late rates ($P < 0.001$ for all comparisons). **(E)** Example images shown with four RFs on the edge of the array, together with the γ or β fold-changes. **(F)** Fold-changes in neural activity for images with man-made content (e.g. buildings) in the RF and images with nature content (e.g. trees, bushes) in the RF. Comparison was significant for β ($P < 0.001$) and γ ($P < 0.001$), but not for early and late rates ($P = 0.5$ and $P = 0.9$, T-Test). **(G)** Fold-changes in neural activity for image patches that contain object boundaries versus image patches that do not. Comparisons were significant for β ($P < 0.001$) and γ ($P < 0.001$), but not for early and late rates ($P = 0.94$ and $P = 0.73$, T-test). **(H)** Left: Percentage of sites with a β or a γ peak, across all randomly selected images. Right: Distribution of spatial predictability across randomly selected images that belong to the categories used in this study (see Methods).

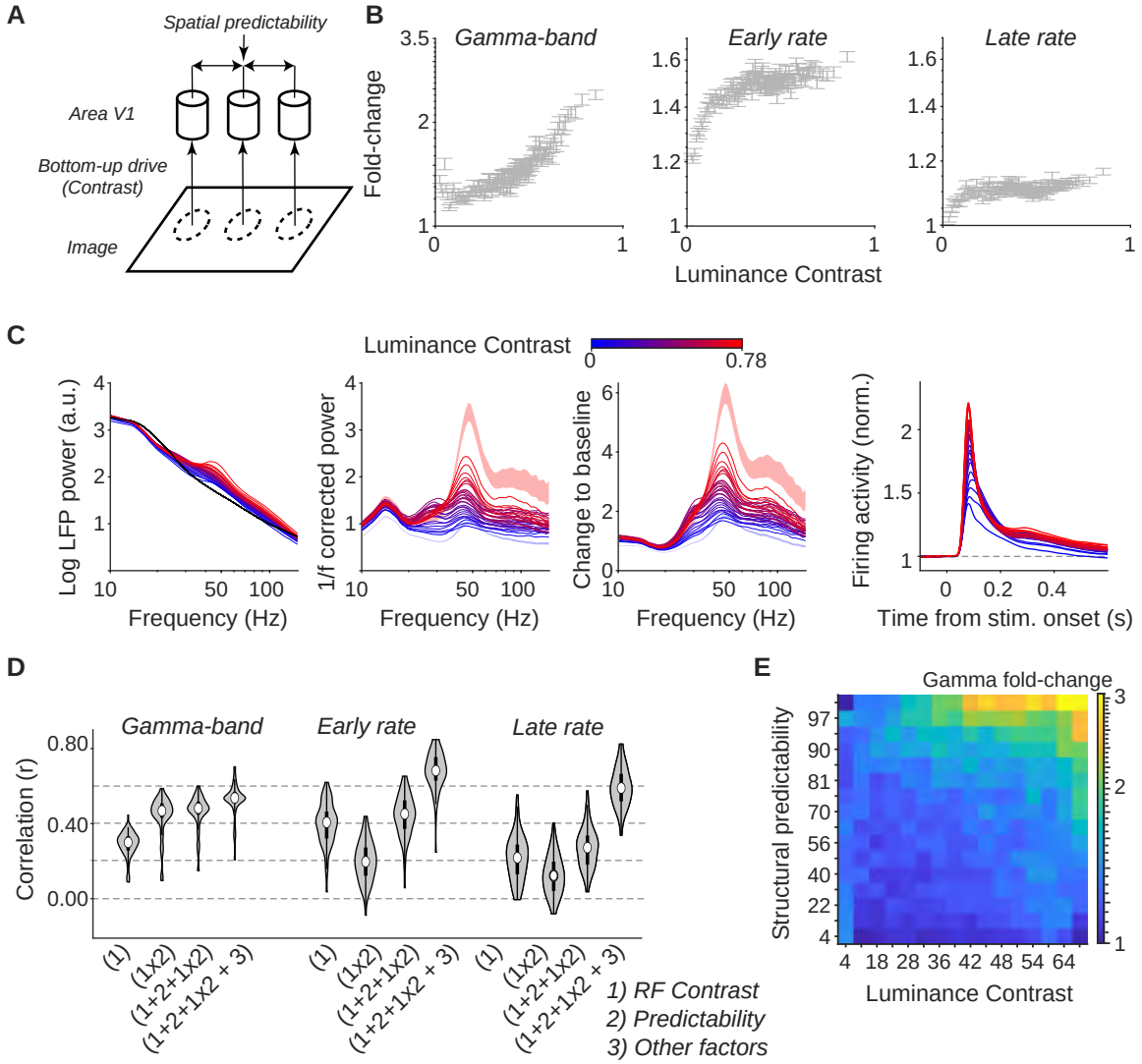

Figure 3: Dependence of firing rates and synchronization on luminance contrast. **(A)** Illustration of interaction between predictability and bottom-up inputs. Firing rates and synchronization can be modified both by bottom-up inputs and image predictability. The influence of spatial predictability is mediated by horizontal, recurrent interactions or top-down inputs. **(B)** Left: γ-peak amplitude for different levels of luminance contrast. Middle: Same for early MU firing rates (50-150ms). Right: Same for late MU firing rates (200-600ms). **(C)** Left: Average LFP power spectra for monkey H. Black line indicates the pre-stimulus period. Second panel: Average 1/f-corrected LFP power spectra, with SEMs shown for the highest and lowest level of luminance contrast. Third panel: Average baseline-corrected LFP power spectra. Right: Multi-unit firing rate for different levels of luminance contrast. **(D)** Average correlation across sites of γ and firing rate with image factors. Left-to-right: Luminance contrast (1); the product of contrast and predictability (1x2); predictability; contrast, predictability $\times$ interaction, (1+2+1x2); and a model with additional low-level features (1+2+1x2+3), namely spatial frequency, luminance and orientation (see Methods). The correlation was computed for each recording site separately across all images presented across sessions. All correlations were significantly different from zero ($P < 0.001$, paired T-test). For γ, the difference between (1x2) and (1+2+1x2) was significant ($P < 0.001$), but the difference between (1+2+1x2) and (1+2+1x2+3) was not ($P > 0.05$). For early rates, all comparisons were significant ($P < 0.05$). For late rates, all comparisons except for (1+2+1x2) vs. (1x2) were significant at $P < 0.05$. **(E)** Heatmap of γ fold-changes for different quantiles of luminance-contrast and predictability.

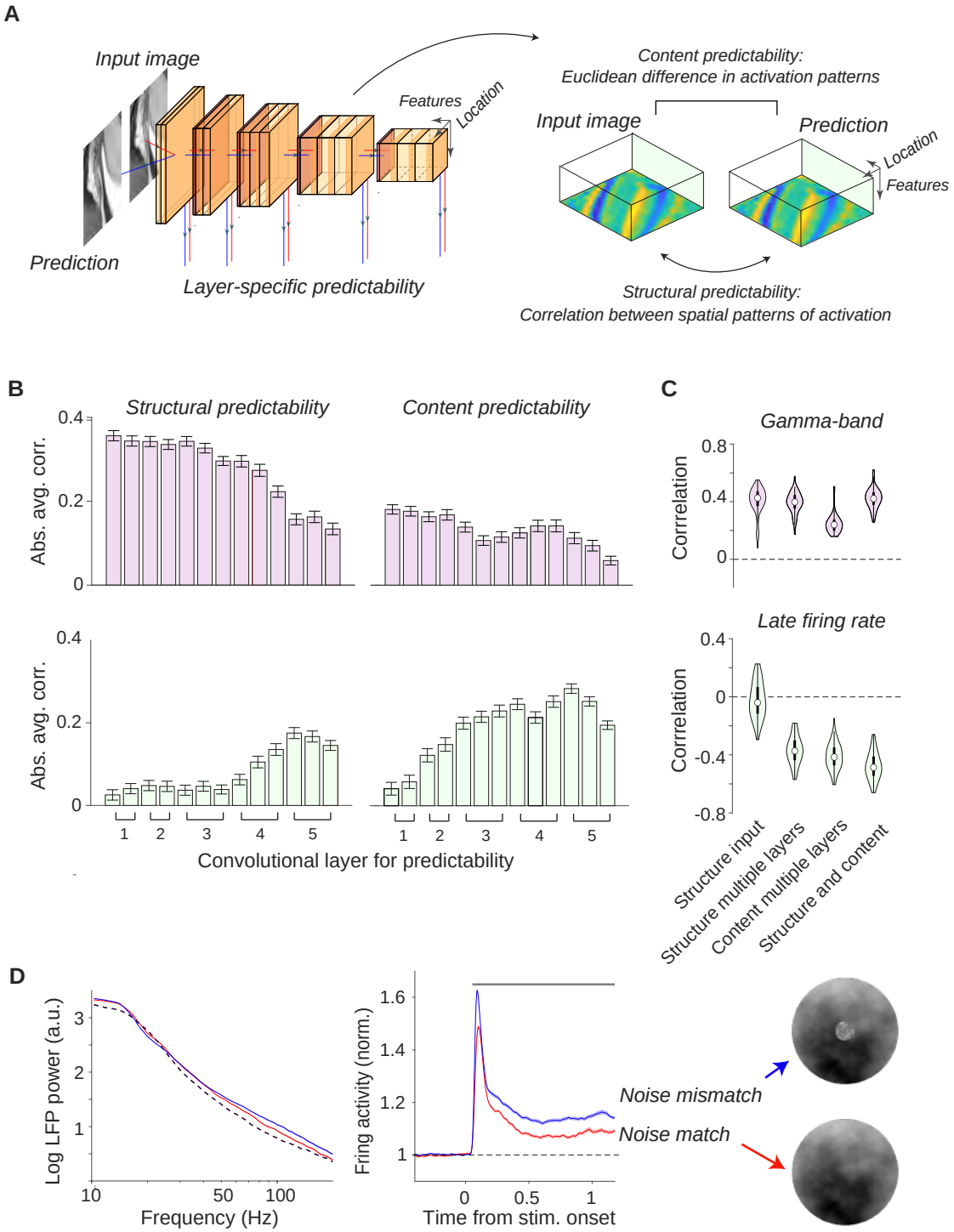

Figure 4: Modulation of V1 activity by low and high-level predictability. **(A)** Illustration of the VGG-16 network used to define low- and high-level predictability. Responses of VGG-16 units in different layers were computed both for the input (ground-truth) and predicted image-patch. For each VGG-16 layer, we then computed content (based on Euclidean distance) and structural (based on spatial correlations) predictability. **(B)** Correlation of late firing rate (200-600ms, green) and peak γ-power (red) with content and structural predictability across VGG-16 layers. Note that we first computed average correlations and show here the absolute average value of these correlations; correlations were positive for γ and negative for firing rates. Late firing rates showed a significant increase in absolute correlation across VGG-16 layers, both for structural and content predictability (structure: $R = 0.85$; content: $R = 0.81$, $P < 0.001$ for both). By contrast, γ showed a significant decrease (structure: $R = 0.9$; content: $R = 0.84$, $P < 0.001$). The average correlation (across layers) for structural predictability was significantly higher than content predictability for γ ($P < 0.001$, Paired T-test). For firing rates, the average correlation with content predictability was higher than structure ($P < 0.001$, Paired T-test). **(C)** Left-to-right: (i) Correlation with structural predictability at the input image level; (ii) regression model based on structural predictability across layers 1conv2, 2conv2, 3conv3, 4conv3, 5conv3; (iii) regression model based on content predictability in those layers; (iv) regression model based on both content and structure predictability in those 5 layers. For gamma, the full regression model did not explain more variance than structural predictability at the input level ($P = 0.40$), and structure explained more variance than content ($P < 0.001$, paired T-test). For rates, the model with content explained more variance than structure ($P < 0.001$), and the model with both content and structure explained more variance than either content or structure ($P < 0.001$ for both). **(D)** Neural activity for noise mismatch and noise match stimuli. Stimuli had white, pink or brown noise in the 1 degree center, and white, pink or brown noise in the surround. Stimulus size was 6 degrees. Stimuli were centered on the recording site's RF. Left: LFP power-spectra for noise mismatch and noise match stimuli. Note that the broad-band increase in LFP power at high frequencies is typical for spike-bleed in (Ray and Maunsell, 2011). The dashed line represents the baseline pre-stimulus period. Right: Normalized MU firing-rates for noise match and mismatch stimuli. Firing rates were significantly different between noise and noise-mismatch stimuli (gray bar: $P < 0.05$, T-test, N=16 recording sites).

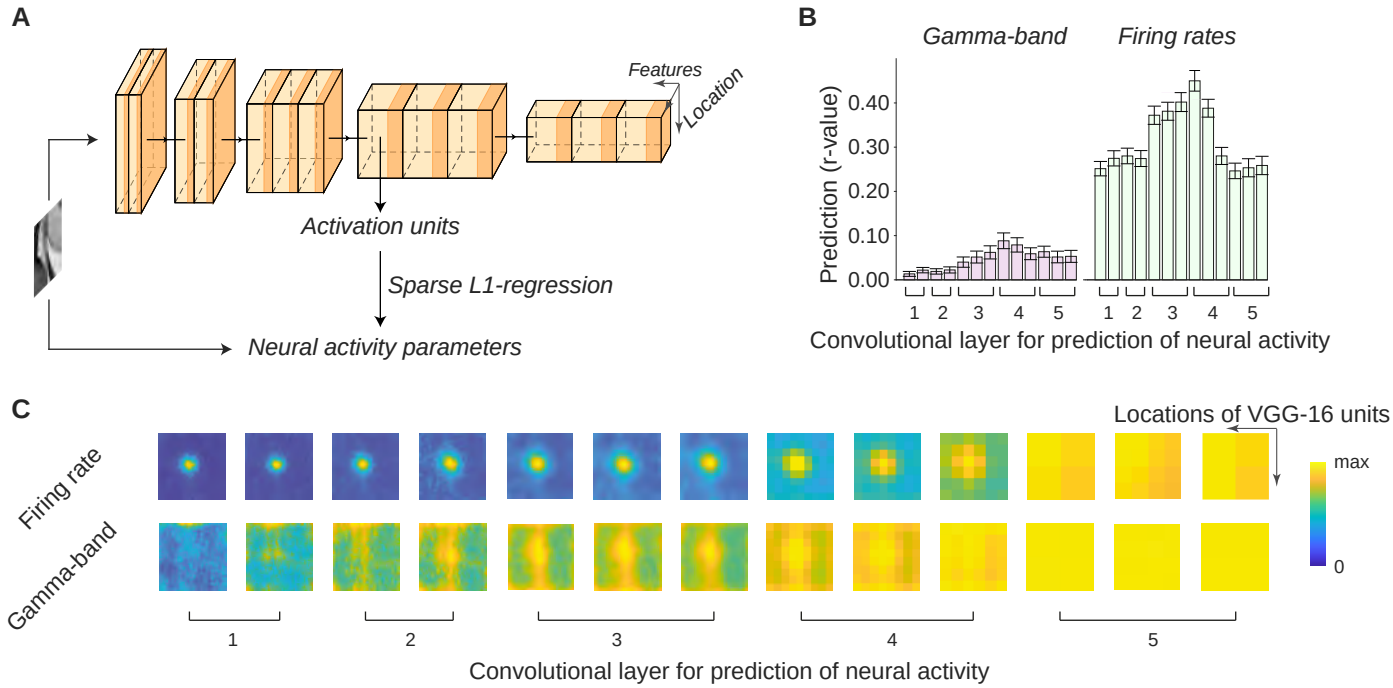

Figure 5: Prediction of neural activity based on activation of neurons in VGG-16 network for object recognition. **(A)** For each recording site, we determined different neural activity parameters. The image patch centered on the RF of the recording site was then passed into the VGG-16, and we computed the activation of every VGG-16 artificial neuron (AN). Sparse L1-regression with cross-validation was used to predict neural activity from VGG-16 ANs with RFs at the center of the image. **(B)** Prediction accuracy of different neural activity parameters depending on VGG-16 layer. Prediction accuracy for late firing rates was significantly higher for middle (5-9) than early (1-4) and deep (10-13) convolutional layers ($P < 0.001$, paired T-test). For γ, prediction accuracy was significantly higher for middle ($P < 0.001$) and deep ($P < 0.05$) than early layers. **(C)** Prediction accuracy depending on the RF location of VGG-16 ANs in the image.

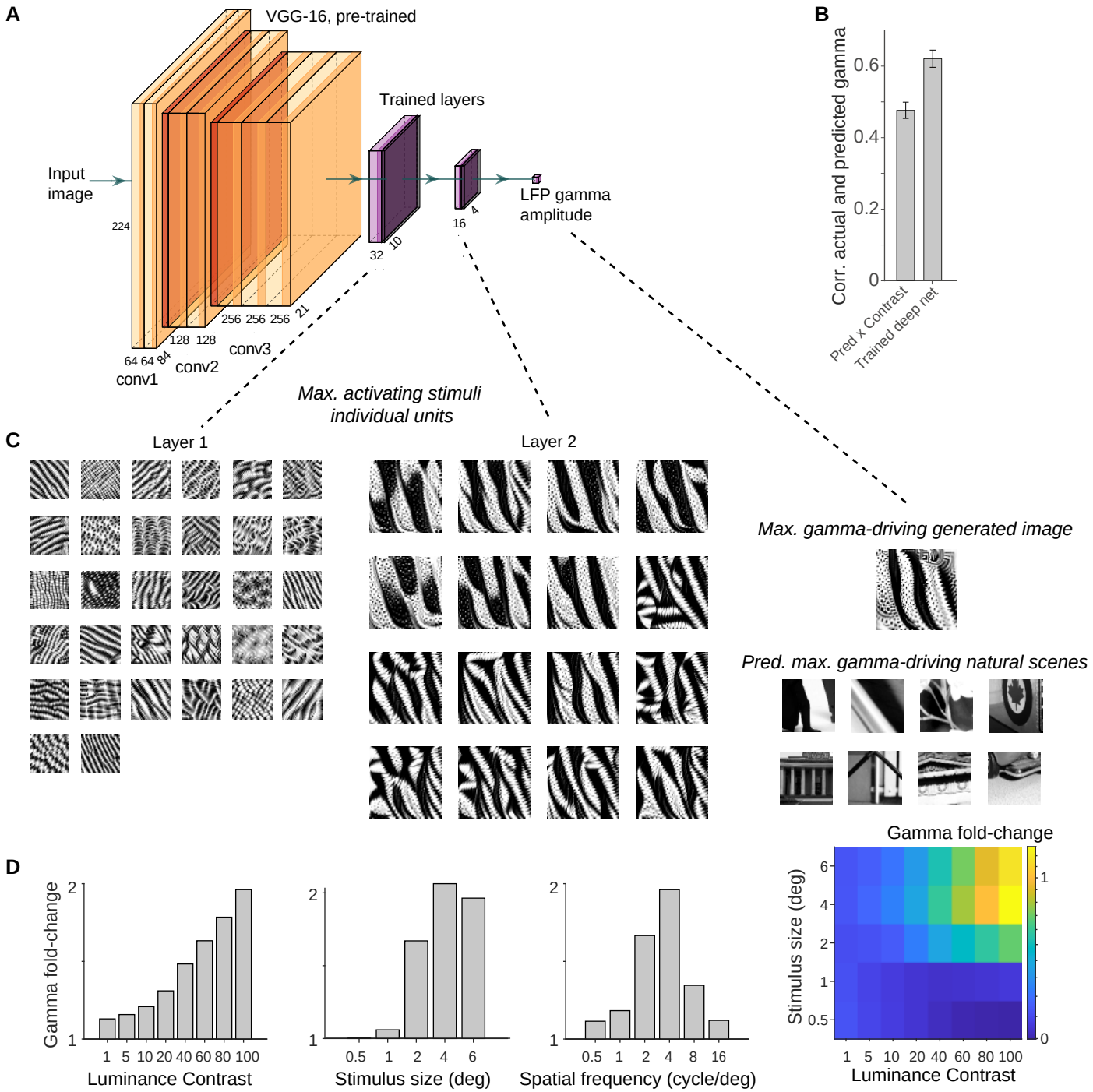

Figure 6: Prediction of gamma-synchronization with a neural network trained *de novo*. **(A)** Architecture of neural network ("gamma-net"). The image was preprocessed with the VGG-16 network, and then passed into a network with two layers that were trained *de novo* on images presented to the monkeys. The network was trained with cross-validation and dropout. **(B)** Average prediction accuracy (r-value) of gamma net. Network loss was minimized by preprocessing the image until 3conv3 of the VGG-16. SEMs were computed across cross-validation runs. For comparison, we also include the correlation of γ with the product of contrast and predictability, as shown in Figure 3D. The error bar for gamma corresponds to 25-75% quantiles across channels as shown in Figure 3D. **(C)** Left: Maximum driving images for the first layer of the gamma-prediction network. Middle: Maximum driving images for the second layer of gamma-prediction network. Right, top: Generated image that predicted maximum output (i.e. γ) in the network. This image had a grating-like structure, even though the network was never trained on gratings, but only on natural images. Right, bottom: Natural images that were predicted to yield maximum γ oscillations. **(D)** We characterized the tuning of gamma-net to gratings as a function of different parameters: Size, luminance contrast and spatial frequency. Shown are averages across 16 orientations.

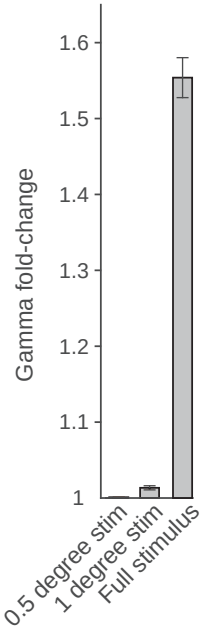

Figure Extended Data 1: Comparison of γ fold change for images at 0.5, 1 degrees of visual angle (dva) or the full natural image. In the 0.5 and 1dva conditions, we presented only part of the image (around the RF center of the recording sites) with a circular aperture of 0.5 or 1dva diameter. γ oscillations effectively disappear for small stimuli that do not extend in the surround, even though these elicit higher firing rates (Coen-Cagli et al., 2015).

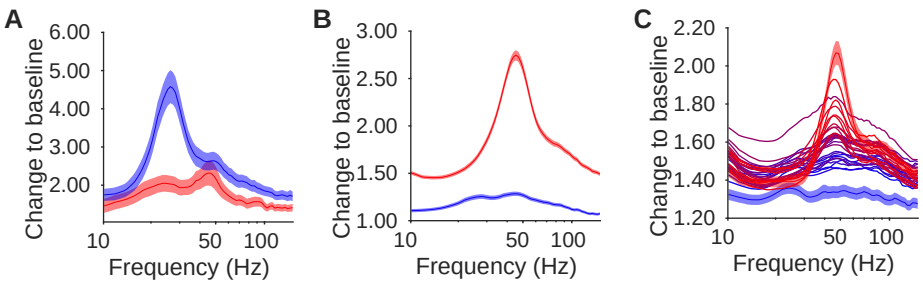

Figure Extended Data 2: **(A)** Average MUA power spectrum for sites (across channels and sessions) with 25% highest and 25% lowest β-peaks. Like the LFP, the MUA shows rhythmic fluctuations in the beta-band with a peak around 26 Hz. **(B)** Same as (A), but now for γ. **(C)** MUA power as a function of the product of spatial predictability and luminance-contrast. MUA power shows a gradual increase in γ amplitude. All panels are for monkey H.

.

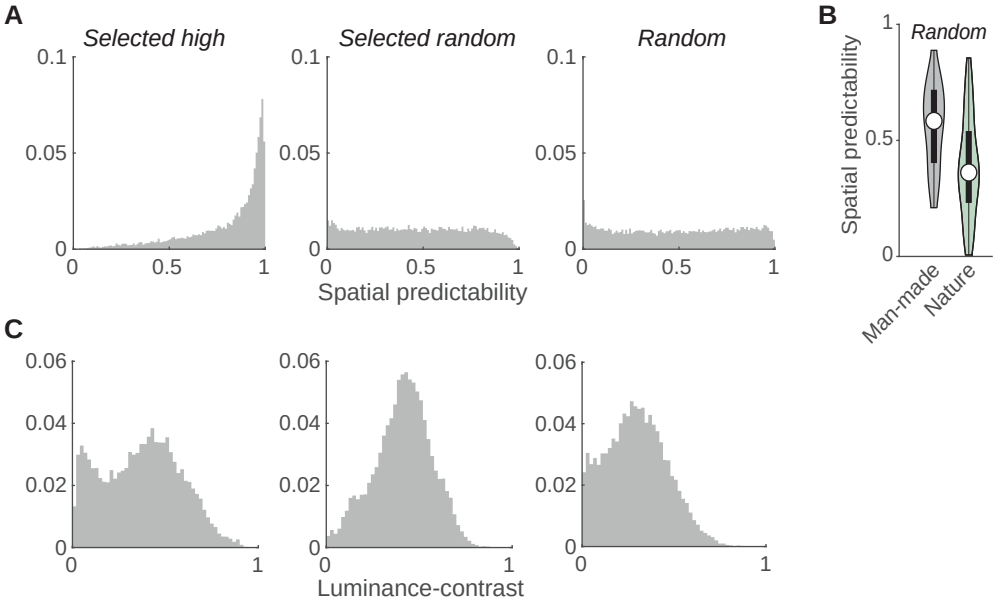

Figure Extended Data 3: **(A)** Histograms of spatial (structural) predictability for three image categories. High predictability: 10/20 images that were randomly selected with additional criteria on predictability in the RFs of the neurons. Random, selected: images that fulfilled criteria the criteria as stated in the Methods section, but had no additional criteria on spatial predictability (10/20 images per session). Random: images that fullfilled the criteria on global luminance contrast (> 0.2), luminance (between 40 and 200) and spatial frequency (average centroid greater than 0.5 degrees of visual angle). **(B)** Average structural predictability for the random images category, separate for nature and man-made contents. Shown are the median, 25/75% percentiles and data distribution across images. **(C)** Same as (A), but now for luminance contrast.

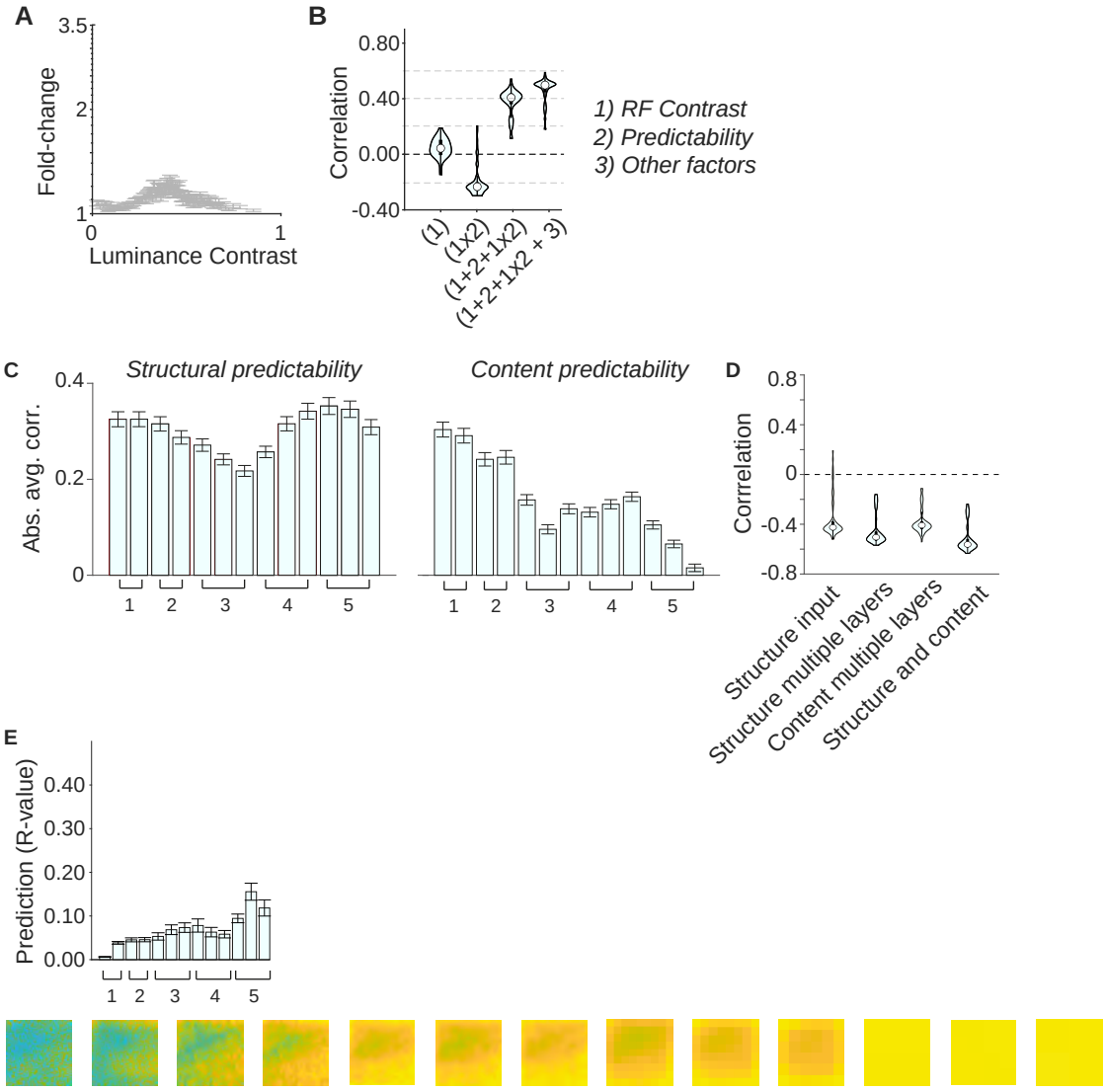

Figure Extended Data 4: Main results for β synchronization. **(A)** Average 1/f corrected β-peak amplitude for different levels of luminance contrast, similar to Figure 3B. β power peaks at intermediate luminance contrasts. **(B)** Similar to Figure 3D. Beta was very weakly correlated with luminance contrast and strongest for intermediate contrasts. Including other image factors as predictors led to a small increase in explained variance for β synchronization. **(C)** Correlation of β-power with structural (left) and content (right) predictability across VGG-16 layers. β showed a significant decrease in absolute correlation with predictability across layers for content ($R = 0.9$, $P < 0.001$). **(D)** For β synchronization, we found that explained variance was also low (peaking around 0.15) compared to firing rates. Prediction accuracy increased with VGG-16 layer, and β synchronization was best predicted from the activity of units in the deeper layers of the VGG-16 (5conv2) (similar to Figure 5) (deep vs. early and middle VGG-16 layers: $P < 0.001$, paired T-test). This is consistent with the idea that β synchronization propagates from higher areas to early visual cortex (Bressler et al., 2006, 2007; Bastos et al., 2015; Richter et al., 2018). **(E)** RFs for β synchronization did not show a localized RF structure that was related to VGG-16 units.

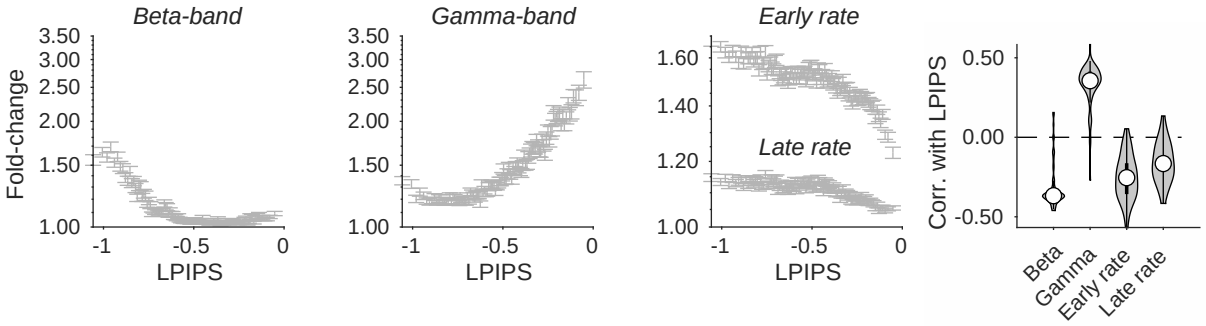

Figure Extended Data 5: Correlations of firing rates and neural synchronization with spatial predictability defined as LPIPS (Learned Perceptual Image Patch Similarity), a perceptual similarity measure based on deep neural networks for object recognition (Zhang et al., 2018). Left three figures: Same as Figure 2 Main text, but now with spatial predictability defined as LPIPS. Right: Correlations across all images, averaged across recording sites; correlations were significant for all variables ($P < 0.001$). Note that correlations for β and γ are comparable to structural predictability (Figure 2), but that the relationships for firing rates are monotonic and significant. This likely reflects the fact that LPIPS is based on VGG-16 activations across all layers, and that γ is well correlated with predictability in terms of low VGG-16 layers, whereas firing rates are well correlated with predictability in terms of deep VGG-16 layers (see Figure 4)

.

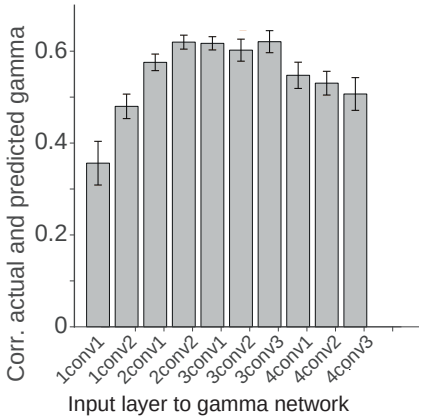

Figure Extended Data 6: Performance of gamma-net depending on the size of the pre-trained VGG-16 block that was used before the additional layers that were trained de novo. Network performance peaked when initializing the VGG-16 in middle layers.

layer	1c1	1c2	2c1	2c2	3c1	3c2	3c3	4c1	4c2	4c3	5c1	5c2	5c3
pixels	3	5	10	14	24	32	40	60	76	92	132	164	196
dva (A & I)	0.07	0.12	0.25	0.35	0.60	0.80	1.00	1.50	1.90	2.30	3.30	4.10	4.90
dva (H)	0.06	0.10	0.19	0.27	0.46	0.62	0.77	1.15	1.46	1.77	2.54	3.15	3.77

Table 1: Table showing Receptive Field sizes in the VGG-16 across layers. The abbreviation dva stands for degrees in visual angle. This was shown separately for monkey H and monkey A and I because of the differences in the number of pixels per degree.