Top-down sharpening of hierarchical visual feature representations

Authors:

Mohamed Abdelhack[1,2] and Yukiyasu Kamitani[1,2]*

Affiliations:

[1]Graduate School of Informatics, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan

[2]ATR Computational Neuroscience Laboratories, 2-2-2 Hikaridai, Seika, Soraku, Kyoto 619-0288, Japan

***Corresponding author's email address: kamitani@i.kyoto-u.ac.jp**

Abstract

The robustness of the visual system lies in its ability to perceive degraded images. This is achieved through interacting bottom-up and top-down pathways that process the visual input in concordance with stored prior information. The interaction mechanism by which they integrate visual input and prior information is still enigmatic. We present a new approach using deep neural network (DNN) representation to reveal the effects of such integration on degraded visual inputs. We transformed measured human brain activity resulting from viewing blurred images to the hierarchical representation space derived from a feedforward DNN. Transformed representations were found to veer towards the original non-blurred image and away from the blurred stimulus image. This indicated deblurring or sharpening in the neural representation, and possibly in our perception, from top-down modulation. We anticipate these results will help unravel the interplay mechanism between bottom-up and top-down pathways, leading to more comprehensive models of vision.

Keywords: human vision, top-down modulation, deep neural networks

Introduction

Perception is the process by which humans and other animals make sense of the environment around them. It involves integrating different sensory cues with prior knowledge to arrive at a meaningful interpretation of the surroundings. This integration is achieved by means of two neuronal pathways: a bottom-up pathway, which processes sensory information hierarchically, and a top-down pathway, which projects prior information down the hierarchy (Arnal & Giraud, 2012; Clark, 2013; Summerfield & de Lange, 2014). In the visual cortex, top-down signals are believed to propagate from higher visual and cognitive areas to the lower visual areas (Friston, 2005; Heeger, 2017). They cause modulation of the neural response of the sensory signal, which in turn is processed up the hierarchy again, to the higher visual areas; the process is repeated until convergence on a perceptual result. The interplay mechanism between these two pathways is still an open question. It is believed that the imbalance between the bottom-up and top-down pathways results in the hallucinatory symptoms associated with psychological disorders such as schizophrenia (for review, see Friston, Brown, Siemerkus, & Stephan, 2016; Jardri et al., 2016).

Previous studies have given rise to two main hypotheses to explain the top-down modulation process. The sharpening hypothesis states that top-down signals enhance the neural representation in the lower visual areas, thus improving the quality of the degraded sensory signal. The enhanced signals are then processed further up the hierarchy, leading eventually to a perceptual result. Several studies have shown evidence for sharpened signals driven by prediction in the lower visual areas (Gayet et al., 2017; Hsieh, Vul, & Kanwisher, 2010; Kok, Jehee, & de Lange, 2012; Lee &

Mumford, 2003). Conversely, the prediction error hypothesis (which originates from computer science ideas; Shi, Sun, & Huifang (2008) states that top-down signals provide expected signal information that is uninteresting, and therefore gets subtracted (Mumford, 1992; Rao & Ballard, 1999). This results in an error signal that is further processed to update the prediction signal. This process repeats until the error signal reaches zero, which corresponds with achieving a perceptual result. This idea was triggered by the finding of an associated reduction in brain activity in the early visual areas in response to the top-down modulation of activity (Alink, Schwiedrzik, Kohler, Singer, & Muckli, 2010; den Ouden, Friston, Daw, McIntosh, & Stephan, 2009; den Ouden, Kok, & de Lange, 2012; Gordon, Koenig-Robert, Tsuchiya, van Boxtel, & Hohwy, 2017; Kok et al., 2012; Meyer & Olson, 2011; Murray, Kersten, Olshausen, Schrater, & Woods, 2002; Todorovic, van Ede, Maris, & de Lange, 2011). It was found, however, that this reduction in the macroscopic activity level is accompanied by improved fine-scale neural representations in V1 (Kok et al., 2012). Modelling studies have generally incorporated both mechanisms in their proposed models of the interaction between bottom-up and top-down modulation, in order to achieve a Bayesian-like perception model (Friston, 2005; Heeger, 2017).

Most of the top-down modulation studies utilized expectation of a previously known visual stimuli as a method to drive the operation of top-down pathways, and have hence focused on lower visual areas.. Expectation-of-stimulus tasks facilitate comparison of a visualized stimulus and an expected stimulus at the lower visual feature level. Kok, Mostert, & De Lange (2017) have shown that expectation tasks elicit pre-stimulus neural representation of the expected stimulus in the lower visual areas, which is not

the case in normal exploratory visual recognition. While such studies have provided an empirical framework for the operation of top-down modulation driven by expectation in the lower visual areas, they have not revealed its operation, and nor have they adjudicated between sharpening and prediction error models in regular recognition-targeting visual tasks. Previous investigations on the operation mechanism of top-down modulation in recognition-targeting tasks have argued that the visual cortex generates predictions by rapid processing of the low spatial frequency components of the image (Ahissar & Hochstein, 2004; Hochstein & Ahissar, 2002) and some global structures coded in the high spatial frequencies (Oliva & Torralba, 2006). This early processing generates predictions that lead to the fine-tuning of signals that propagate down the hierarchy, enhancing sensory representations in earlier visual areas (Ahissar & Hochstein, 2004; Hochstein & Ahissar, 2002; Morgan, Petro, & Muckli, 2016; Revina, Petro, & Muckli, 2017).

In this study, we tackle this question by investigating top-down pathway operation during a natural-image visual recognition task. We drive the operation of top-down modulation by applying degradation to natural images by blurring them. When visual images are degraded, the visual sensory signal is less reliable, and the visual cortex therefore depends more heavily on prior knowledge driving the top-down pathway operation. It also depends on others sensory cues, however, in our setting we do not provide any other sensory information about the stimulus so the subject has to depend solely on prior knowledge. To unmask the top-down effect, we investigate how the neural representations of viewing blurred images deviate from a pure feedforward representation. A sharpening mechanism would cause the neural representations to try

to mimic the expected stimulus and thus cause the deviation to veer towards non-blurred representations. Conversely, a prediction error mechanism would cause the expected neural representations to be subtracted from the input stream resulting in an opposite deviation. Moreover, successful recognition in the sharpening realm would lead to even more similarity with the non-blurred representations, as opposed to the opposite occurring under the prediction error hypothesis where the resulting prediction error would reach a value of zero.

To make such a comparison, we measured and analyzed brain activity from functional magnetic resonance imaging (fMRI) brain data from different regions of the lower and higher visual areas, to visualize the degradation effect on different levels of neural processing. We utilized deep neural network (DNN) feature space as a proxy for hierarchical representation. This was introduced earlier by Horikawa & Kamitani (2017a) to build a generic object decoder and show the architectural homology between the visual processing system and DNNs. Moreover, DNNs for image classification were shown to be a good representation of hierarchical processing, and to resemble the hierarchical neural processing in the visual cortex (Cadieu et al., 2014; Güçlü & van Gerven, 2015; Horikawa & Kamitani, 2017b; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). In this study, we use a feature decoding method devised by Horikawa & Kamitani (2017a) to predict DNN representations from brain activity. We then further isolate the effect of top-down modulation by subtracting the hierarchical bottom-up processing component inherent in the feedforward-only DNN model chosen. We also use this method to find the level of visual processing at which the top-down effect is most visible. As the DNN model we use is feedforward only, it lacks a feedback

–or top-down– system; the representations extracted from it are representative of a bottom-up-only system. The deviation of decoded features from the brain data of these representations could be attributed to a top-down modulation effect which is caused by the image degradation. The decoded features are analyzed for their similarity to the feedforward-only DNN features of the stimulus images and original non-blurred images. These similarities are then compared to their counterpart noisy DNN features, which account for decoding errors as a baseline for pure-feedforward behavior, to find whether predicted features deviate towards those of original images or away from them. We also compare the case where the image content is successfully recognized with the one where it is not. If image sharpening were in operation, it would be expected that the top-down effect would be boosted due to successful perception.

Results

DNN feature decoding

We posed a question on how top-down modulation in the visual cortex affects the neural representation of blurred images. To address this question, we measured brain activity while presenting blurred images. The protocol involved the presentation of stimuli in blurred-to-original image sequences. Each sequence consisted of stimuli showing different blur levels of the same image presented in the order of the most blurred to the non-blurred original image (Figure 1A). Subjects vocally reported the perceived object in each stimulus, while also reporting their certainty of their perception. We conducted two experiments using this protocol. In the first experiment, each image (stimulus sequence) was chosen from a random object category and the subject had no prior information of the object category (no-prior condition). In the second experiment, the stimulus sequences were chosen from five predefined object categories (airplane, bird, car, cat, and dog). The subjects were informed about the object categories of the set, but not of each stimulus (category-prior condition). Using these two conditions we can analyze the effect of adding prior information on the top-down effect in different visual areas.

To examine the effect of top-down modulation, we investigated the neural representation of blurred images via the proxy of a hierarchical feedforward-only representational space (Horikawa & Kamitani 2017a). Deep neural network (DNN) models provide this capability as they run in a hierarchical feedforward fashion. We extracted feature values of the stimulus images using AlexNet (Krizhevsky, Sutskever, &

Hinton, 2017), a standard DNN implementation for object classification. This consists of eight layers with a hierarchical structure, and emulates the structure of the visual cortex. The first five layers (DNN1–5) are convolutional while the last three (DNN6–8) are fully connected. The DNN was trained to classify images into one of 1000 categories defined in the ImageNet large-scale object recognition challenge (http://www.image-net.org/challenges/LSVRC/).

To transform brain data into the DNN feature space, we trained multivoxel decoders to predict DNN features from brain activity data. A separate training stimulus dataset consisting of 1000 natural non-blurred images was used to train those decoders, with one image for each classification category of the DNN. We selected 1000 features from each layer of the DNN and constructed a decoder for each feature, totaling 8000 decoders. Using a sparse linear regression algorithm (SLR; Bishop 2006), each decoder was trained to use the fMRI responses to these images to predict the feature values (normalized to a mean value of zero in each feature). Note that the decoders were trained only on independent non-blurred images.

Using the trained decoders, the brain activity pattern induced by each stimulus in the blurred-to-original sequences was decoded (transformed) into the DNN feature space. For each stimulus image, the Pearson correlation coefficient between its decoded feature vector and the true features of the same stimulus image (r_s) at each layer was computed. In addition, the correlation between the decoded feature vector and the true features of the corresponding non-blurred original image (r_o) was computed (Figure 1B). For non-blurred stimuli, r_s and r_o are identical.

Feature gain computation

The correlation with stimulus image features (r_s) reflects the degree to which image features resulting from feedforward processing are faithfully decoded from brain activity, while the correlation with original images (r_o) reflects the degree to which the decoded features are "sharpened" by top-down processing, to be similar to those of the non-blurred images. Figure 2A shows a scatter plot depicting a representative result for prediction of the DNN layer 6 feature vector from the region of interest comprised of all the visual areas (Visual cortex, VC) of Subject 4. DNN6 is a higher middle layer of AlexNet where we can visualize the top-down effect on mid-level representations of visual stimuli. Each point represents a stimulus image pooling both category-prior and no-prior conditions, and disregarding behavioral data. Figure 2B shows the mean of the results in Figure 2A, grouping stimuli by different blur levels. From the results of r_s and r_o, we define Δr_{decode} as the difference between them. We notice from the representative data that decoded features have higher correlation with the original image features than with the stimulus image features, except when the blurring effect becomes too large, as in the 25% blur level. This suggests that a sharpening effect occurring in the visual cortex causes the neural representations of viewing the blurred image to mimic those of a less blurred version of it.

One shortcoming of this measure (Δr_{decode}) is that it does not have an appropriate baseline for sharpening. A value of Δr_{decode} equal to zero implies that decoded features are equally similar to stimulus and original image features, but it does not mean that there is no sharpening. Thus, we defined a baseline for no sharpening according to the

behavior of feedforward-only processing. Decoded features from feedforward-only processing were modeled by stimulus image features plus Gaussian noise. The noise level was determined to match the decoding errors with the non-blurred images used as stimuli, in which no sharpening was assumed to be involved. Noise was added to the point where the decoded and noise-added features had nearly identical correlations to the original image features (Figure 2B and C; 0% blur level in each). The same level of noise (the mean across images in each subject and DNN layer) was added to the stimulus features of the blurred images. We then computed r_s and r_o for the noise-matched features, from which we could obtain the noise-matched baseline Δr_{noise} (Figure 2C).

By comparing B and C in Figure 2, it is possible to note an opposite trend in how the features are correlated. As the decoder was trained to predict image features, the natural trend for Δr_{decode} would be negative, similar to Δr_{noise}. This indicates a level of alteration in the neural representation of the blurred images, to improve the match with the original images.

By subtracting the noise baseline from Δr_{decode}, we obtained the “feature gain” incurred by top-down processing. The value of the feature gain indicates how the top-down pathways affect the predicted features in comparison with pure feedforward behavior. Figure 2D shows the results of the mean feature gain for different subjects for each layer. We can observe positive significant feature gains for most of the DNN layers and blur levels (17 out of 24 DNN layer/blur level combinations; *t*-test across subjects, $p < 0.05$). This suggests that top-down processing modulates neural representations to bias

them towards the original images. We also noticed that the fully-connected layers DNN6–8 had more pronounced positive feature gains than the convolutional layers. Another notable issue is that the 12% blur level shows better feature gain relative to both 6% and 25% blur levels in higher visual areas. One possible explanation is that 12% blur level is situated in a middle ground between the other levels where 6% blur level is quite similar to the original image while 25% blur level is too degraded to provide any useful information.

One possible cause for this result is the training scheme of the decoders, which only used natural non-blurred images. This could have biased the output features to those resembling natural images. In this case, the features could be correlated to any natural image features. We investigated this possibility by measuring the content specificity of the predicted features. We computed the correlation of predicted features (excluding those with a 0% blur level) with the corresponding original image feature (r_o). This was then compared with the mean correlation of the same predicted features, but with the original features of different images. This measure provided information on how tightly the predicted features were associated with the presented stimulus content, as opposed to natural images in general. Figure 3 shows the result of such a content specificity analysis. There are significant differences between correlations with the same image features and mean correlations with different image features in all layers (*t*-test across subjects, $p < 0.05$), indicating a tight association of the predicted features with the stimulus image content, and ruling out a decoder bias explanation.

As mentioned before, the DNN model used in this study implements hierarchical

processing that is synonymous with that happening in the visual cortex. Previous studies have shown homology between the features of the DNNs and the representations in the visual cortex (Cadieu et al., 2014; Güçlü & van Gerven, 2015; Horikawa & Kamitani, 2017a; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). To this point, we have shown the results of features predicted from the collection of all denoted visual areas (VC). We further investigated the separate visual areas of the lower, intermediate, and higher visual areas, to examine the homology between the feature gain and the visual cortex hierarchy (Figure 4). We showed that the feature gain also shows similar homologies to the visual hierarchy, in that we could observe that shallower DNN layers showed larger feature gain from the lower visual areas (V1–3), while deeper DNN layers showed larger feature gain from the higher visual areas (LOC, FFA, PPA). The results are significantly positive for most of the layers and regions of interest (ROI), especially in the higher visual areas and fully connected layers (*t*-test, $p < 0.05$). However, DNN1 did not show significantly positive feature gains. These results imply that feature gain also follows the same visual homology in the visual cortex areas, and that the top-down effect is more pronounced in higher visual areas.

Effects of prior knowledge and recognition

In the previous analyses, the data from different experimental conditions were pooled together. We then further investigated the difference between the category-prior and no-prior conditions. We compared the feature gain means grouped according to the experimental condition (category-prior vs. no-prior) while pooling all the behavioral responses (Figure 5). We performed two-way ANOVA on the feature gain data using the ROI and the experimental conditions as the independent variables. The addition of a

prior caused significant enhancement to the feature gain in layers DNN4–8 ($p < 0.05$). The difference was most pronounced in DNN8 ($p = 0.000014$). This result indicates that addition of prior information enhances top-down modulation, thereby causing an increase in feature gain. This implies augmented sharpening of neural representations.

This result, however, pooled both correctly and incorrectly reported results. When considering behavioral data, there are considerable differences between category-prior and no-prior conditions. The category-prior condition was characterized by a higher number of correct responses (235 out of 300 total instances for five subjects) compared with the no-prior condition (92 out of 300 total instances for five subjects). However, in the category prior condition, the task was to choose one of five categories. This could lead to false positives, as if a subject responded in a random manner, 20% of the responses would be likely to be correct. In some cases when the stimulus was highly degraded, the best guess response by the subjects could be random. To attempt to curb this problem, we could use the certainty level as an indicator of correctness, especially for the category prior. We found from the behavioral results that nearly all the trials labelled as certain were also correctly recognized (category prior: 138 out of 139 certain trials were correct; no-prior: 57 out of 70 certain trials were correct). This further supports the observation that adding priors aids recognition.

We further analyzed our data by grouping it according to both experimental condition (category-prior and no-prior) and recognition performance (correct and incorrect). We show the results of the mean feature gain over subject means for each DNN layer in Figure 6. For each experimental condition, we performed a three-way ANOVA test using

ROI, recognition performance, and blur level as independent variables. For the category-prior condition, we found significant enhancement in feature gain when an image was correctly recognized in DNN4–6 and 8, while for the no-prior condition significant enhancement was found in DNN2. From these results, we notice that the effect of recognition leads to enhancement in the feature gain, which is most pronounced in the intermediate convolutional layers of the DNN.

We also analyzed our data by grouping it according to certainty level (certain and uncertain). We show the results of mean feature gain over subject means for each DNN layer of this analysis in Figure 7. For the category-prior condition, when an image was recognized with certainty we found significant enhancement in feature gain in DNN2, 4, 5, and 8, while for the no-prior condition significant enhancement was found in DNN1.

From the results of Figures 6 and 7, we can observe that in some layers and conditions, recognition has a significant boosting effect on feature gain. However, we also found a considerable feature gain even without recognition, which indicates a sharpening effect that is not guided by subjective recognition. This could be caused by a lower-level sharpening associated with local similarity or object component sharpening, which could be common across different objects (like body parts in animals).

Discussion

In this study, we have demonstrated that top-down modulation induces sharpening of the neural representations of blurred visual stimuli. This sharpening can assist the visual system in achieving successful prediction. Compared with pure-feedforward behavior, the neural representations of blurred images tended to be biased towards those of corresponding original images, even though the original images had not yet been viewed. This sharpening effect was also found to follow a visual hierarchy similar to that in the visual cortex. We found that this sharpening was content-specific, and not just due to a natural image bias. It was also shown to be boosted by giving category information to the subject prior to stimulus viewing. This indicated that adding a more specific prior does not lead to the expected signal being explained away as a prediction error hypothesis. However, we did not find that recognition had a strong role in boosting the enhancement process.

In our experimental protocol, the subjects viewed blurred stimuli in randomly organized sequences. In each sequence, different levels of blur of the same image were shown, ordered from the most blurred to the non-blurred stimulus (Figure 1A). This ensured that subjects did not have pixel level information. Nonetheless, the results show a tendency for the blurred images' neural representations to correlate with the original images (Figure 2A and B). Conversely, the feedforward behavior demonstrated by the noisy DNN output showed an opposite tendency (Figure 2C). We computed the feature gain to investigate how the predicted DNN features deviated from pure feedforward behavior. Feature gain analysis showed that the predicted features are rather correlated with the original image features (Figure 2D). This indicates that a sharpening effect

happens across the visual cortex, leading to a more natural-image-like neural representation. This representation was also confirmed as not being due to a natural-image bias caused by the decoder training dataset, which consisted of natural unaltered images (Figure 3). These results are in line with previous studies showing that neural representations are improved due to a top-down effect (Gayet et al., 2017; Hsieh et al., 2010; Kok et al., 2012; Lee & Mumford, 2003). Kok et al. (2012) demonstrated that even though the overall neural activation weakens, the neural representations improve when a stimulus is in agreement with the expectation. Gayet et al., (2017) also showed that visual working memory enhances the neural representations of viewed stimuli. The reverse hierarchy theory also suggests that top-down modulation serves to fine-tune sensory signals by means of predictions initially made using lower spatial frequency features (Ahissar & Hochstein, 2004; Hochstein & Ahissar, 2002). Furthermore, Revina et al. (2017) showed that blurred stimuli can generate top-down processes that generalize to higher spatial frequencies.

Our analysis also shows that feature gain follows a similar hierarchy to the visual cortex (Figure 4). This indicates that the sharpening process occurs in the same hierarchical processing localization as normal processing where low level sharpening occurs in lower visual areas and enhancement of higher level features occurs in the higher visual areas. It could be indicative of a convergent mechanism by which bottom-up and top-down pathways are integrated into a single neural representation of the stimulus. This suggestion could be supported by previous reports on the prediction of visual features. Horikawa & Kamitani (2017a) demonstrated that visual perception and mental imagery yielded feature prediction that was homologous with that of the visual cortex hierarchy.

Horikawa & Kamitani (2017b) also showed similar results from dream induced brain activity. Earlier studies showed strong representational similarities between the deeper layers of DNN and the brain activity in the inferior temporal cortex (IT; Cadieu et al., 2014; Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al. 2014). This could be further investigated by high resolution imaging to reveal the layered structures in the visual cortex and analyze the neural representations in each layer. We demonstrate that top-down effects also show similar homology, thus suggesting that DNN-based methods are useful for studying visual top-down pathways since it can reveal the localization of the sharpening by means of DNN layer feature gain.

When we added a category-prior to the task, the number of competing categories for recognition decreased, thus the subjects tended to have a more directed top-down effect, due to the fewer number of competing stimuli (Bar & Aminoff, 2003). This led to a higher feature gain, which was especially noted in the higher layers (Figure 5). This further supports the idea of neural representation sharpening when given a prior describing the stimulus content, as the top-down signal would be more correlated with the correct recognition results, thus leading to a stronger feature gain.

We also found that when subjects successfully recognized the image content, the feature gain in some layers predicted from lower visual areas was significantly improved. However, this was not salient as a general trend (Figures 6, 7). It was expected that recognition would lead to a boost in feature gain from the sensory competition perspective, as the subject would attend to the successfully recognized

object in the stimulus image, leading to a directed top-down effect (Kastner, De Weerd, Desimone, & Ungerleider, 1998; Kastner, Pinsk, De Weerd, Desimone, & Ungerleider, 1999; Moran & Desimone, 1985). Hsieh et al. (2010) also showed that successful recognition of binary images leads to a neural representation that is more correlated with the natural image, although in their study, recognition was driven by ground truth image viewing before watching the degraded stimulus at a later time.

Our results could be justified by the findings in de-Wit, Kubilius, Wagemans, & Op de Beeck (2012), where top-down prediction was shown to be topologically inaccurate as it led to activity reduction in the whole of V1, rather than at the predicted location. Revina et al. (2017) also found that brain patterns due to top-down modulation did not share information with the corresponding bottom-up signals. In the prediction error realm, successful prediction would lead to zero error in the higher visual areas, and thus feature gain would decrease. Our results show an opposite, albeit weak effect, which nonetheless supports the representation-sharpening hypothesis and rules out the prediction error hypothesis. The sharpening effect without recognition may be driven by more localized and lower-level feature mechanisms. These mechanisms would enhance features corresponding to local components of the main objects that were common across many objects (i.e. eyes in animals). These local enhancement effects could lead to different recognition results. This was shown to be true for computer vision DNN-based deblurring algorithms, where the results of the enhancement process can lead to different results, according to the desired object (Bansal, Sheikh, & Ramanan, 2017).

From these results, we can deduce that top-down modulation is in operation when visual input is degraded, even in the absence of a memory or expectation prior. Previous studies have proposed that the brain makes an initial processing step using low spatial frequency information, which generates predictions of the content of the image in the orbitofrontal cortex; these predictions are then used to drive the top-down modulation effect (Bar et al., 2006; Bar & Aminoff, 2003; Breitmeyer, 2014; Kveraga, Boshyan, & Bar, 2007). This top-down effect comes about in the form of sharpening of neural representations resulting from viewing degraded images. The mechanisms by which this effect materializes have been mostly overlooked in previous literature, due to the difficulty in finding a baseline for measurement. There has been more focus on the source of this top-down modulation effect than on how it materializes in the visual cortex (Bar et al., 2006; Chiou & Lambon Ralph, 2016). As we demonstrate here, the DNN representations could offer a plausible proxy for representing brain activity and for attaining a pure-feedforward baseline that can be used for measuring top-down effects. The illustrated enhancement was shown to be affected by the presence of prior semantic information, leading to a boost in the enhancement effect, which was more visible in higher-level features. To the contrary, successful recognition did not also cause an overall boost in neural representation enhancement. Our results contribute to the long-standing question of top-down pathways affect bottom-up signals to achieve successful perception. We have examined the question from a more general perspective of vision, which has allowed us to achieve a more comprehensive understanding of the vision process.

Methods and materials

Subjects

Five healthy subjects (three males and two females, aged between 22 and 33) with normal or corrected-to-normal vision took part in the fMRI experiments. The study protocol was approved by the Ethics Committee of ATR. All the subjects provided written informed consent for their participation in the experiments.

Visual stimuli

Both original and blurred image stimuli were shown. The images were selected from the ImageNet online database (Jia Deng et al., 2009), which is the database used for training, testing, and validation of the pre-trained DNN model used in this study (see below). The database contains images that are categorized by a semantic word hierarchy organized in WordNet (Fellbaum, 2012). First, images with a resolution lower than 500 pixels were excluded, then the remaining images were further filtered to select only those that showed the main object at or close to the midpoint of the image. The selected images were then cropped to a square that is centered on the midpoint. If no acceptable image remained after this filtration process, another image was obtained from the worldwide web through an image search.

We created three different levels of blurring for the blurred image stimuli. Blurring was conducted by running a square-shaped averaging filter over the whole image. The size of the filter relative to the image size dictated the degradation level. The three degradation filters used had a side length of 6%, 12%, and 25% of the side length of the stimulus image.

We then added the original stimulus image represented by a level of 0% (Figure 1A).

Experimental design

Experiments were divided into: 1) the decoder training runs where natural undegraded images were presented, and 2) the test image runs where the blurred images were presented. In the decoder training runs, we selected one stimulus image for each of the AlexNet classification categories defined in the last layer, resulting in a total of 1000 stimulus images. This training stimulus set selection was conducted to avoid any bias to certain categories in the decoder. This dataset was divided into 20 runs of 50 images each. The subject was instructed to press a button when the image was a repeat of the image shown one-back. In each run, 5 of the 50 images were repeated in the following trial to form the one-back task. Each image was shown once to the subject (except for the one-back repetitions).

The test image runs consisted of two conditions. In the first condition, the subjects did not have any prior information about the stimuli presented (no-prior condition). In the second condition, the subjects were provided a semantic prior in the form of category choices (category-prior condition). The stimuli in the category-prior condition consisted of images from one of five object categories (airplane, bird, car, cat, or dog). The subject was informed of these categories prior to the experiment, but not the order in which they were to be presented.

The stimuli in both of the test conditions were presented in sequences of maximum blurring to original image (25%, 12%, 6%, and 0% blurring). Each sequence consisted

of stimuli representing all four levels of blurring of the same original image. We selected this order of presentation to avoid the subjects having a memory-prior of the less blurred stimuli when viewing the more blurred ones. For each condition, the sequences for 80 images were randomly distributed across two runs (40 images each). The runs belonging to the same experimental condition were conducted in the same experimental session.

All image presentation was performed using Psychtoolbox (Kleiner et al., 2007). Each image (12 × 12 degrees) was presented in a flashing sequence for 8 s at 1 Hz (500 ms on time). Images were displayed in the center of the display with a white central fixation point. The fixation point changed from white to red 500 ms before each new stimulus appeared. A 32-s pre-rest and 6-s post rest period were added at the beginning and end of each run respectively. Subjects were required to fixate on the central point. For test runs, subjects were required to provide voice feedback of their best guess of the perceived content of the stimulus. They were also required to report the certainty level of that guess by pressing one of two buttons, one indicating certainty and the other indicating uncertainty.

MRI acquisition

FMRI data was collected using a 3-Tesla MAGNETOM Verio (Siemens Medical Systems, Erlangen, Germany) MRI scanner located in Kokoro Research Center, Kyoto University. For image presentation experiments, an interleaved T2*-weighted multiband accelerated EPI scan was performed to obtain images covering the whole brain. The scanning parameters were TR = 2000 ms; TE = 43 ms; flip angle = 80°; FOV =

192 × 192 mm; voxel size = 2 × 2 × 2 mm; slice gap = 0 mm; number of slices = 76; multiband factor = 4. For localizer experiments, an interleaved T2*-weighted gradient-EPI scan was performed with the following parameters TR = 3000 ms; TE = 30 ms; flip angle = 80°; FOV = 192 × 192 mm; voxel size = 3 × 3 × 3 mm; slice gap = 0 mm; number of slices = 46. For retinotopy experiments, an interleaved T2*-weighted gradient-EPI scan was also performed where the scanning parameters were TR = 2000 ms; TE = 30 ms; flip angle = 80°; FOV = 192 × 192 mm; voxel size = 3 × 3 × 3 mm; slice gap = 0 mm; number of slices = 30. T2-weighted turbo spin echo (TSE) images with the same slice positions as the EPI images were also acquired, to act as high-resolution anatomical images. The parameters for the anatomical sequences matching the image presentation acquisition were TR = 11 000 ms; TE = 59 ms; flip angle = 160°; FOV = 192 × 192 mm; voxel size = 0.75 × 0.75 × 2.0 mm; slice gap = 0 mm; number of slices = 76. For the localizer experiment, the TSE parameters were TR = 7020 ms; TE = 69 ms flip angle = 160°; FOV = 192 × 192 mm; voxel size = 0.75 × 0.75 × 3.0 mm; slice gap = 0 mm; number of slices = 48. For the retinotopy TSE acquisition the parameters were TR = 6000 ms; TE = 58 ms; flip angle = 160°; FOV = 192 × 192 mm; voxel size = 0.75 × 0.75 × 3.0 mm. T1-weighted magnetization-prepared rapid acquisition gradient-echo (MP-Rage) fine-structural images of the entire head were also acquired. The scanning parameters for these were TR = 2250 ms; TE = 3.06 ms; TI = 900 ms; flip angle = 9°; FOV = 256 × 256 mm; voxel size = 1 × 1 × 1 mm number of slices = 208.

MRI data preprocessing

After rejection of the first 8 seconds of each acquisition to avoid scanner instability effects, the fMRI scans were preprocessed using SPM8

509 (http://www.fil.ion.ucl.ac.uk/spm), including 3D motion correction, slice-timing correction,
510 and co-registration to the appropriate high resolution anatomical images. Both scans
511 were then also co-registered to the T1 anatomical image. The EPI data were then
512 interpolated to 2 × 2 × 2 mm voxels and further processed using Brain Decoder Toolbox
513 2 (https://github.com/KamitaniLab/BrainDecoderToolbox2). Volumes were shifted by 2 s
514 (1 volume) to compensate for hemodynamic delays, then the linear trend was removed
515 from each run and the data were normalized. As each image was presented for 8 s, it
516 was represented by four fMRI volumes. These four volumes were then averaged to
517 provide a single image with increased signal to noise ratio for each stimulus image. The
518 averaged voxel values for each stimulus block were used as an input feature vector for
519 the decoding analysis.
520

521 **Region of interest construction**

522 Regions of interest (ROIs) were created for several regions in the visual cortex,
523 including the lower visual areas V1, V2, and V3, the intermediate area V4, and the
524 higher visual areas consisting of the lateral occipital complex (LOC), parahippocampal
525 place area (PPA), and fusiform face area (FFA).
526 First, anatomical 3D volumes and surfaces were reconstructed from T1 images using
527 the FreeSurfer reconstruction and segmentation tool
528 (https://surfer.nmr.mgh.harvard.edu/). To delineate the areas V1–4, a retinotopy
529 experiment was conducted following a standard protocol (Engel et al., 1994; Sereno et
530 al., 1995) involving a rotating double wedge flickering checkerboard pattern. The brain
531 activity data for this experiment was analyzed using the FreeSurfer Fsfast retinotopy
532 analysis tool (https://surfer.nmr.mgh.harvard.edu/fswiki/FsFast). The analysis results

were visually examined and ROIs were delineated on a 3D inflated image of the cortical surface. Voxels comprising areas V1–3 were selected to form the lower visual cortex (LVC) ROI.

Functional localizer experiments were conducted for the higher visual areas. Each subject undertook eight runs of 12 stimulus blocks. For each block, intact and pixel-scrambled images of face, object, and scene categories were presented in the center of the screen (10 × 10 degrees). Each block contained 20 images from one of the previous six categories. Each image was presented for 0.3 s followed by 0.45 s of blank gray background. This led to each block having a duration of 15 seconds. Two blocks of intact and scrambled images of the same category were always displayed consecutively (the order of scrambled and intact images was randomly chosen), followed by a 15- s rest period with a uniform gray background. Pre-rest and post-rest periods of 24 s and 6 s respectively were added to each run. The brain response to the localizer experiment was analyzed using the Freesurfer Fsfast event related analysis tool. Voxels showing the highest activation response to intact images for each of the face, scene, and object categories in comparison with their scrambled counterparts were visualized on a 3D inflated image of the cortical surface and delineated to form FFA, PPA, and LOC regions respectively. Voxels constituting the areas FFA, PPA, and LOC were then selected to form the higher visual cortex (HVC) ROI and the aggregation of LVC, V4, and HVC was used to form the visual cortex (VC) ROI.

Selected ROIs for both retinotopy and localizer experiments were transformed back into the original coordinates of the EPI images.

Deep neural network model

The neural representations were transformed into a deep neural network (DNN) feature proxy using the AlexNet DNN model (Krizhevsky et al., 2017). The Caffe implementation of the network packaged for the MatConvNet tool for MATLAB (Vedaldi & Lenc, 2015) was used for implementation. This network was trained to classify 1000 different image categories with images from the ImageNet database. The model consisted of 8 layers; the first five of which were convolutional layers, while the last three were fully connected layers. One thousand features were extracted from each layer, with the features with the highest feature decoding accuracy according to the mean accuracy of the five subjects' data in Horikawa & Kamitani (2017a) being selected. All the feature units in the last layer were selected, as this layer contained 1000 units in total. The features from each layer were labelled as DNN1–DNN8.

DNN feature decoding

Multiple linear regression decoders were constructed to predict each feature extracted from the voxels of each ROI from the layers of the DNN. The decoders were constructed using sparse linear regression (SLR; Bishop 2006). This algorithm assumes that each feature can be predicted using a sparse number of voxels and selects the most significant voxels for predicting the features (For details, see Horikawa & Kamitani 2017a).

A decoder was constructed for each feature. Voxel selection was undertaken for each ROI, to select the 500 voxels with the highest correlations with each feature value. FMRI data and features of the training image dataset were first normalized to a zero mean with one standard deviation. The mean and standard deviation values subtracted

were also recorded. The decoders were then trained on the normalized fMRI data and DNN features. The recorded mean and standard deviation from the training fMRI data were then used to normalize the test data before decoding the features. The resulting features were not denormalized, to avoid the effect of baseline correlation in the subsequent data analysis. For the correlation analysis, the feature vectors emanating from the DNN were normalized using the mean and standard deviation of the training dataset, to match the predicted features. These normalized feature vectors are referred to as “true” feature vectors in this study.

The feature pattern correlation was computed for each stimulus image by aligning the predicted 1000 features from each DNN layer and computing their Pearson correlation coefficient with the corresponding true feature vector.

Noise matching

Gaussian noise was added to the true features to simulate errors in DNN feature decoding. The true feature power was first calculated, then white Gaussian noise was added at varying levels of signal-to-noise ratio (SNR). The correlation between the noisy and noiseless features of original images was calculated to find the matching noise level. The matching noise level was considered to be the one that yielded equal values in this correlation, and the correlation between the decoded features with original image stimuli and the noiseless features of the original images. This matching noise level was calculated for each ROI/DNN layer pair in each subject.

Feature gain

After determining the matching noise level for each ROI/DNN layer pair, noise was added to the true features of the stimulus images to generate noisy true features. The difference between a) the correlation between the noisy true features and noiseless original image features $r_{o_{noise}}$, and b) the correlation between noisy true features and noiseless stimulus image features $r_{s_{noise}}$, produced the noise baseline as follows

$$\Delta r_{noise} = r_{o_{noise}} - r_{s_{noise}}$$

This noise baseline represented the pure feedforward behavior towards the blurred images fitted to the noisy observations from the fMRI.

The predicted feature correlation difference was calculated between: a) the correlation between predicted features and original image features, and b) the correlation between predicted features and stimulus image features.

$$\Delta r_{decode} = r_o - r_s$$

The feature gain was the difference between the predicted correlation difference and the noise baseline.

$$Feature\ Gain = \Delta r_{decode} - \Delta r_{noise}$$

This represented the effect of top-down modulation in the features. A positive gain meant features were more biased towards the original image features.

Content specificity

To estimate the content specificity of the predicted features from the VC, their correlation with the original image features was compared to that with the other original image features. The correlation of predicted features for each stimulus was calculated

for each of the non-corresponding original images in the test dataset (n = 39), and the mean correlation was then calculated. The mean over all the stimulus images from the stimuli grouped by DNN layer with all blur levels pooled was calculated (Different image correlation), and compared with the mean of the correlations with the corresponding original images (Same image correlation). To compare this to the baseline correlation between different images, the mean of the correlation between each stimulus image vector and the feature vectors of other original images was calculated (True feature correlation).

Behavioral data extraction

The subject acquisition response was recorded manually from the voice recordings. The written record was then revised with each subject to ensure accuracy. The record was written as incorrect in the cases where the subject missed giving a voice response. In the cases when the subject missed giving a button response, the previous button response from the same sequence was used, except when the stimulus was the last (original image) or the first (most degraded image) in the sequence, when the response was set to certain and uncertain respectively. Correct responses were the ones identical to the response of the last stimulus in each sequence (original image).

Acknowledgments

The authors thank Tomoyasu Horikawa, Kei Majima, Mitsuaki Tsukamoto, and Shuntaro Aoki for help with data collection and analysis. We also thank the members of Kamitani Lab for their valuable comments on this manuscript. This research was supported by grants from JSPS KAKENHI Grant numbers JP15H05920, JP15H05710, and ImPACT Program of Council for Science, Technology and Innovation (Cabinet Office, Government of Japan), and the New Energy and Industrial Technology Development Organization (NEDO), and Japanese Government Scholarship (MEXT).

We thank Karl Embleton, PhD, from Edanz Group (www.edanzediting.com/ac) for editing a draft of this manuscript.

This study was conducted using the MRI scanner and related facilities of Kokoro Research Center, Kyoto University

Additional information

Funding (Funder; Grant reference number; Author)

JSPS KAKENHI; JP15H05920; Mohamed Abdelhack, Yukiyasu Kamitani

JSPS KAKENHI; JP15H05710; Mohamed Abdelhack, Yukiyasu Kamitani

ImPACT Program of Council for Science, Technology and Innovation; N/A; Yukiyasu Kamitani

The New Energy and Industrial Technology Development Organization (NEDO); N/A; Yukiyasu Kamitani

Japanese Ministry of Education, Culture, Sports, Science and Technology Scholarship (MEXT); N/A; Mohamed Abdelhack

The funders had no role in study design, data collection, and interpretation, or the

decision to submit the work for publication.

Author contributions

MA and YK designed the study. MA performed experiments and analyses. MA and YK wrote the paper.

Author ORCIDs

Mohamed Abdelhack, https://orcid.org/0000-0002-6753-3237

Yukiyasu Kamitani, http://orcid.org/0000-0002-9300-8268

Ethics

Human subjects: All subjects gave informed consent. The research protocol was approved by the Ethics Committee of ATR. IRB number 16-106, 17-106.

References

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences*, *8*(10), 457–464. doi:10.1016/j.tics.2004.08.011

Alink, A., Schwiedrzik, C. M., Kohler, A., Singer, W., & Muckli, L. (2010). Stimulus predictability reduces responses in primary visual cortex. *The Journal of Neuroscience*, *30*(8), 2960–2966. doi:10.1523/JNEUROSCI.3730-10.2010

Arnal, L. H., & Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences*, *16*(7), 390–398. doi:10.1016/j.tics.2012.05.003

Bansal, A., Sheikh, Y., & Ramanan, D. (2017). PixelNN: Example-based Image Synthesis. *arXiv Preprint arXiv:1708.05349*.

Bar, M, Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., … Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(2), 449–454. doi:10.1073/pnas.0507062103

Bar, Moshe, & Aminoff, E. (2003). Cortical analysis of visual context. *Neuron*, *38*(2), 347–358. doi:10.1016/S0896-6273(03)00167-3

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

Breitmeyer, B. G. (2014). Contributions of magno- and parvocellular channels to conscious and non-conscious vision. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *369*(1641), 20130213. doi:10.1098/rstb.2013.0213

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., … DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate

IT cortex for core visual object recognition. *PLoS Computational Biology*, *10*(12), e1003963. doi:10.1371/journal.pcbi.1003963

Chiou, R., & Lambon Ralph, M. A. (2016). The anterior temporal cortex is a primary semantic source of top-down influences on object recognition. *Cortex*, *79*, 75–86. doi:10.1016/j.cortex.2016.03.007

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. doi:10.1017/S0140525X12000477

de-Wit, L. H., Kubilius, J., Wagemans, J., & Op de Beeck, H. P. (2012). Bistable Gestalts reduce activity in the whole of V1, not just the retinotopically predicted parts. *Journal of Vision*, *12*(11). doi:10.1167/12.11.12

Den Ouden, H. E. M., Friston, K. J., Daw, N. D., McIntosh, A. R., & Stephan, K. E. (2009). A dual role for prediction error in associative learning. *Cerebral Cortex*, *19*(5), 1175–1185. doi:10.1093/cercor/bhn161

Den Ouden, H. E. M., Kok, P., & de Lange, F. P. (2012). How prediction errors shape perception, attention, and motivation. *Frontiers in Psychology*, *3*, 548. doi:10.3389/fpsyg.2012.00548

Engel, S. A., Rumelhart, D. E., Wandell, B. A., Lee, A. T., Glover, G. H., Chichilnisky, E. J., & Shadlen, M. N. (1994). fMRI of human visual cortex. *Nature*, *369*(6481), 525. doi:10.1038/369525a0

Fellbaum, C. (2012). WordNet. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/9781405198431.wbeal1285

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the*

Royal Society of London. Series B, Biological Sciences, *360*(1456), 815–836. doi:10.1098/rstb.2005.1622

Friston, K., Brown, H. R., Siemerkus, J., & Stephan, K. E. (2016). The dysconnection hypothesis (2016). *Schizophrenia Research*, *176*(2-3), 83–94. doi:10.1016/j.schres.2016.07.014

Gayet, S., Guggenmos, M., Christophel, T. B., Haynes, J.-D., Paffen, C. L. E., Van der Stigchel, S., & Sterzer, P. (2017). Visual working memory enhances the neural response to matching visual input. *The Journal of Neuroscience*, *37*(28), 6638–6647. doi:10.1523/JNEUROSCI.3418-16.2017

Gordon, N., Koenig-Robert, R., Tsuchiya, N., van Boxtel, J. J., & Hohwy, J. (2017). Neural markers of predictive coding under perceptual uncertainty revealed with Hierarchical Frequency Tagging. *eLife*, *6*. doi:10.7554/eLife.22749

Güçlü, U., & van Gerven, M. A. J. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *The Journal of Neuroscience*, *35*(27), 10005–10014. doi:10.1523/JNEUROSCI.5023-14.2015

Heeger, D. J. (2017). Theory of cortical function. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(8), 1773–1782. doi:10.1073/pnas.1619788114

Hochstein, S., & Ahissar, M. (2002). View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron*, *36*(5), 791–804.

Horikawa, T., & Kamitani, Y. (2017a). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, *8*, 15037. doi:10.1038/ncomms15037

Horikawa, T., & Kamitani, Y. (2017b). Hierarchical Neural Representation of Dreamed Objects Revealed by Brain Decoding with Deep Neural Network Features. *Frontiers in Computational Neuroscience*, *11*, 4. doi:10.3389/fncom.2017.00004

Hsieh, P. J., Vul, E., & Kanwisher, N. (2010). Recognition alters the spatial pattern of FMRI activation in early retinotopic cortex. *Journal of Neurophysiology*, *103*(3), 1501–1507. doi:10.1152/jn.00812.2009

Jardri, R., Hugdahl, K., Hughes, M., Brunelin, J., Waters, F., Alderson-Day, B., … Denève, S. (2016). Are Hallucinations Due to an Imbalance Between Excitatory and Inhibitory Influences on the Brain? *Schizophrenia Bulletin*, *42*(5), 1124–1134. doi:10.1093/schbul/sbw075

Jia Deng, Wei Dong, Socher, R., Li-Jia Li, Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). IEEE. doi:10.1109/CVPR.2009.5206848

Kastner, S., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1998). Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI. *Science*, *282*(5386), 108–111. doi:10.1126/science.282.5386.108

Kastner, S., Pinsk, M. A., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron*, *22*(4), 751–761. doi:10.1016/S0896-6273(00)80734-5

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS*

Computational Biology, *10*(11), e1003915. doi:10.1371/journal.pcbi.1003915

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in psychtoolbox-3. *Perception*, *36*(14).

Kok, P., Jehee, J. F. M., & de Lange, F. P. (2012). Less is more: expectation sharpens representations in the primary visual cortex. *Neuron*, *75*(2), 265–270. doi:10.1016/j.neuron.2012.04.034

Kok, P., Mostert, P., & De Lange, F. P. (2017). Prior expectations induce pre-stimulus sensory templates. *BioRxiv*. doi:10.1101/119073

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. doi:10.1145/3065386

Kveraga, K., Boshyan, J., & Bar, M. (2007). Magnocellular projections as the trigger of top-down facilitation in recognition. *The Journal of Neuroscience*, *27*(48), 13232–13240. doi:10.1523/JNEUROSCI.3481-07.2007

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, *20*(7), 1434–1448. doi:10.1364/JOSAA.20.001434

Meyer, T., & Olson, C. R. (2011). Statistical learning of visual transitions in monkey inferotemporal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(48), 19401–19406. doi:10.1073/pnas.1112895108

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, *229*(4715), 782–784. doi:10.1126/science.4023713

Morgan, A. T., Petro, L. S., & Muckli, L. (2016). Cortical feedback to V1 and V2 contains unique information about high-level scene structure. *BioRxiv*. doi:10.1101/041186

Mumford, D. (1992). On the computational architecture of the neocortex. *Biological Cybernetics*, *66*(3), 241–251. doi:10.1007/BF00198477

Murray, S. O., Kersten, D., Olshausen, B. A., Schrater, P., & Woods, D. L. (2002). Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(23), 15164–15169. doi:10.1073/pnas.192579399

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. *Progress in Brain Research*, *155*, 23–36. doi:10.1016/S0079-6123(06)55002-2

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. doi:10.1038/4580

Revina, Y., Petro, L. S., & Muckli, L. (2017). Cortical feedback signals generalise across different spatial frequencies of feedforward inputs. *Neuroimage*. doi:10.1016/j.neuroimage.2017.09.047

Sereno, M. I., Dale, A. M., Reppas, J. B., Kwong, K. K., Belliveau, J. W., Brady, T. J., … Tootell, R. B. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, *268*(5212), 889–893. doi:10.1126/science.7754376

Shi, Sun, Y. Q., & Huifang. (2008). *Image and Video Compression for Multimedia Engineering: Fundamentals, Algorithms, and Standards*. CRC Press, Inc.

Summerfield, C., & de Lange, F. P. (2014). Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews. Neuroscience*, *15*(11), 745–756. doi:10.1038/nrn3838

Todorovic, A., van Ede, F., Maris, E., & de Lange, F. P. (2011). Prior expectation mediates neural adaptation to repeated sounds in the auditory cortex: an MEG study. *The Journal of Neuroscience*, *31*(25), 9118–9123. doi:10.1523/JNEUROSCI.1425-11.2011

Vedaldi, A., & Lenc, K. (2015). Matconvnet: convolutional neural networks for MATLAB. In *Proceedings of the 23rd ACM international conference on Multimedia - MM' ' '15* (pp. 689–692). New York, New York, USA: ACM Press. doi:10.1145/2733373.2807412

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8619–8624. doi:10.1073/pnas.1403112111

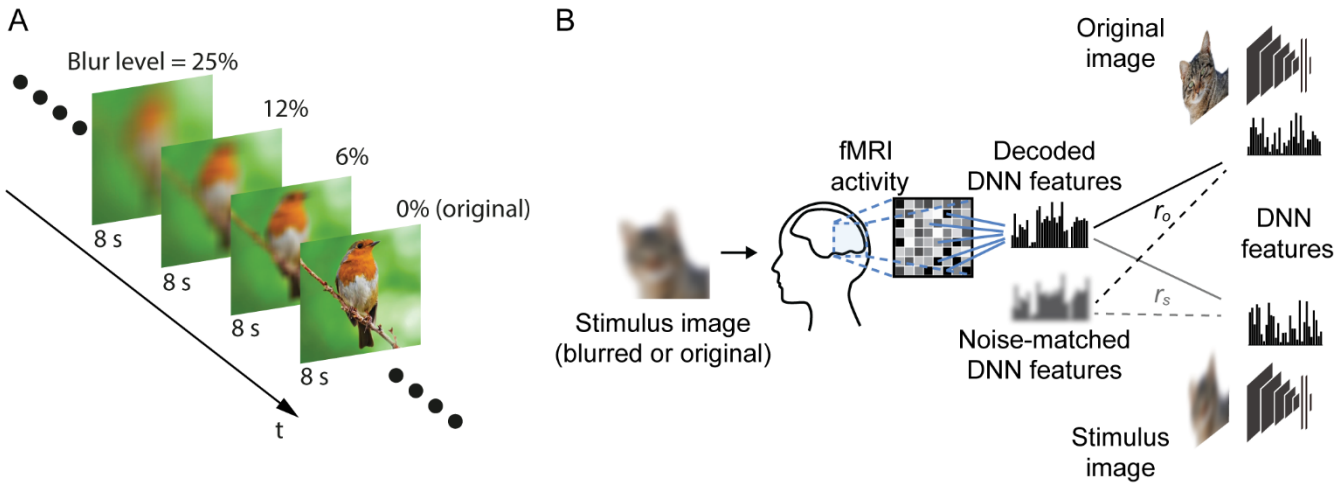

Figure 1 | Study design: (A) The stimulus sequence was divided into sequences of four stimuli each. Stimuli in the same sequence contained different blur levels of the same image organized from the highest blur level (25%) to the lowest (0%). Each stimulus was presented for 8 seconds. (B) Overview of the feature decoding analysis protocol; fMRI activity was measured as the subjects viewed the stimulus images presented, described in A. Trained decoders were used to predict DNN features from fMRI activity patterns. The decoded features were then analyzed for their similarity with the true DNN features of both the original image (r_o) and stimulus image (r_s). The same procedure was also conducted for noise matched DNN features which are composed of true DNN features with additional Gaussian noise to match predicted features from fMRI.

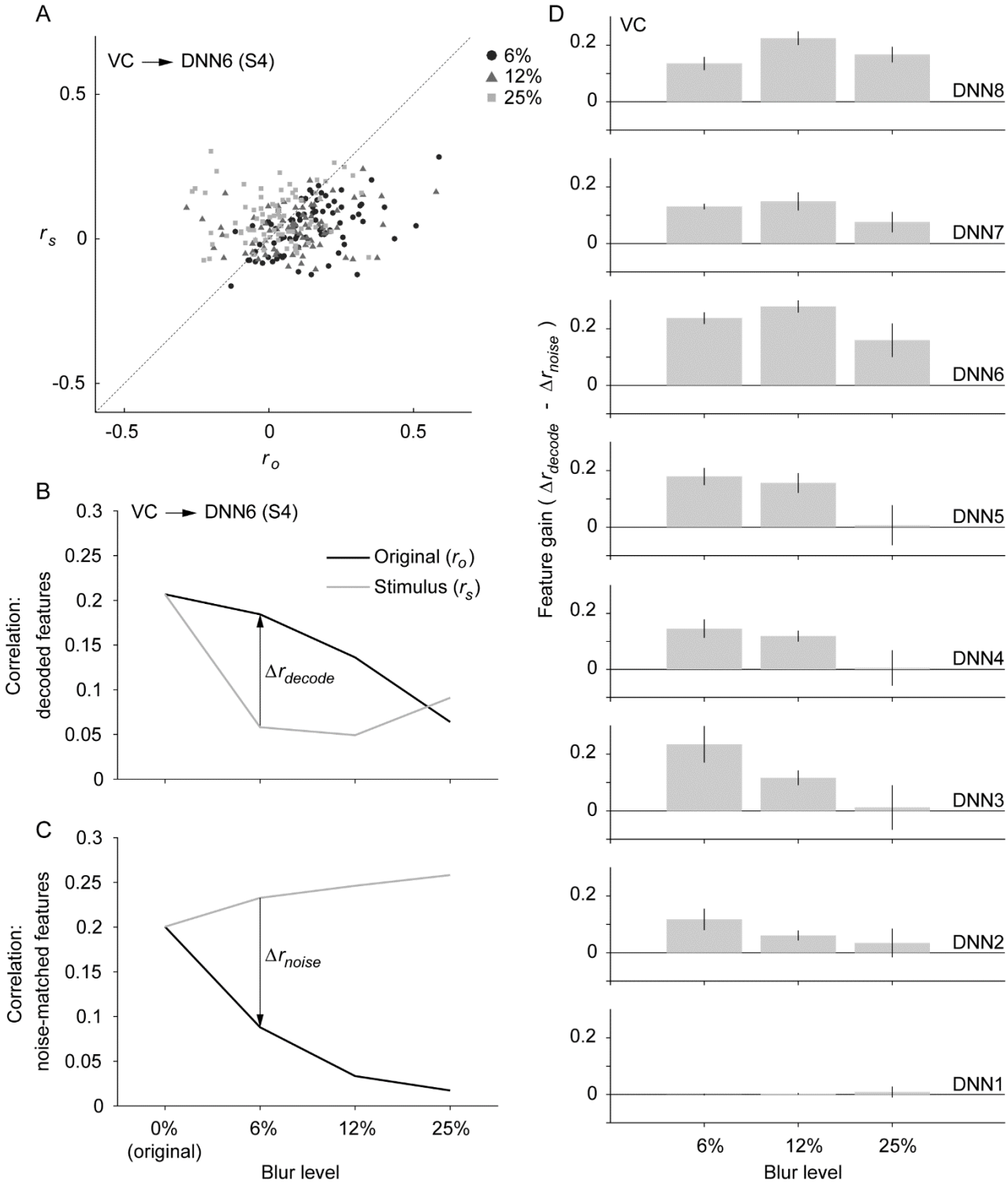

Figure 2 | Correlation of decoded features with original and stimulus image features: (A) Scatter plot showing feature correlation of DNN6 features decoded from

the whole visual cortex (VC) of subject 4, with original image features (r_o; x-axis) and stimulus image features (r_s; y-axis). Each point represents a stimulus image for all blurring levels except 0%. Diagonal dotted line represents the line of equal correlation ($\Delta r_{decode} = 0$). (B) Representative result from DNN6 features decoded from the whole visual cortex (VC) of subject 4. Solid lines represent the mean correlation at different blur levels while pooling different experimental conditions and behavioral response data. The difference between r_o and r_s is labelled as Δr_{decode}. (C) Representative result showing mean noise-matched feature correlation with the original and stimulus image features for different blur levels. Noise-matching was performed to match the correlation of the DNN6 predicted features of the 0% blur stimuli decoded from VC of subject 4 (thus obtaining equal values with the decoded features at the 0% level). The difference between r_o and r_s yields the noise baseline (Δr_{noise}). (D) Feature gain is defined as the difference between Δr_{decode} and Δr_{noise}. Mean feature gain is indicated for each DNN layer for features decoded from VC at different stimulus blur levels (excluding the 0% level). Error bars indicate 95% confidence interval (CI) across five subjects.

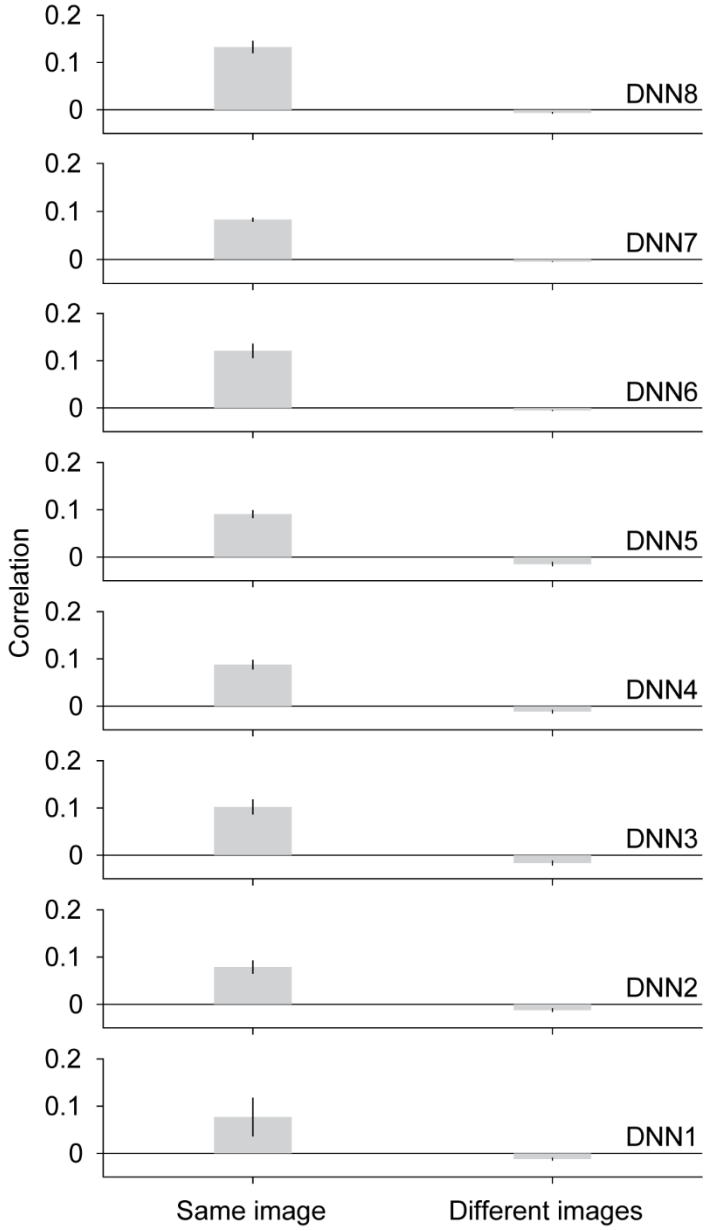

Figure 3 | Content specificity of decoded features with blurred images: Same image correlation indicates correlation of predicted features (blur levels pooled, excluding 0%) with corresponding original image features. Different images correlation indicates the mean of correlations of the same predicted features with original image features of different images. The mean correlation is shown for different DNN layers. Error bars indicate 95% CI across five subjects.

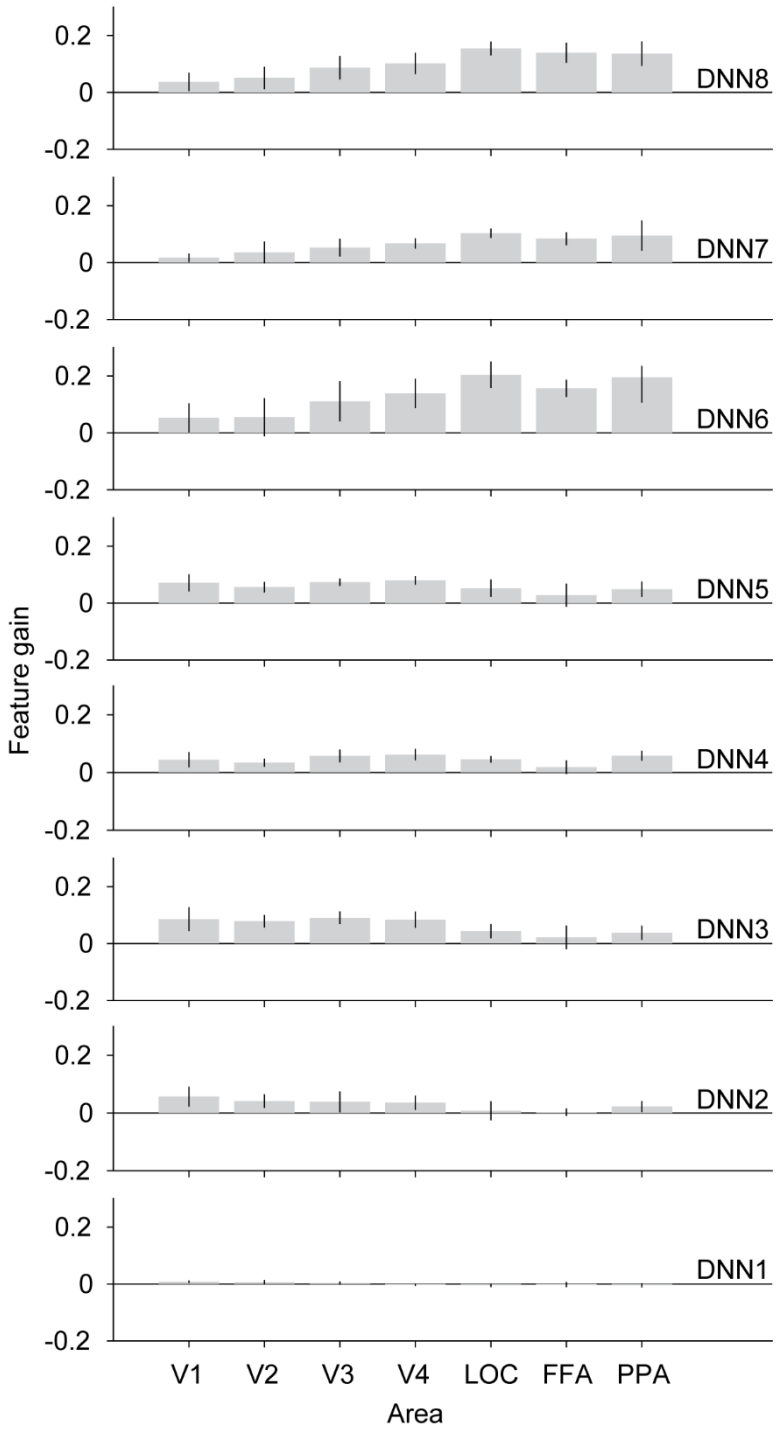

Figure 4 | Feature gain across visual areas: Feature gain for features predicted from different visual areas. Mean feature gain is indicated for each DNN layer (blur levels pooled, 0% excluded). Error bars indicate 95% CI across five subjects.

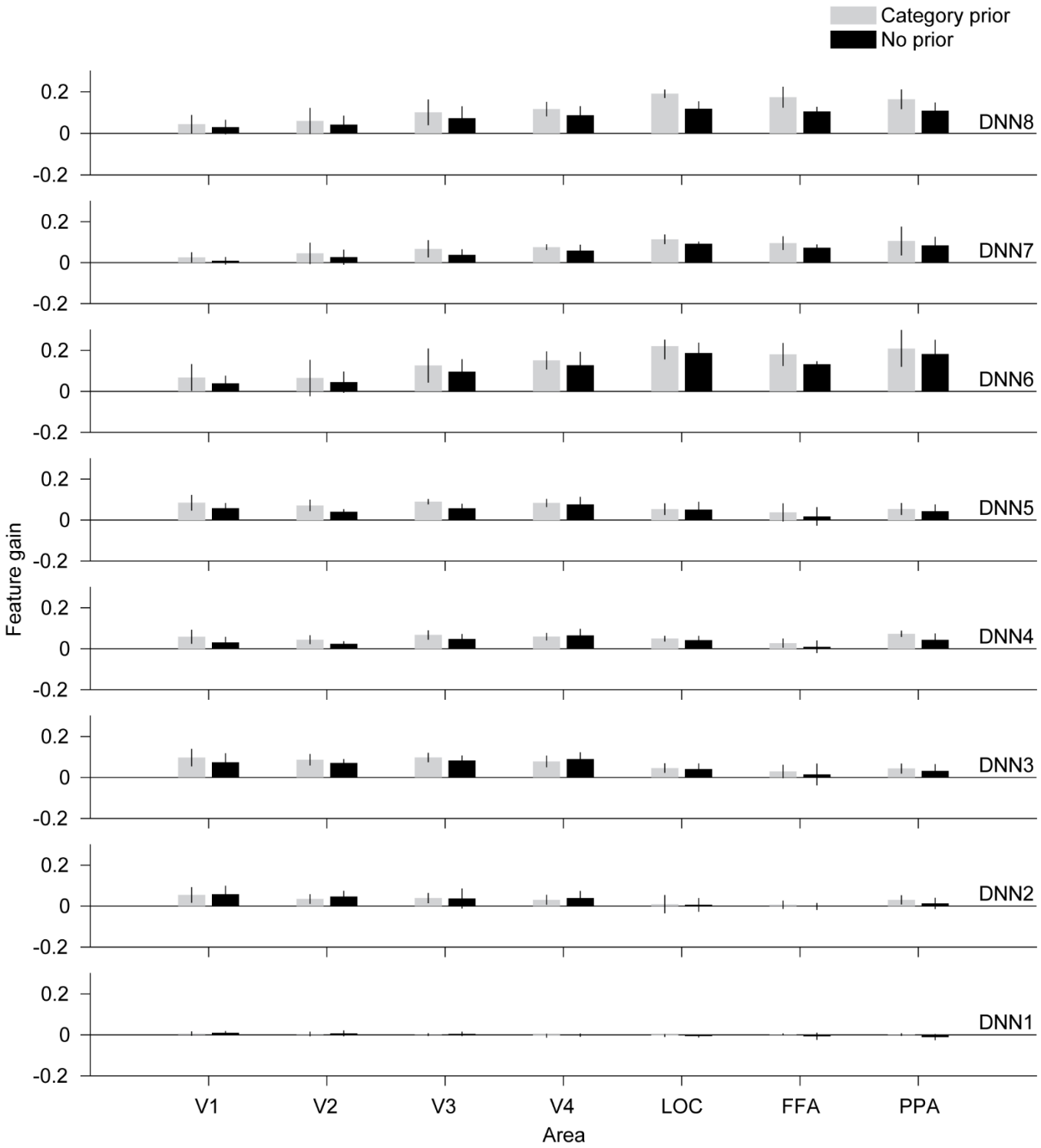

Figure 5 | Effect of category prior: Feature gain for features predicted from different visual areas grouped by experimental condition (category-prior vs. no-prior). Mean feature gain is indicated for each DNN layer (blur levels pooled, 0% excluded). Error bars indicate 95% CI across five subjects.

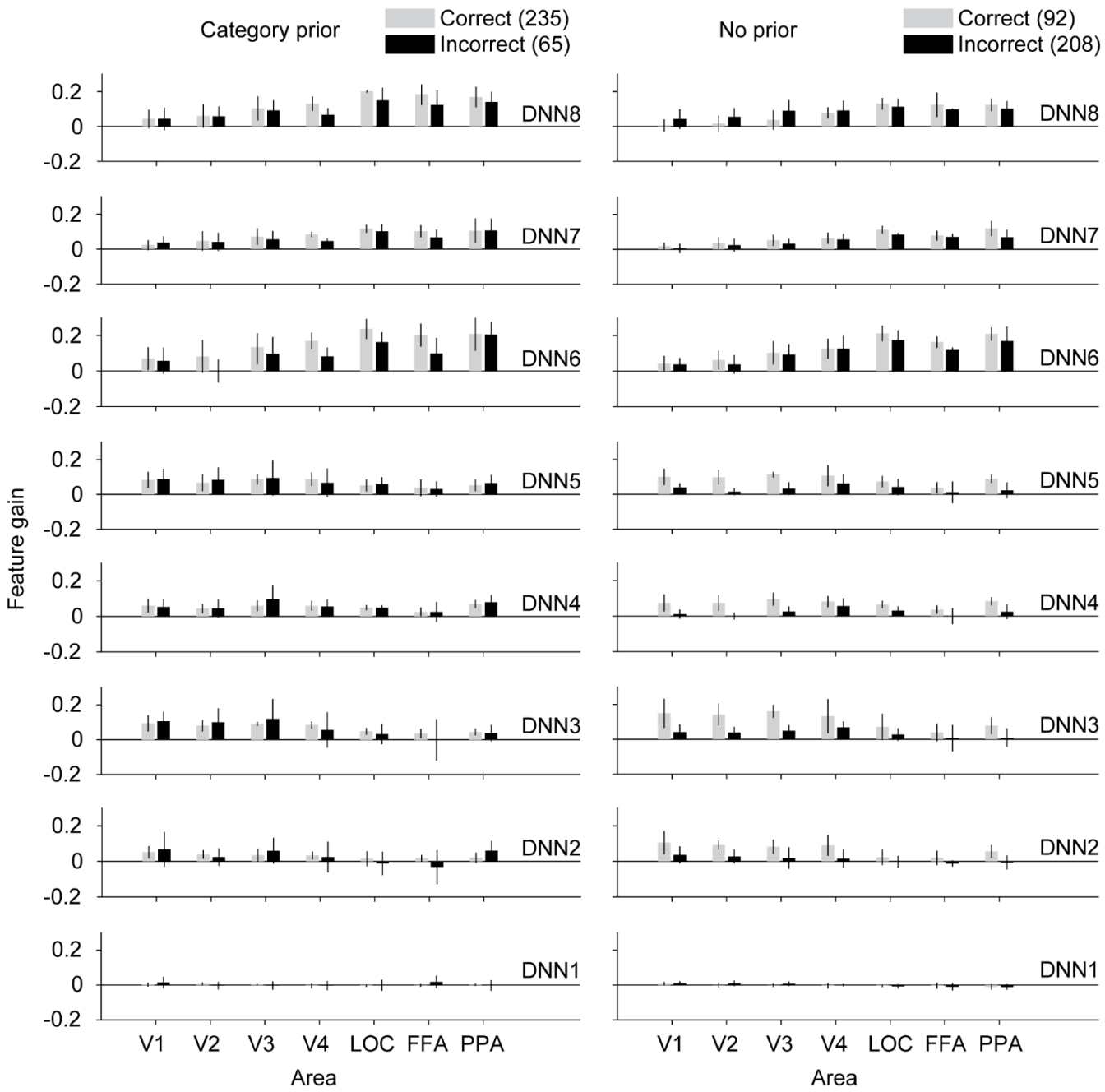

Figure 6 | Effect of behavioral performance: Feature gain for features predicted from different visual areas grouped by experimental condition (category-prior vs. no-prior) and recognition (correct vs. incorrect). Legends include the total number of occurrences of each response across subjects. Mean feature gain is indicated for each DNN layer (blur levels pooled, 0% excluded). Error bars indicate 95% CI across five subjects.

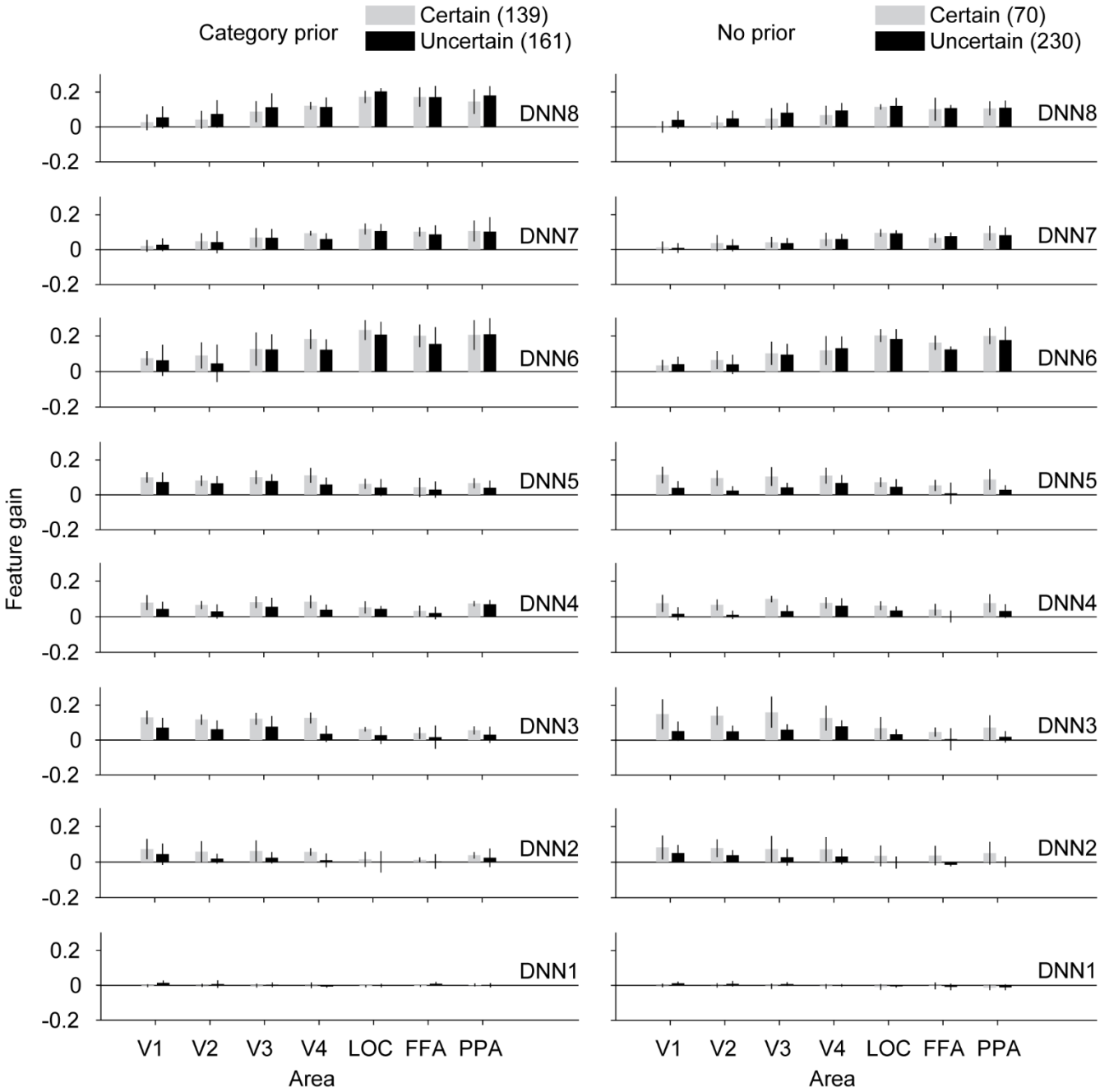

Figure 7 | Effect of confidence level: Feature gain for features predicted from different visual areas grouped by experimental condition (category-prior vs. no-prior) and confidence level (certain vs. uncertain). Legends include the total number of occurrences of each response across subjects. Mean feature gain is indicated for each DNN layer (blur levels pooled, 0% excluded). Error bars indicate 95% CI across five subjects.