

1 TITLE

2 Posterior inferotemporal cortex cells use multiple visual pathways to complement fine and
3 coarse discriminations

4 **AUTHORS:** Ponce C.R. *, Lomber S.G. and Livingstone M.S.

5 * Corresponding author

6 ABSTRACT

7 In the macaque monkey brain, posterior inferior temporal cortex (PIT) cells are responsible
8 for visual object recognition. They receive concurrent inputs from visual areas V4, V3 and V2. We
9 asked how these different anatomical pathways contribute to PIT response properties by
10 deactivating them while monitoring PIT activity. Using cortical cooling of areas V2/V3 or V4 and
11 a hierarchical model of visual recognition, we conclude that these distinct pathways do not
12 transmit different classes of visual features, but serve instead to maintain a balance of local- and
13 global-feature selectivity in IT.

14 INTRODUCTION

15 Posterior IT (PIT) neurons are the penultimate stage of the ventral visual processing
16 stream, comprising cortical areas V1→V2→V4→PIT→anterior IT (AIT). In addition to this main
17 pathway, PIT also receives direct feedforward projections from V3 and V2¹, and V4 receives direct
18 inputs from V1. These short projections (V1→V4→PIT and V1→V2→PIT) have been called
19 bypass pathways², and they represent a significant fraction of the inputs to PIT: 14% of all neurons
20 in the brain projecting to PIT are located in areas V2|3 (for context, 26% of inputs to PIT arrive
21 from V4; only 1% of inputs to V1 come from the LGN)^{3,4}. The remaining projections arise from AIT
22 and the dorsal pathway. The goal of this study is to define the role of these different input
23 pathways in PIT selectivity.

24 We recorded from an unbiased sample of PIT neurons while deactivating, by cooling,
25 areas V2-V3 (together) and V4 (**Fig. 1a**). We measured PIT firing rates before and during cooling
26 of V4 or V2|3, and quantified changes in the representational capacity of PIT. Using firing rate
27 statistics and linear classifiers (i.e. support vector machines), we found that while V4-dependent
28 inputs were more important for preserving the representation of the identity of images in PIT, the
29 different concurrent pathways did not transmit different types of visual features (such as different
30 proportions of curvature or spatial frequency). We modeled the contributions of short- and long-
31 pathways using the standard model of visual recognition⁵, and observed that short pathways were
32 well-positioned for fine feature discriminations. We confirmed that fine-feature discrimination was
33 relatively better preserved during cooling compared to coarse-feature discrimination. We
34 simulated the effects of cooling on decoding accuracy using various random cooling effects
35 models, and found that while these random models predicted an overall loss of accuracy, they
36 did not predict the preservation of decoding accuracy for fine discriminations. Thus we conclude
37 that short pathways are helpful in fine discrimination, because their receptive field weights match
38 simpler image elements. By introducing units with simpler preferences into PIT, the short
39 pathways create a diversity of feature preferences available for downstream perceptual
40 operations.

41 RESULTS

42 Cooling affected portions of PIT response fields

43 We implanted floating microelectrode arrays in the PIT of two adult male monkeys (2
 44 arrays in monkey R and 3 arrays in monkey G), along with cryoloops on dorsal V2, V3 and V4^{6,7}.
 45 The arrays were placed anteriorly to the inferior occipital sulcus, the cryoloops were placed within
 46 the lunate sulcus and over the predorsal gyrus (**Fig. 1b**). We activated the cryoloops
 47 intraoperatively, using thermal imaging to plot the extent of cooling and found that the lower
 48 thermal region was limited to 1-1.5 mm around and within the cryoloop (**Fig. 1c**). The flat area of
 49 cortex directly cooled by the cryoloops was around 13 x 5 mm². The electrode arrays were at least
 50 5 mm anterior to the prelunate cryoloop, and anterior to the inferior occipital sulcus. Three weeks
 51 after each surgery, we measured the retinotopic response fields of the arrays and the retinotopic
 52 size of the cooling scotomas: the animals maintained their gaze on a 0.4°-diameter central black
 53 circle, while a 2°-diameter image was flashed randomly within a 16 x 16° radial grid. We collected
 54 spike data before, during and after activation of the V2/V3 or V4 cryoloops, counterbalancing the
 55 order of the V4 vs. V2|3 deactivations. The mean PIT population response fields were centered
 56 on the perifoveal upper contralateral hemifield (**Fig. 1d**). In contrast, the V4 and V2|3 scotomas
 57 were centered towards the perifoveal lower hemifield, as predicted by the dorsal retinotopic
 58 representations of V2, V3 and V4. In subsequent experiments, stimuli were sized to fit within the
 59 overlapping region of both scotomas (**Fig 1e**) (1.4°-wide images for monkey G, 2.0°-wide images
 60 for monkey R).

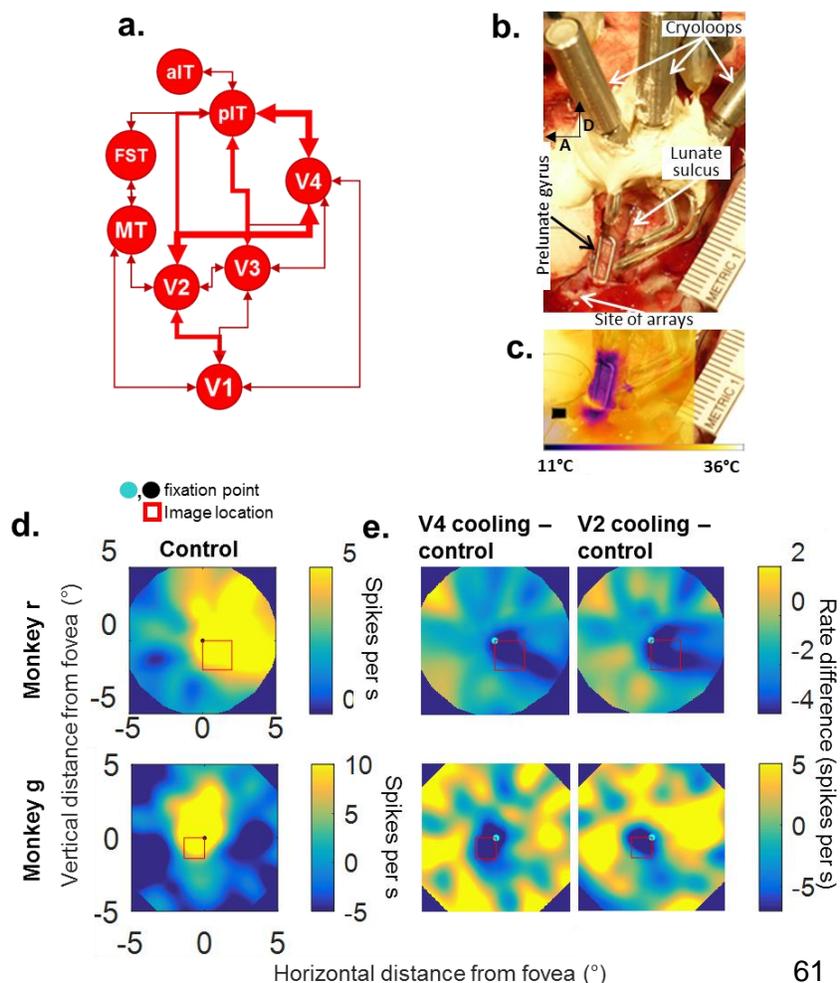


Figure 1. Cooling affected portions of the aggregate PIT response fields.

a. Input network to posterior IT (PIT).

b. Intraoperative picture, monkey R left hemisphere. Posterior craniotomy showing the location of the lunate sulcus, prelunate gyrus and V2/V4 cryoloops. The arrays were later implanted where shown by the label. A: anterior, D: dorsal.

c. Composite image showing the superimposed thermal and visible light images, while the prelunate gyrus loop is active. The black square shows the approximate location of the arrays.

d. Average firing rate for all units, evoked by flashing an image in a 8 x 8° grid, during the control (warm) condition.

e. Difference in activity during cooling of area V4 (left column) or V2|3 (right column) – the control map was subtracted from the cooling map. Dark regions show reductions in firing rate.

62

63 Cooling reduced firing rates in PIT units and decoding accuracy by linear classifiers

64 In all the following experiments, we showed the fixating animals 293 images belonging to
65 15 different categories (angles, animals, artificial objects, curves, faces, radial and linear gabors,
66 joint angles, plants, places, noise textures and tristar; the entire image set is shown in
67 **Supplementary Fig. 1**). When we cooled either set of coils, PIT multiunits reduced their visually
68 evoked spike rates (**Fig 2a**). During control conditions, PIT multiunits showed a median visual
69 response of 18 ± 1 (monkey R) and 22 ± 2 (Monkey G) spikes/s (range of -3 to 120 spikes per
70 second relative to baseline for monkey R; -12 to 106 spikes per s for monkey G). When the V2/V3
71 loops were cooled, the overall average rate was reduced to 13 ± 1 and 14 ± 1 spikes/s (monkeys R,
72 G). When the V4 cryoloops were cooled, the overall rate was reduced to 12 ± 1 and 15 ± 1 spikes/s
73 (monkeys R, G; probability that all means come from same distribution $P = 6 \times 10^{-4}$ and 1×10^{-4} ,
74 one-way ANOVA, $N_{\text{sites}} = 300$ and 256 , $F(2, 897) = 8.1$, $F(2, 765) = 8.9$). In one animal, we cooled
75 both V4 and V2|3 loops concurrently, measuring a similar reduction in firing rate (38%); cooling
76 both sets of loops did not silence PIT activity. Another measure of input strength is response
77 latency, and here we similarly observed little difference between V2/V3 cooling and V4 cooling
78 (see **Supplementary Section 1**).

79 Next we used pattern analysis to quantify the encoding capacity of PIT during V4 or V2|3
80 cooling. We trained statistical classifiers (support vector machines with a linear kernel, or SVMs)
81 using data from each experimental condition (before cooling, during V4 or V2|3 cooling). SVMs
82 were used in an all-vs.-all approach, so chance performance was 50% per comparison. We had
83 few trials per image during each cooling condition (4-6) and so we used leave-one-out cross-
84 validation for each paired comparison. To further guard against unreliable values due to the small
85 sample number, we also trained SVMs using the same data but shuffling the image labels. Thus,
86 accuracy was defined as the mean of all cross-validation cycles using the correct labels minus
87 the mean of cross-validation cycles using the shuffled labels, so a range of 0-0.5 is equivalent to
88 50-100% accuracy.

89 First, this decoding analysis showed that faces elicited the highest classification accuracy
90 in both animals, which was notable because we did not pre-select the array implantation sites by
91 proximity to fMRI-defined face patches. SVMs showed a median accuracy value of 0.25 ± 0.01 and
92 0.26 ± 0.01 above baseline (monkeys R, G, standard error of the median). During V4 deactivation,
93 median accuracy dropped to 0.16 ± 0.01 and 0.17 ± 0.01 ; during V2|3 deactivation, median
94 accuracy dropped to 0.19 ± 0.01 and 0.20 ± 0.01 . These median accuracy values were statistically
95 different at the group level (monkeys R, G: $P = 5 \times 10^{-12}$, 3×10^{-23} , one-way Kruskal-Wallis, $N_{\text{images}} =$
96 $347, 293$, $\chi^2(2, 1038) = 52$, $\chi^2(2, 876) = 104$). The differences in median values between V4 and V2|3
97 cooling were statistically reliable in both animals (monkeys R and G: $P = 2 \times 10^{-5}$, 4×10^{-3} , two-tailed
98 Wilcoxon sign rank test, $N = 347, 293$ accuracy values, Z-stats: -4.3, -2.9, **Fig. 2b**, bottom, and
99 **Fig. 2c**). We also performed an SVM category classification analysis, grouping responses by
100 category, not as individual images. We also found that deactivating V4-inputs reduced category
101 classification accuracy more than did deactivating V2|3 (**Supplementary Section 2, Fig. S2 and**
102 **S3**).

103 In summary, PIT multiunits showed similar mean firing rate reductions during V2|3 or V4
104 deactivation, but in both monkeys, SVM accuracy was reduced more by V4 deactivation. This
105 suggests that V4 direct inputs are more important for image identification and categorization, and
106 that this cannot be explained by a simple reduction in mean spike rate. We further explored the

107 reasons why decoding accuracy shows a quantitatively stronger role for V4 using a
 108 multidimensional projection analysis, described in **Supplementary Section 3**.

109

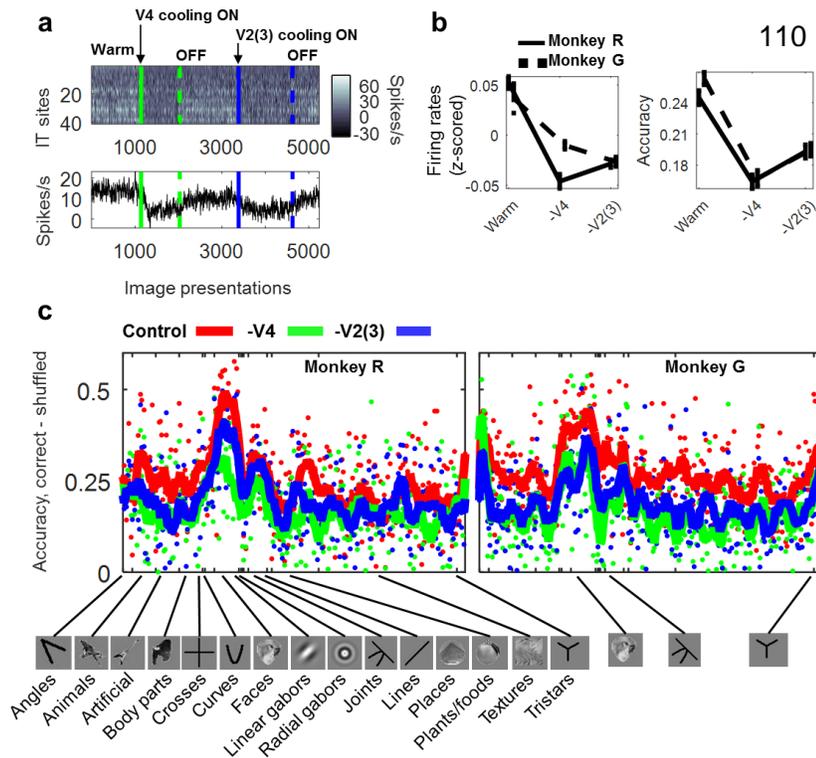


Figure 2. Effects of cooling on firing rate and classification accuracy.

a. (top) Data from one cooling session (monkey R, day 1). The top figure shows the evoked spike rates from 50 PIT sites (rows) recorded concurrently before, during and after cooling of V4 and V2/3. Each column represents one image presentation. Solid lines mark the onset of each cooling condition, broken lines show the onset of the rewarming periods. (bottom) Mean firing rate across each temperature condition.

b. (top) Average firing rate activity (z-scored) for all channels during each temperature condition. (bottom) Average classification accuracy for all images during each temperature condition.

c. Mean cross-validated accuracy for each image before and during cooling (red = control, green = V4 cooling, blue = V2/3 cooling). The x-axis shows all 293 images listed alphabetically by their category.

111 Cooling effects on median accuracy were simulated by random processes

112 To gain a better understanding of the drop in decoding performance during input cooling,
 113 we asked how much of the accuracy impairment could be attributed to non-specific fractional
 114 reductions in firing rate. We modeled three mechanisms of cooling rate reductions: 1) each IT
 115 multiunit underwent a given fractional reduction across all of its responses (a *site-by-site* cooling
 116 mechanism); 2) all multiunits underwent the same fractional reduction, but the reduction could
 117 vary over time (a *temporal* mechanism); finally, 3) each multiunit underwent a mix of *site-by-site*
 118 and *temporally dependent* reductions. We first examined the effects of each cooling mechanism
 119 using a model. In neural activity space, where images are represented as coordinate positions in
 120 a multi-dimensional space (**Fig. 3a**), these cooling effects would induce different geometric
 121 transformations and thus affect decoding accuracy differently. Consider the first mechanism of
 122 site-by-site cooling: this has no effect on classification accuracy, because each site's responses
 123 are normalized before classification; it makes no difference if the lengths of the population
 124 response vectors change, as long as they keep the same direction (**Fig. 3b-c, i**). This is a
 125 plausible compensation because normalization is a common mechanism in cortical computations.
 126 The second mechanism, temporally dependent reductions, where cooling imposes a different
 127 fractional value on the whole population vector at different times, pulls and stretches the pre-
 128 cooling response vector groups towards a minimum (**Fig. 3b-c, ii**). The third mechanism, where
 129 individual multi-unit sites undergo different fractional reductions over time, is interesting because

130 it changes the direction of the vectors, directly affecting the image representations in activity
 131 space (**Fig. 3b-c, iii**).

132 We applied each mechanism to our control (warm) data to simulate the cooling drop in
 133 decoding performance. First we measured the distribution of fractional cooling changes during
 134 cooling. In monkey R, the median fractional reduction during V4 cooling was 0.68 (25th, 75th
 135 quantiles: 0.51-0.80), and during V2|3 cooling, 0.73 (25th, 75th quantiles 0.59-0.84). For monkey
 136 G, the median fractional reduction during V4 cooling was 0.67 (25th, 75th quantiles 0.54-0.82),
 137 and during V2|3 cooling 0.62 (25th, 75th quantiles 0.48-0.74). Note that some of these fractional
 138 changes included increased firing rates during cooling, but this was expected from weakly firing
 139 multiunits. Next, we sampled these fractional distributions (with replacement) and multiplied each
 140 sampled fraction times the control (warm) response counts. These multiplications were done
 141 using either the site-by-site cooling mechanism, the temporally dependent cooling mechanism, or
 142 the mixed mechanism. As a form of cross-validation, the fractional gain distributions came from
 143 different days from the control data. We created 20 “cooling” populations, based on V4 and V2|3
 144 cooling, per monkey.

145 Before cooling, the mean decoding accuracy for both animals was $26 \pm 1\%$ over baseline,
 146 $16 \pm 1\%$ during V4 cooling and $19 \pm 1\%$ during V2|3 cooling. We found that the site-by-site spike
 147 mechanism did not reduce decoding accuracy (its value remained $26 \pm 1\%$, same as control). The
 148 temporally dependent mechanism reduced decoding accuracy to $15 \pm 1\%$ over baseline. The
 149 mixed mechanism lowered accuracy as a function of the number of responses that were randomly
 150 affected: if we multiplied 100% of all responses by the random fractions, decoding accuracy was
 151 reduced to $16 \pm 1\%$. To match the experimental reductions in accuracy, we had to affect between
 152 40-80% of all responses, which resulted in 18-22% decoding accuracy, (**Fig. 3d**). In summary, we
 153 could not account for the observed reduction in decoding accuracy by a uniform fractional in
 154 spikes within each multiunit site, but a temporally varying reduction or a mix of site-by-site and
 155 temporally varying reductions could account for the mean decoding accuracy loss during cooling.

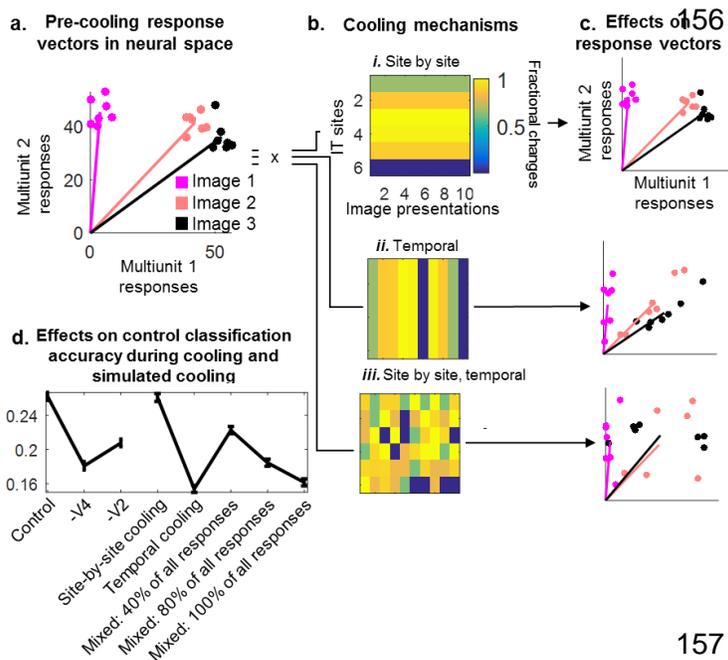


Figure 3. Cooling simulations.

a. Hypothetical responses of two units (axes) to three different images (colors), each presented seven times each (points). **b.** Different mechanisms of cooling. Cooling may impose a fixed fractional change in a channel-by-channel basis, a temporal basis, or a combination of the two. **c.** Responses in (A) transformed by each “cooling” reduction rate mechanism. **d.** Mean accuracy values (baseline-subtracted, \pm standard error) measured before cooling (“control,” “-V4,” “-V2|3”, both monkeys combined), and after different cooling simulations (“site-by-site”, “temporal”, and “mixed”). For the “mixed” simulations, each percentage shows the fraction of responses that were randomly multiplied by the gain values.

157

158

Cooling reduced selectivity of individual PIT multiunits

159 Cooling inputs to PIT reduced classification accuracy at the population level. To examine
160 accuracy at the level of individual sites, we measured selectivity using an F-test. Let us say that
161 PIT units were selective to specific images if the mean variance of their spike counts to different
162 images was greater than the mean variance of their spike counts to each image (F-statistic); the
163 value 1 suggests no selectivity; the greater the value, the more selective. We called the F-test
164 statistic per channel before cooling $F_{control}$, during V4 cooling F_{-V4} and during V2|3 cooling, $F_{-V2|3}$.
165 If the distributions of F_{-V4} and $F_{-V2|3}$ are closer to the non-selectivity value of 1 compared to the
166 $F_{control}$ distribution, this would suggest that PIT sites became less selective during cooling.

167 Each PIT site gave one $F_{control}$, one F_{-V4} and one $F_{-V2|3}$ value. We plotted each control F
168 value against its counterparts and found that cooling F-stats were mostly lower than the warm
169 distribution (**Fig. 4c**, Monkey R, median $F \pm$ standard error, Warm: 1.13 ± 0.03 , -V4: 1.10 ± 0.01 , -
170 V2|3: 1.14 ± 0.02 ; Monkey G, Warm: 1.21 ± 0.05 , -V4: 1.13 ± 0.02 , -V2|3: 1.14 ± 0.02). However, there
171 was no statistical difference among these values ($P = 0.20, 0.20$, Kruskal-Wallis, $N = 300, 256$
172 sites, $\chi^2(2,897)=3, \chi^2(2,765)=3$). The reason there was no statistical difference is that many of the
173 PIT sites were not that selective to start with, having pre-cooling F stats already bottomed out at
174 1. Still, units with higher pre-cooling F values showed greater changes during cooling. To quantify
175 this observation, we asked if the slope describing the relationship between the pre-cooling and
176 cooling F values was statistically different from unity. We used a bootstrap approach. For 1,000
177 iterations, we re-sampled sites with replacement and used their $F_{control}$, F_{-V4} and $F_{-V2|3}$ values to fit
178 linear regression lines between control and -V4 values, and between control and -V2|3 values.
179 This analysis resulted in 1,000 slopes describing the $F_{control}$ and F_{-V4} relationship, and another
180 1,000 slopes describing the $F_{control}$ and $F_{-V2|3}$ relationship. None of these slope values overlapped
181 the line of unity (Monkey R, mean slope \pm SEM, control vs. -V4: 0.64 ± 0.03 , control vs. -V2|3:
182 0.82 ± 0.04 ; Monkey G, control vs. -V4: 0.53 ± 0.04 , control vs. -V2|3: 0.59 ± 0.03). We also noticed
183 that the mean $F_{control} / F_{cooling}$ slope was shallower during V4 cooling than V2|3 cooling in both
184 animals. This implied that PIT multiunits became less selective during V4 cooling than during V2|3
185 cooling. This difference in slope was significant using a randomization test, which shuffled $F_{cooling}$
186 values from the V4 and V2|3 conditions and asked if this mixed distribution could produce the
187 observed slope (one-tailed randomization test, $P = 0.001$ and 0.02 ; see Methods for details).

188 We further asked whether there was any relationship between the retinotopic location of
189 a PIT response field relative to the cooling scotomas, and its subsequent selectivity (F-statistic)
190 change. The images were presented at the intersection of the V4 and V2|3 scotomas. Therefore,
191 some individual multiunit PIT response fields (RFs) would cover more of the stimulus than others.
192 For each PIT site, we measured the fraction of its RF that overlapped the stimulus/scotoma, and
193 correlated this value against the subsequent change in selectivity (F-statistic). The RF overlap
194 measure was computed using data from different recording days. There was a small but
195 statistically reliable correlation of RF overlap with selectivity change (selectivity change was
196 defined as $F_{control} - F_{cooling}$): during V4 cooling, the Pearson correlation coefficient was 0.19 and
197 0.32 (monkeys R and G: $P = 1.2 \times 10^{-3}$ and 1.3×10^{-7} , Student's t-test, $N = 300, 256$ sites). During
198 V2|3 cooling, the correlation coefficient was 0.11 and 0.26 ($P = 0.06$ and 2.3×10^{-5}). Note that the
199 stimuli were placed in the same overlapping region between both -V4 and -V2|3 scotomas, so
200 the lower correlation values for V2|3 cooling are not due to differences in scotoma overlap; rather
201 it is because the selectivity change is less pronounced for V2|3 cooling (if there was no selectivity
202 change, the correlation would be zero). We conclude that PIT multiunits lost selectivity across
203 images as a function of RF location.

204

205

a. F-ratio per cooling condition

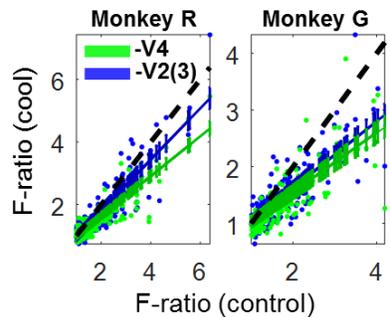
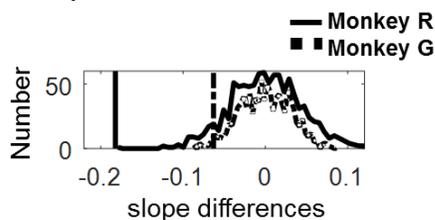


Figure 4. Selectivity of individual PIT sites before and during cooling.

(a) Scatterplots of F statistics (warm vs. -V4: green, warm vs. V2|3, blue). Each point shows the paired F-ratios for a given channel, measured before and during cooling. The solid colored lines show the mean slope describing the control and cooling F-ratio distributions. The error bars around each slope show the standard error. The broken black line shows unity.

(b) Differences in slopes expected given a mixed temperature distribution (solid curve = distribution from monkey R data, broken line = monkey G). The vertical lines show the experimental difference.

b. Slopes from randomized V4-V2 F-stats



206

Cooling did not reveal shape-specific deficits

207

208 We used regression analyses to find image features especially affected by cooling. These

209 features could be anything encoded by PIT, including luminance, contrast⁸, orientation content,

210 curvature content⁹ and categorical membership (e.g. “faces,” “body parts,” “tristars”, **Fig. 5**). In

211 addition to these features, we used principal component analysis on our images to extract 87

212 different quantitative descriptors for each of our 293 images (see **Supplementary Fig. 5**). We

213 also used experimental predictors, such as the mean population spike rate per image (before

214 cooling), and the mean decoding accuracy evoked per image (before cooling). We used all 87

215 features in a linear regression analysis that could explain the change in decoding accuracy per

216 image. In both monkeys, the only consistent predictor of V4- or V2|3-cooling accuracy loss was

217 the magnitude of classification accuracy before deactivation: the larger the classification accuracy

218 for each image before deactivation, the larger the subsequent reduction in accuracy. During V4

219 cooling, the percentage of variation explained by each model was 45-55% (monkeys R and G: R^2

220 = 0.45 and 0.55, $P = 4 \times 10^{-6}$ and 1.74×10^{-7} , $F(100,246)$: 2.0, $F(100,192)$: 2.4); during V2 cooling,

221 the R^2 values were 0.48 and 0.56 ($P = 1 \times 10^{-7}$ and 3×10^{-8} , $F(100,246)$: 2.3, $F(100,192)$: 2.5). The

222 linear model could not account for differences in decoding accuracy between V4 and V2|3 cooling

223 ($R^2 = 29\%$, $P = 0.84$, 0.91, $F(100,246)$: 0.8, $F(100,192)$: 0.8). In summary, no shape- or category

224 feature showed a statistical relationship with decoding accuracy reductions during cooling,

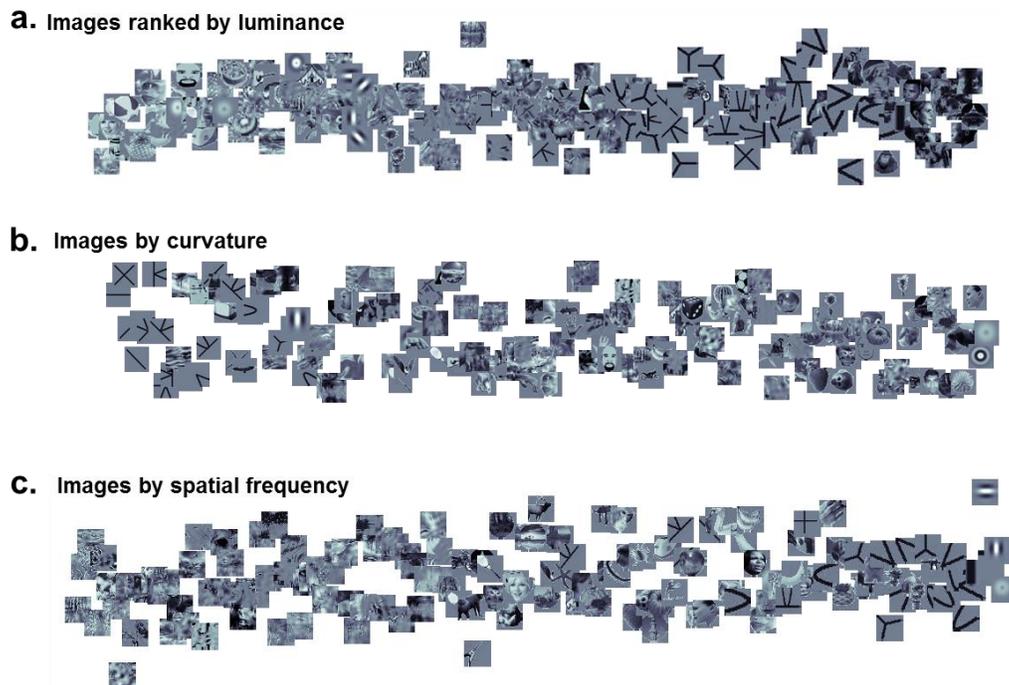
225 suggesting that a yet-undiscovered image property is differentially represented among the

226 pathways, or that image feature encoding does not differ between them. This oriented us to the

227 possibility that the primary difference between these pathways is not between the bypass

228 pathways, but between the long pathway and the bypass pathways themselves.

228



229

Figure 5. Examples of three image features used to predict changes in accuracy. A. Images sorted by their luminance value. **B.** Images sorted by their curvature content. **C.** Images sorted by the first Fourier-transform principal component (high vs. low spatial frequency content).

230

Cooling parallels in the Standard Model of Visual Recognition

231 The previous results oriented us to compare PIT units that received inputs from the long
232 (V1→V2→V4→PIT) pathway versus PIT units that received inputs from the short (V1→V2→PIT
233 or V1→V4→PIT) pathways. This was outside our experimental reach, because both of our cooling
234 interventions affected cells that depended on continuous information flow through the long
235 pathway. Thus we proceeded by exploring simulated versions of these two PIT types, using the
236 Standard Model of Visual Recognition, specifically the HMAX version of Serre, Oliva and Poggio
237 (2007). This is a hierarchical, feedforward-only model inspired by the visual system^{10,11},
238 comprising comprises multiple layers (*areas*), each with many filters (*receptive fields*) of different
239 sizes. The mean filter size per layer increases along the hierarchy, with small V1-like RF sizes in
240 the first layer, and large AIT-like RF sizes in the last layer. Each layer performs two serial
241 operations: a convolutional tuning operation followed by a pooling (invariance) operation. Inspired
242 by the simple/complex cells in V1, the convolutional operation provides a measure of similarity
243 between the input pattern and the “synaptic” weights of its filter. The outputs of different
244 convolutional steps are then pooled using a maximum operation; this is a non-linear step that
245 reduces multiple inputs into a single output, like a complex cell responding with its most active
246 simple cell input. This pooling step results in fewer responses feeding into the next layer and more
247 invariance to scale/position changes. At the highest layers of the model, there emerges a sparse
248 population of units, whose activations encode an abstract representation of the original image.
249 This response vector can be used in a final classification step to measure the accuracy of
250 representation of the original image.

251 Our version of this model had three alternative pathways that could provide input to a
252 given simulated PIT unit: the long pathway had four layers (V1→V2→V4→PIT); the second

253 pathway skipped the second layer (V1→V4→PIT), and the third pathway skipped the third layer
 254 (V1→V2|3→PIT, **Fig. 6a**). Filter widths doubled at each layer, but for the bypass pathways, RF
 255 size quadrupled at the bypass stage. This insured that all filters in a given area had the same
 256 range of sizes, irrespective of their inputs. The key difference in filter shapes between the long-
 257 and short-pathway inputs was the level of detail present in the filters: for example, V4 units shaped
 258 from V2 inputs were more abstract than V4 units shaped from V1 inputs, because the latter set of
 259 filters sampled visual activity which had undergone one fewer round of max pooling. At the end
 260 of the network, we decoded responses using SVMs in an all vs. all approach, with leave-one-out
 261 cross-validation and shuffled-label control.

262

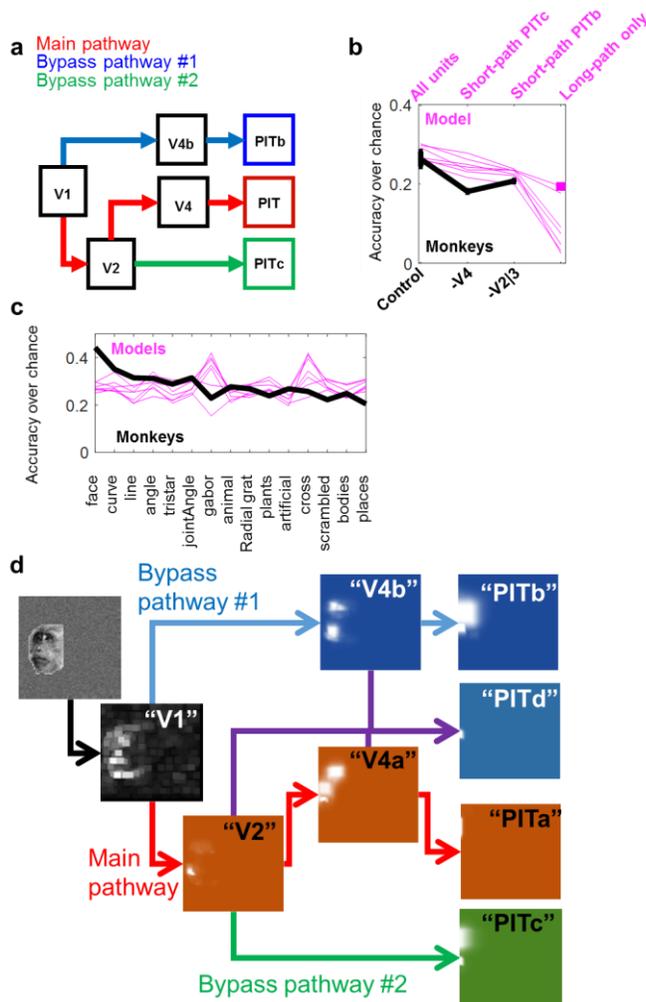


Figure 6. a. Modified standard model of visual recognition (HMAX architecture).

b. Mean classification accuracy over chance (\pm SEM), for both animals, during the warm, V4 and V2|3 deactivation (black) and for all simulated PIT populations in seven models (pink lines). The label “All units” refers to mixed short- and long-pathway PIT cells. The square under the “Long-path only” label shows the accuracy achieved when units were pooled from seven networks, in order to match cell count.

c. Accuracy per category for all seven models (pink). The black line shows the mean accuracy per category from the monkey data.

d. Example of activation of units at different layers in the hierarchy.

263

264 There were 58 units total at the final layer: 4 long-pathway units, 25 V1→V4→PIT units,
 265 25 V2→PIT units, and 4 units that received mixed inputs from long- and short-pathways. SVMs
 266 performed best at classifying each image when using a mix of long- and short-pathway units (N
 267 = 58), scoring $28 \pm 1\%$ after baseline subtraction, mean of seven models \pm SEM). For comparison,
 268 SVMs trained on the control monkey data showed an average of $26 \pm 2\%$ after baseline
 269 subtraction. When SVMs were trained using data from only one population of simulated PIT cells,
 270 they performed worse than when using the mixed population (PITc: $24 \pm 1\%$, PITb: $24 \pm 1\%$, PIT:
 271 $9 \pm 3\%$; **Fig. 6b**). Within each model, the number of simulated PIT units at each short-pathway was

272 the same ($N = 25$ units per final bypass layer), but the number of long-pathway units was lower
273 because of the additional pooling stage ($N = 4$); this partially explains why SVMs performed so
274 low when using long-pathway values only. To correct for this, we also trained SVMs using all the
275 long-pathway units from seven different networks: pooled together, these 28 units still performed
276 worse ($19 \pm 1\%$ above baseline) than SVMs trained on a mixed population. These reductions in
277 simulated PIT cell diversity were similar to the accuracy from cooling data (mean $-V4$ deactivation
278 performance was 16% and $-V2|3$ deactivation performance 19%, averaging both monkeys).

279 Unlike the monkey PIT units, which led to the highest classification accuracy for faces, the
280 entire population of simulated PIT model units showed highest accuracy values for line shapes,
281 bodies and artificial objects (24-34% over chance for all categories, **Fig 6c**). Visual examination
282 of the activation patterns within each layer highlighted interesting differences (**Fig. 6d**). Units in
283 the early layers ($V2$, $V4$) of the long pathway lit up local features, like eyes, but in subsequent
284 layers (PIT) of the long pathway, these features were pooled into more complex feature
285 selectivities. These long-pathway units looked like they would be good at discriminating complex
286 images, but at the expense of representing more primitive geometric features. In contrast, PIT
287 units receiving inputs from the bypass pathways seemed to preserve the explicit representation
288 of the more primitive sub features. We tested this observation as follows. We compared the image
289 selectivity to local and global features of two classes of simulated PIT cells – those at the terminus
290 of the long pathway or the short $V1 \rightarrow V4 \rightarrow$ PIT pathway. We used images of 10 faces and 10
291 quadrupeds from our 293-image set. We used SVMs to interrogate the long- and short-pathway
292 units on four tasks, where success in each task depended on discriminating local vs. global
293 features. The tasks were to classify 1) faces vs. quadrupeds, 2) heads vs. faceless heads, 3)
294 faces vs eyeless faces, and 4) quadrupeds vs. legless quadrupeds (**Fig. 7**). The first two
295 comparisons involved many local features (a global comparison), the second two comparisons
296 involved a local feature. The pathways performed differently: for the global task, the long-pathway
297 units showed better performance than the short-pathway units (accuracy classifying faces vs.
298 heads, long-pathway: 0.58 ± 0.05 , short-pathway: 0.54 ± 0.06 ; faces vs quadrupeds: long-pathway:
299 0.58 ± 0.05 , short-pathway: 0.57 ± 0.07). For the local-feature tasks, the short-pathway units
300 allowed better performance than the long-pathway units (accuracy detecting missing eyes, long
301 pathway: 0.58 ± 0.05 , short pathway: 0.64 ± 0.04 ; accuracy detecting missing legs: long pathway:
302 0.53 ± 0.06 , short pathway: 0.66 ± 0.06). This suggested that in the macaque brain, short pathways
303 could be helpful for fine, local-feature discriminations, while the long pathways could implement
304 Gestalt-like, global discriminations.

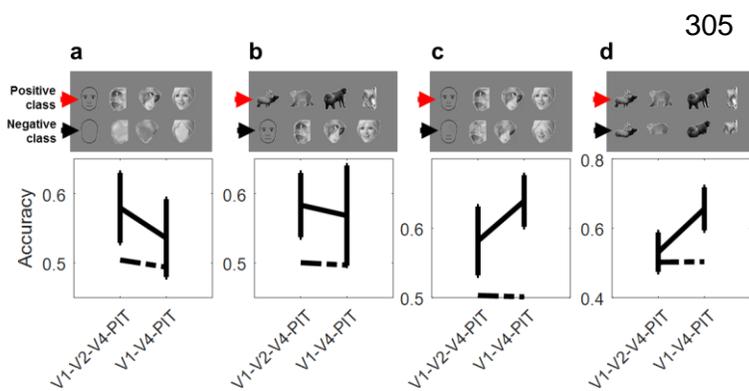


Figure 7. Sensitivity to image parts by long- and short-pathway units.

Classification accuracy differences among different pathways. Solid lines show mean SVM accuracy values, error bars show standard error of the mean. Broken lines show accuracy after label shuffling. **a.** Heads vs. faceless heads. **b.** Faces vs. quadrupeds. **c.** Faces vs. eyeless faces. **d.** Quadrupeds vs. legless quadrupeds.

306

307 Cooling affected fine discrimination less than coarse discrimination

308 The cooling interventions always disrupted the long pathway but preserved at least one
309 short pathway. If the short pathways are more important for local (fine) discriminations and the
310 long pathway more important for global discriminations, then during cooling, SVMs should perform
311 better with local discriminations because of these remaining short pathways. We wanted to test
312 this hypothesis, so first we had to define “local” vs. “global” discriminations in the context of our
313 data.

314 The decoding accuracy analysis (**Fig. 2**) showed that monkey PIT populations
315 discriminated some images more accurately than others. The image pairs that were hard to
316 discriminate must have had similar locations in the neural activity space: the closer the images,
317 the more likely that individual trials were on the wrong side of the dividing linear hyperplane. We
318 ranked the decoding accuracy for each image as a function of the Euclidean distance to its
319 neighbors (in the original 100+ dimensional activity space). This plot showed that decoding
320 accuracy for two images increased with their distance in activity space. It also showed that very
321 close neighbors were systematically misclassified (**Fig. 8a**, bottom). Thus we can interpret
322 distance as a continuous metric for local-feature vs. global-feature discriminations: close image
323 neighbors activate similar feature detectors in PIT relative to distant image pairs.

324 We then asked whether the cooling reduction in accuracy varied with neighbor distance,
325 and if so, whether that change could be predicted by a random cooling mechanism like those of
326 **Figure 3**. These random cooling mechanisms were blind to the identity of each image and to their
327 distances in activity space. This created two hypotheses: the first hypothesis was that the change
328 in decoding accuracy as a function of distance will be either a simple multiplicative or subtractive
329 change, captured by random cooling mechanisms. The second hypothesis is that the change in
330 decoding accuracy as a function of distance will not be explained by a simple gain change, but in
331 fact will be better preserved within short distances, as predicted by the convolutional model of
332 visual recognition.

333 We plotted the decoding accuracy as a function of distance for the random cooling
334 mechanisms and found that they showed a series of multiplicative changes to the control (warm)
335 accuracy-distance plot (**Fig. 8b**); these curves showed small changes in accuracy for near-
336 neighbors, and larger accuracy changes for far neighbors. In contrast, when we plotted the cooling
337 data’s decoding accuracy as a function of distance, we found that these curves differed from the
338 simulated-cooling curves within the range of short distances (**Fig. 8c-d**). The random cooling
339 mechanism curves predicted small decrements in misclassification, but the cooling data showed
340 a considerable improvement in classification. The cooling data and the simulated cooling data
341 curves were otherwise well-matched at the mid- to far range of neighbor distance. To compute
342 the statistical reliability of this observation, we derived the probability that the cooling median
343 accuracy at each distance could be emitted by the simulation (we ran a Wilcoxon rank-sum test
344 at the i th neighbor position, asking if the median accuracy measured in the cooling distribution
345 was higher than the median accuracy derived from the simulation). We found that the nearest 28
346 neighbors were reliably better classified than predicted by the simulation (P values ranged from
347 4×10^{-35} to 0.03 within the nearest 28 positions, one-tailed paired Wilcoxon signed rank test, $N =$
348 640 images, Z -values = 1.9-12.3). This resilience in decoding accuracy for near neighbors is
349 consistent with the interpretation that short pathways are most helpful with fine discrimination.

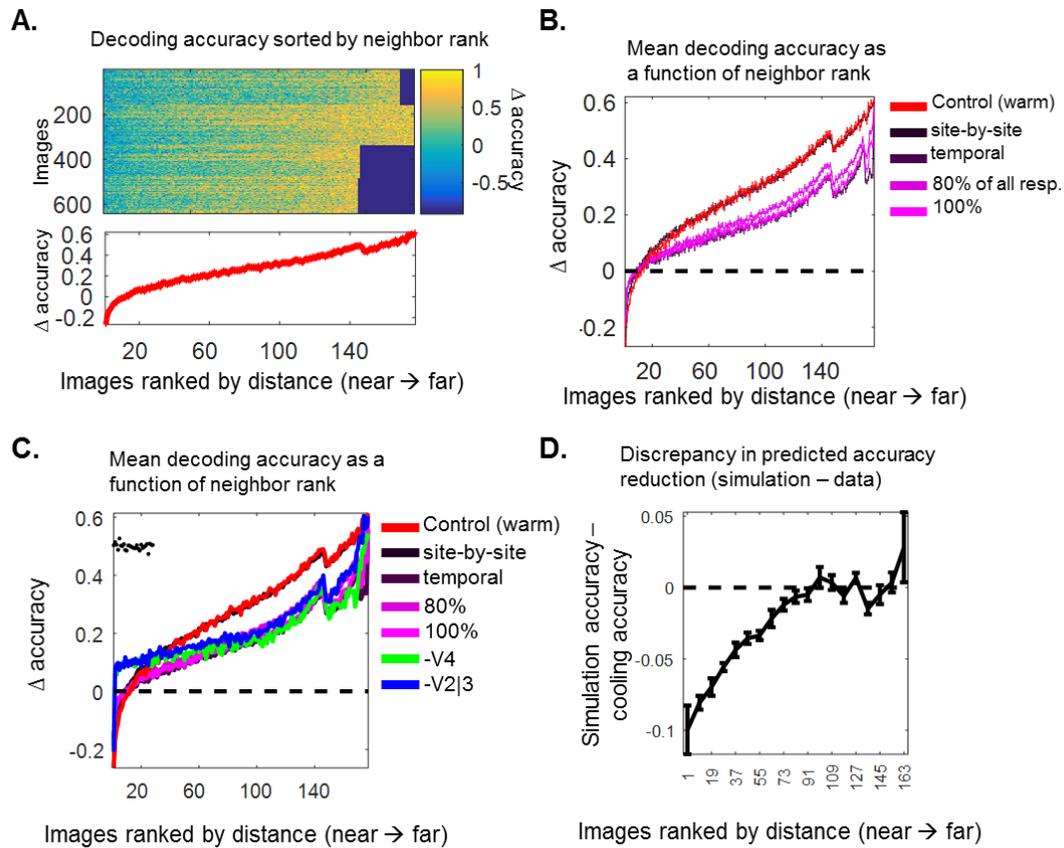


Figure 8. Differences in decoding accuracy for near- vs. far neighbors.

a. Linear decoding accuracy for every image (rows) when compared to its neighbors, as sorted by rank distance (columns). Top figure shows the decoding accuracy for all images, across days and monkeys. Bottom figure shows the mean decoding accuracy for each rank position (±SEM). Colors show the decoding accuracy minus the shuffled-label accuracy.

b. Mean decoding accuracy (minus shuffled baseline, ± SEM) measured before cooling (red). The additional curves, colored in magenta, show the values for the different cooling simulations (see Figure 3 for a description of each mechanism).

c. Same as (c), but with data from V4 cooling (green) and V2|3 cooling (blue). Black dots show which image ranks show statistically differences between the cooling data and trial-by-trial model.

d. Discrepancy in predicted accuracy reduction between the simulation (trial-by-trial) and V4 cooling data. Each point shows the average for 10 neighbors (±SEM), from nearest to furthest.

350

351

DISCUSSION

352

We investigated how posterior inferotemporal cortex cells combine information from areas V2, V3 and V4. We implanted microelectrode arrays in PIT while cooling areas V2/V3 or area V4. PIT multiunits showed a reduction in firing rate that was similar across both types of interventions. Support vector machines showed that classification accuracy was reduced more during V4 deactivation compared to V2|3 deactivation. Changes in classification accuracy could not be predicted by any class of visual features, such as contrast or spatial frequency content. This finding is consistent with previous iterations of the Standard Model of Visual Recognition, which implemented a hierarchical convolutional network where different bypass projections contain the

359

360 same types of visual features. This previous model also modeled bypass pathways and predicted
361 a role for fine discrimination, and here we explore that idea further². We confirmed that these
362 pathways provide an advantage to simulated PIT units in fine discriminations. According to the
363 model, the key contrast was between long-pathway PIT cells (V1→V2→V4→PIT) vs. short-
364 pathway PIT cells (V1→V4→PIT), and these populations could not be isolated using our cooling
365 technique. However, we found that images that are represented similarly in PIT activity space
366 were better discriminated during cooling; this advantage was not predicted by non-specific cooling
367 simulations.

368 Our principal conclusion is that input pathways of different lengths create a diversity of
369 functional preferences in a given cortical region. We argue that multiple pathways allow some
370 units to respond to complex features, and some units to simpler features. This raises several
371 questions. First, in the context of a fine discrimination task, why would units with complex image
372 preferences be less helpful than units with simpler image preferences? Cells with complex image
373 preferences can be less responsive to variations of their preferred image: Kobatake and Tanaka
374 have shown that IT cells have a minimum number of “critical features,” a minimal combination of
375 image parts required to elicit a response¹². If any of those features are missing, the cell may not
376 respond at all. Thus in the context of detecting the presence of simple parts within a complex
377 image, there may be an advantage of having cells tuned for the complex image along with cells
378 tuned for its simple parts. This type of simpler preference is more likely to arise earlier in the visual
379 hierarchy and become preserved by short bypass pathways. Did we miss any classes of 2-D,
380 luminance shape features that might be differentially represented between the V2|3 and V4
381 paths? While possible, there are no strong theoretical candidates for such features. Hegde and
382 Van Essen (2007) compared the relative shape selectivities in neurons from V1, V2 and V4, and
383 showed that these cells responded similarly to the same set of simple and complex images,
384 offering little qualitative diversity¹³.

385 The V1-V2|3-V4 input pathway is the most important source of excitatory activity to PIT,
386 and one might expect that its disruption would extinguish nearly *all* activity within the PIT scotoma.
387 However, PIT units within the scotoma continued to respond, albeit weakly. This was because we
388 did not cool all of V2, V3 or V4 – the scotomas covered only a few degrees of central vision, and
389 near-foveal representations receive a disproportionate amount of cortical real estate in the brain.
390 Indeed, cortical areas deep in the visual hierarchy receive too many inputs and may be practically
391 impossible to silence without bilateral V1 resection, silencing of lateral connections and feedback:
392 as an extreme example, anterior IT cells show no statistical change in overall firing rate after
393 surgical resection of areas V4 and PIT¹⁴. Ultimately, we will get more answers to the problem of
394 multiple pathways in macaque by recording from single units with known anatomical input profiles,
395 a goal that may be met with the use of chemo- or optogenetics.

396 **METHODS**

397 All procedures were approved by the Harvard Medical School Institutional Animal Care
398 and Use Committee, following the Guide for the Care and Use of Laboratory Animals (8th edition,
399 National Academies Press). This paper conforms to the ARRIVE Guidelines checklist.

400 **Behavior.** Two adult male macaques (10 -17 kg) from the New England Primate Center
401 were trained to perform a fixation task. The task required them to stare at a 0.5°-wide red or black
402 square in the middle of the screen, keeping their gaze within $\pm 1.3^\circ$ from the fixation spot. We used
403 an ISCAN eye monitoring system to keep track of eye movements (www.iscanin.com). The trial
404 timeline was as follows: at the start of each trial, the fixation target appeared and the animal had
405 up to eight seconds to direct its gaze to the fixation target. Once fixation was acquired, a small

406 reward could be delivered to encourage the animal. Within a random period between 17- 117 ms
407 after fixation onset, an image appeared perifoveally for 200 ms, then disappeared for 200 ms until
408 a new image appeared. This on-off cycle could be repeated with 3-5 different images per trial. If
409 the animal held fixation until the end of the final on-off cycle, a reward was dispensed. The reward
410 size increased by 25% of the initial reward size every 100 trials.

411 **Animal care.** The animals were housed in a vivarium with a 12-h light/dark cycle, under
412 social pairing. They worked during the daytime. The monkeys had no previous major surgical
413 history.

414 **Visual stimuli.** We used MonkeyLogic to control experimental workflow
415 (<http://www.brown.edu/Research/monkeylogic/>). We used 293 high-resolution images, including
416 photographs and simple shapes. Photographs came from Google Images, and our choice of
417 pictures were guided by categories used in previous IT studies^{15,16} including animals, artificial
418 gadgets, body parts, faces, places, plants/fruits(20-21 examples per category; faces and body
419 parts were evenly distributed between monkey and human). Most of these images were used to
420 create scrambled counterparts via the Portilla and Simoncelli visual texture model¹⁷, which
421 transforms white noise into textures that share pairwise joint statistical constraints as the original
422 intact images. These textures convincingly replicate small shape primitives present in the original
423 images and scatters them throughout the image (118 textures). We also used 54 simpler line
424 shapes such as lines, curves, tristars, radial and linear gabors, and simple combinations of lines
425 and curves (joint angles). These line shapes were generated using the Cogent Matlab toolbox
426 (developed by John Romaya at the LON at the Wellcome Department of Imaging Neuroscience,
427 http://www.vislab.ucl.ac.uk/cogent_graphics.php). Images were 1.4° in width (for monkey G) or
428 2.0° (for monkey R) at their longest axis. The images were not normalized for luminance, contrast
429 or other visual properties.

430 **Implanted devices.** The cryoloops were manufactured in the laboratory of Stephen
431 Lomber and are described in Lomber, Payne and Horel (1999)¹⁸. Cryoloops were composed of
432 23-gauge hypodermic stainless steel tubing, shaped to fit the individual curvature of each animal's
433 occipitotemporal shapes as determined by structural magnetic resonance images. The cryoloops
434 were 3.5 mm wide and between 4-11 mm long. A microthermocouple sensor was attached to the
435 stem of the cryoloop to monitor its temperature. The bodies of the cryoloops were wrapped in
436 Teflon tubing except at the loop. The loops contained protected inlet/outlet ports that permitted
437 the daily connection of Teflon tubes carrying chilled methanol, as driven by FMI "Q" Pumps (Model
438 QG150, fluidmetering.com). The methanol was contained within the tubing system and could not
439 cause any chemical harm to the tissue. The custom floating microelectrode arrays were
440 manufactured by MicroProbes for Life Sciences (Gaithersburg, MD); each had 32 platinum/iridium
441 electrodes per ceramic base, electrode lengths of 4-16 mm, impedances between 0.7- 1.0 MΩ,
442 all connected to a 36-channel Omnetics connector (allowing for two additional grounds and two
443 reference electrodes).

444 **Surgical procedures.** Both animals were implanted with custom-made titanium
445 headposts before fixation training. After several weeks of post-surgical recovery and fixation
446 training, the animals underwent a second surgery for the implantation of cryoloops and floating
447 microelectrode arrays. Animals were anesthetized using ketamine/xylazine (I.M.) and isoflurane;
448 buprenorphine/non-steroidal anti-inflammatories were used for pain control. In each animal, we
449 performed a craniotomy centered at the lunate sulcus and extending antero-laterally. Monkey R
450 received three cryoloops, two placed within the left lunate sulcus and one over the prelunate
451 gyrus. The medial lunate sulcus loop was located 20 mm from the midline, traveled 7 mm deep
452 into the sulcus and was 3 mm wide; the lateral lunate sulcus loop traveled 4.5 mm into the sulcus

453 and was 3.5 mm wide); the prelunate gyrus loop was placed anteriorly to the lunate sulcus loops,
454 was 11 mm long and 3 mm wide. Monkey G received two cryoloops, one over the prelunate gyrus
455 and one within the lunate sulcus. The lunate sulcus loop was placed 2.1 cm from the midline, was
456 11 mm long with this axis running in the mediolateral axis within the lunate sulcus, 3 mm wide,
457 and its most dorsal edge was 1.5 mm deep. The prelunate gyrus loop was also placed 2.1 cm
458 from the midline, anteriorly to the lunate sulcus loop, ran 10.5 mm long and was 3 mm wide. We
459 collected thermal images to map the spread of cooling from the tubing, and confirmed that it was
460 limited to 1-3 mm radially, as first shown in previous publications¹⁹. Two to three floating
461 microelectrode arrays were implanted within the same intraoperative session, after placement of
462 the cryoloops. Their insertion sites were determined using three guidelines: they had to be anterior
463 to the inferior occipital sulcus, many millimeters away from the prelunate gyrus cryoloop, and
464 avoid large vasculature. All arrays were implanted caudal to the posterior middle temporal sulcus.
465 We implanted two 32-channel arrays in monkey R, and one 32-channel plus two 16-channel
466 arrays in monkey G.

467 **Experimental session workflow.** All data reported was collected within a couple of
468 months after implantation. Each day, the animal would be head-fixed and its implants connected
469 to the experimental rig: first the cryoloops were connected to the chilled-methanol-bath tubing and
470 temperature sensors, then the microelectrode arrays were attached to their headstages. The first
471 step each day was to calibrate our measurements of the animal's gaze using the built-in
472 MonkeyLogic routine. We used the Plexon Multichannel Acquisition Processor (MAP) Data
473 Acquisition System to collect electrophysiological information, including high-frequency ("spike")
474 events, local field potentials and other experimental variables, such as eye position, reward rate
475 and photodiode outputs tracking monitor frame display timing. Each channel was auto-configured
476 daily for the optimal gain and threshold; we collected all electrical events that crossed a threshold
477 of 2.5 standard deviations from the mean peak height of the distribution of electrical signal
478 amplitudes per channel. These signals included typical single-unit waveforms, multi-unit
479 waveform bursts, and visually-active hash.

480 The animal began its fixation task while we collected responses from the arrays with the
481 cryoloops at body temperature (36-37°C; this is what we call "control" or the "warm" condition).
482 After ~20 minutes of data collection to permit ~4-6 repetitions of each image, we activated either
483 the V2|3 or V4 cryoloops, bringing the temperature of the cryoloops to ~9°C, which lowered the
484 temperature of the adjacent cortex to 16-18°C. We waited for another 5 repetitions of each image
485 to pass, and then turned off the cryoloop pumps and collected 1-2 more repetitions under this first
486 re-warming session. We then paused the fixation task for 10 minutes to allow the tissue
487 temperature to return to normal and preserve the animal's motivation for a second round of
488 cooling. After 10 minutes, the temperature reported by the cryoloops was around 34°C, and we
489 re-started our experiment. We repeated each image presentation 3-4 times and then activated
490 the second set of cryoloop(s) (~8°C), waited for 5 repetitions and turned off the cryoloops. We
491 then collected data until the animal was satiated. We balanced the order of the V2|3 vs. V4
492 cryoloop activations: if on the first day we activated the V4 cryoloop first and V2|3 cryoloops
493 second, the next day we activated the V2|3 cryoloops first and V4 cryoloop second. There was
494 an even number of days for each cooling order.

495 **Electrophysiology data preparation.** The raw data files comprised event ("spike") times
496 per channel for the entire experimental session (the number of channels available per day were
497 64, but not all provided reliable signal-to-noise qualities). We divided each daily data set into
498 thousands of raster plots defined by the onset of each image presentation and labeled each raster
499 plot with its corresponding channel, image name and temperature condition. We defined three

500 windows of analysis: the baseline period lasted from 0-50 ms after image onset, the early period
501 from 51-150 ms after image onset, the late period from 151-250 ms after image onset, a full image
502 presentation window was 51-400 ms after image onset. We found that multiunit responses could
503 last almost 400 ms, although their peak responses always occurred within the early window. Here
504 we report responses within the full window minus the activity within the baseline window (we call
505 these evoked responses). For all multivariate analyses, we normalized the activity of each site by
506 transforming its evoked responses to z-scores: all evoked responses emitted by a single site
507 during an experimental daily session were averaged, this mean response was subtracted from all
508 individual evoked rates, and each value was then divided by the standard deviation of all evoked
509 responses.

510 Although our full dataset contained 293 images, we did not have enough time to present
511 all images every day and still get the minimum of 15 presentations across the control and cooling
512 conditions. Thus we presented over half of the total image set each day (10 images from each
513 complex category, such as faces and places, along with half of the scrambled textures per day,
514 with most of the simple line shapes, rounding to about ~170-177 unique images per day). The
515 responses of a given channel were correlated across days, but were also statistically different by
516 multivariate descriptors such as multi-dimensional scaling. Because of these differences, we did
517 not combine channel information across days and instead created a multi-day pseudo-population,
518 where sets of concurrently recorded channels ($N = 50-64$) from different days were treated as if
519 recorded at the same time (see Ch. 19 of reference 20)²⁰. Thus the final activity space is defined
520 by firing rates collected across “site-days,” where some dimensions represent responses from the
521 same channel to the same image collected on different days. Because the whole image set was
522 presented on different days, we had two pseudopopulations per animal, each containing different
523 site-day responses to each half of the image set. Each of our pseudopopulations had between
524 100-300 multi-units.

525 **Scotoma mapping experiments.** The goal of these experiments was to identify the parts
526 of the retinotopic field that were captured by our arrays, and the relative location of the response
527 impairment caused by cooling. To achieve this goal, we had the animals fixate while we presented
528 a single image (black-and-gray cartoon face, 2.0°-wide) within all positions in a radial grid (angular
529 coverage of 0-315°, 45° steps; radial coverage of 0-8° from the center of the screen, in 0.5° steps).
530 Three to five positions were randomly chosen per trial. After data collection, we defined evoked
531 responses per position as follows: first we quantified the firing rate per site during the early window
532 of activity (51-151 ms after stimulus onset) and then subtracted the firing rate per site during the
533 baseline window of activity (0-50 ms after stimulus onset). We averaged these evoked responses
534 per position within each site and used the `griddata.m` Matlab function to interpolate the scattered
535 data into a continuous map. This map was smoothed using a 1°-diameter disk filter. This map
536 represented the aggregate receptive field of each multiunit site in our arrays. To identify the overall
537 scotoma, we averaged the response fields of all sites during the control condition and subtracted
538 the average response fields of all sites during V2|3 or V4 deactivation. We measured the size of
539 each scotoma by hand, using the `calcArea.m` function
540 (<http://www.mathworks.com/matlabcentral>).

541 **Firing rate and latency.** The goal of these analyses was to measure changes in the
542 overall firing rate (excitatory drive) of PIT multiunits during input deactivation. These changes
543 included the amplitude of peristimulus rate histograms (PSTHs) and the latency of response. To
544 quantify the changes in evoked response magnitude, we computed the evoked responses per
545 site as described in *Ephys Data Preparation* and averaged these responses across all channels
546 within each temperature condition. We did the same operation using z-scores. We calculated the

547 probability that the median responses emitted during each temperature condition (control, V4 and
548 V2|3 cooling) were sampled from the same distribution using a Kruskal-Wallis one-way analysis
549 of variance. To determine if there was a statistical difference between the V4 and V2|3 cooling
550 condition responses, we used the Wilcoxon signed rank test for zero median. For the latency
551 analyses, we obtained the mean PSTH in response to each image, per site and temperature, and
552 then stacked all image-specific PSTHs in a matrix measuring $N_{\text{images}} \times 400$ (ms after stimulus
553 onset). We identified the time when each PSTH exceeded two standard deviations over baseline
554 and called this *response latency*, with the only acceptance criteria that a plausible response
555 latency would only occur between 30-200 ms after image onset. We also computed the earliest
556 time point when all PSTHs demonstrated the greatest variance in amplitude, as an indicator of
557 the *tuning* latency.

558 **How we identified channels with reliable visually driven activity.** Many electrodes in
559 the arrays reported electrical activity that was not visually driven, possibly because the electrodes
560 were on the pial surface. We repeated this analysis only using channels that showed a statistical
561 difference in mean activity between the baseline and evoked time periods. Using a cross-
562 validation approach, we used 5% of all trials to perform a Wilcoxon signed rank test for the median
563 rate difference during each interval. This told us which channels showed a statistical difference in
564 rate during visual stimulation. We then used the remaining 95% of trials to compute the firing rates
565 during baseline and evoked windows for the selected channels. Monkey R's arrays showed 38
566 out of 64 visually responsive sites; monkey G, 30 out of 64 ($P < 0.05$, two-tailed Wilcoxon signed
567 ranked test for zero median).

568 **Encoding accuracy analyses.** We trained support vector machines with a linear kernel
569 using the Matlab function `fitcsvm.m`. We used an all-vs.-all approach, with SVMs trained to
570 discriminate between pairs of images, using leave-one-out cross-validation. There were 4-5
571 response vectors per class within each comparison (the data used for classification were Z-score
572 vectors; see *Spike Data Preparation*). To estimate the chance accuracy for each paired
573 comparison, we concurrently trained SVMs using the same set of data vectors but with shuffled
574 labels. The number of vectors for each two-class comparison was small, and thus we found that
575 chance accuracy values could vary between 0-1 across all comparisons; the median shuffled-
576 label misclassification rates were 0.60-0.63 for monkeys R and G. We subtracted the chance,
577 shuffled-label accuracy classification rate from the correct-label accuracy classification rates to
578 account for this bias. As an insight to explain this deviation from the expected chance accuracy
579 of 0.5, we trained SVMs to distinguish between stimulus categories (listed in the *Visual Stimuli*
580 section). Each category pair comparison involved 10-20 times as many response vectors as the
581 individual image-vs-image SVM analyses, and the dataset was otherwise identical. Here we found
582 a more reassuring shuffled-label statistical baseline of 0.50 in both animals. Both the category
583 and image-per-image SVM accuracy analyses led to the same conclusions presented in the
584 *Results* section.

585 **Projection analysis.** The goal of this analysis was to reconcile the findings that cooling
586 V2|3 and V4 led to equivalent reductions in PIT population firing rates, but different reductions in
587 classification accuracy. We measured the cooling trajectory traveled by each image during V2|3
588 or V4 cooling and to project it onto the direction of a minimum response vector, where this
589 direction represented a non-specific reduction in firing rate across all sites. We conducted the
590 analysis as follows; first, for every pseudo-population, we defined a minimum response vector
591 $\mathbf{v}_{\min} = \{\min \arg(\mathbf{x}_j)\} = \{\min \arg(x_1), \min \arg(x_2), \dots, \min \arg(x_n)\}$, where \mathbf{x}_j is a variable representing
592 all the mean responses from the j th array site and n is the number of sites, thus the vector length
593 is n . Next, we computed three mean vectors per image: $\mathbf{v}_{\text{warm}} = \text{mean response vector for the}$

594 given image during the control condition, also of length n ; \mathbf{v}_{-v4} = mean response vector for the
595 given image during V4 cooling and \mathbf{v}_{-v2} = mean response vector for the given image during V2|3
596 cooling. We computed the cooling trajectory vector of each image as $\mathbf{V}_{\text{warm-v4}} = \mathbf{V}_{\text{warm}} - \mathbf{V}_{-v4}$ and
597 $\mathbf{V}_{\text{warm-v2}} = \mathbf{V}_{\text{warm}} - \mathbf{V}_{-v2}$. Finally, each individual trajectory was projected onto the vector $\mathbf{V}_{\text{warm}} - \mathbf{V}_{\text{min}}$;
598 the projected (parallel) component was subtracted from the cooling trajectory vector to compute
599 the perpendicular component.

600 **Cooling simulations using the control data.** We simulated the effects of cooling on the
601 control decoding accuracy, by applying fractional reductions to the control spike rates. The
602 approach was to first identify the fractional reductions in firing rate for all channels during cooling
603 in a given pseudo-population and then to use this fractional distribution to simulate cooling
604 changes with data from a different pseudo-population. Fractional reductions were defined as $f_i =$
605 $r_{i, \text{cooling}} / r_{i, \text{control}}$, where r_i is the mean firing rate for a given channel i in a pseudo-population $i = 1-$
606 300 . We sampled values from each fractional distribution with replacement and multiplied warm
607 firing rates in three different ways. In all cases, we can envision the control data set as a matrix
608 of dimensions $r \times c$, where rows are multiunit sites (channels) and columns are individual image
609 presentations. In the first mechanism, *site-by-site cooling*, we sampled r fractional changes and
610 multiplied each sample times all the responses in one channel. In the second mechanism,
611 *temporal cooling*, we sampled c fractional changes and multiplied each fractional change times
612 all elements in the column. Finally, in the third mechanism, we randomly selected a given
613 percentage of responses in the matrix, and multiplied them by an equal number of sampled
614 fractional values (the *mixed site-by-site* and *temporal cooling*). We used each transformed control
615 matrix to train and test support vector machines as described above.

616 **Selectivity analyses (F-statistics).** In this analysis, the F-statistic was used as a
617 measure of selectivity for each multiunit. The F-statistic is a ratio of mean squares, specifically
618 the mean square error estimate for the variance of responses among images, divided by the mean
619 square error estimate for the variance within each image. We computed each F-statistic in a
620 channel by channel basis using the responses to all images within each temperature condition.
621 For each channel, one F-statistic was computed using the warm data (F_{control}), another using the
622 V4 cooling data (F_{-v4}) and another using the V2|3 cooling data ($F_{-v2|3}$). We plotted each F_{control}
623 against its paired F_{-v4} and $F_{-v2|3}$ values. To determine if the slope in each given scatterplot was
624 different from unity, we used a bootstrap approach, where we computed 1,000 different slopes by
625 sampling each channel with replacement (note that we kept each F-ratio trio together; we did not
626 mix warm and cooling F-statistics from different channels). We then asked if the slope distribution
627 from this bootstrap included 1.

628 We used randomization to measure any differences between the mean slopes computed
629 during the V4 and V2|3 cooling conditions (that is, whether there was a difference between the
630 mean $F_{-v4} / F_{\text{control}}$ slope vs. the mean $F_{-v2|3} / F_{\text{control}}$ slope). The null hypothesis is that the mean
631 V4 and V2|3 slopes came from the same distribution. Therefore, we had to create this null
632 distribution. In each of 1,000 passes, we randomly mixed the labels between the V4 and V2|3 F-
633 statistics for each channel and computed a $F_{\text{cooling}} / F_{\text{control}}$ slope. We did that twice per pass, and
634 then subtracted the two slopes. After 1,000 passes, we had 1,000 slope differences that we then
635 compared to the experimental slope difference. We found that these null difference distributions
636 were defined by 5th and 95th percentile values of -0.07 to 0.07 (monkey R) and -0.05 to 0.05
637 (monkey G). The observed differences in mean cooling slopes were -0.18 and -0.06 (monkeys R
638 and G). The probability that the experimental differences in V4 and V2|3 slopes came from such
639 mixed distributions were 0.001 and 0.02.

640 **How we defined response field overlap with the scotoma, for the F-statistic analysis.**

641 For each channel, its RF overlap was defined as the average number of spikes emitted in
642 response to stimuli presented in the stimulus/scotoma region, divided by the total number of
643 spikes emitted in the central $8 \times 8^\circ$. The mean RF overlap value was 0.12 ± 0.01 and 0.15 ± 0.01
644 (monkeys R, G).

645 **Linear regression model.**

646 The goal of this analysis was to determine whether the change in classification accuracy
647 during V4 cooling or during V2|3 cooling could be predicted using different image features. The
648 regression matrix had dimensions of 293×87 (images \times visual features). The features were
649 luminance (defined as the mean pixel value transformed by the monitor's gamma function),
650 contrast (variance of the pixel values transformed by the monitor's gamma function), horizontal
651 vs. vertical power (obtained via a wavelet decomposition analysis using the Matlab function
652 `wavedec2.m`), curvature (defined by the variance of each image's discrete Fourier transform
653 spectral power around all orientations), 50-pixel-based principal components as defined by the
654 `pca.m` function), 30 spatial frequency principal components (`pca.m` applied to the discrete Fourier
655 transformed images), categorical membership (defined *a priori* as angles, animal, artificial,
656 bodies, cross, curve, face, gabors, radial gabors, joint angles, line, places, plants, scrambled,
657 tristar), the mean population control firing rate per image and control classification accuracy per
658 image. Values within each feature group were z-scored before fitting. The dependent variables
659 were either 1) accuracy loss during V4 cooling (control accuracy per image minus $-V4$ accuracy),
660 2) accuracy loss during V2|3 cooling (control accuracy per image minus $-V2|3$ accuracy) or 3)
661 the difference in accuracy loss during V4 minus V2|3 cooling ($[\text{control accuracy per image minus}$
662 $-V4 \text{ accuracy}] - [\text{control accuracy per image minus } -V2|3 \text{ accuracy}]$). The probability that the
663 linear model differed from the constant model was obtained two ways: first, we used the t-statistic
664 provided by the `fitglm.m` function; second, we used a randomization test where the dependent
665 variable was fit with a regression table made up of random numbers, sampled from a flat
666 distribution. The table had the same dimensions as the true data matrix table. The R^2 values of a
667 thousand randomization tests were compared to the R^2 from the regular regression table. To
668 identify the most interesting predictors, we looked at all 87 regression weights and their t-statistics.
669 There was one clear outlier: the control classification accuracy (t-statistics 8.6-8.7). We also fitted
670 a regression analysis that penalized the number of regression weights (the Lasso). To insure that
671 this was not simple regression to the mean, we also divided our control trials such that the control
672 classification tuning curve used for regression was not the same as the control classification
673 tuning curve used to calculate the cooling difference in classification accuracy (the estimated
674 correlation between these cross-validated data sets were 0.33-0.45 for monkeys R and G, $P <$
675 10^{-8} , two-tailed Student's T-test).

676 **Standard model of visual recognition**

677 Our computational model was based on an implementation by Serre, Oliva and Poggio
678 (2007)⁵, available at <http://cbcl.mit.edu/software-datasets/standardmodel/index.html>. This model
679 belongs to the family of hierarchical feedforward models (HMAX) by Riesenhuber and Poggio
680 (1999)²¹ and developed over subsequent publications^{2,22,23}. The model represents the visual
681 object recognition system as a series of convolutional and pooling operations, which transform an
682 image from pixels into neuronal responses. These responses can be used in a statistical classifier
683 to decode their abstracted representational content.

684 The architecture of our network was three to four layers deep and contained three
685 pathways: one main pathway and two bypass pathways. The main pathway had four layers: layer
686 1 (representing V1), layer 2 (V2|3), layer 3 (V4), and layer 4 (PIT units receiving inputs from the

687 main pathway). The second pathway had three layers: layer V1, layer V4b (representing units in
688 V4 receiving direct input from V1) and layer PITb (PIT units receiving input solely from the V1→V4
689 inputs). The third pathway also had three layers: layer V1, layer V2|3 and layer PITc (units in PIT
690 receiving input directly from V2|3). These three types of “PIT” neurons showed different kinds of
691 activation patterns, which we could decode using support vector machines.

692 Layers. Each layer represented a stereotypical set of operations: a convolution/tuning
693 operation and a pair of max operations. The tuning operation is equivalent to a simple cell, which
694 convolves the input with a filter bank via the tuning function $r = \exp(-\frac{1}{2\sigma^2} \sum_{j=1}^{N_{comb}} (w_j - x_j)^2)$,
695 where σ = sharpness parameter, N_{comb} is the number of filters to combine, w = filter weight and x
696 = input image or activity. This simple cell operation describes the Euclidean distance between the
697 RF shape and the incoming input. Different simple cells are characterized by different shapes and
698 sizes of their RF patches. There are more than one receptive field sizes at each layer, and each
699 filter-size convolution is performed in parallel. Several outputs of this tuning operation are then
700 combined in a complex-cell-like operation. *Complex cells* perform a pooling operation: they
701 receive inputs from N_s simple cells with different RF sizes and compute the maximum response
702 emitted by the set. Thus the output of a complex cell layer is sparser than the output of a simple
703 cell layer, because maximum values are repeated across limited areas of response space. These
704 complex layer responses are finally subsampled, imitating the decreasing number of cells that
705 can cover visual space as one moves down the visual pathway.

706 Building the model required two major implementation stages: first we had to create
707 receptive field (RF) patches for each layer and second, use these RF patches to compute
708 responses to our experimental images. As in the 2007 publication, the patches were imprinted
709 using experience-dependent activity.

710 Filters. Each layer contained a set of up to 200 unique filters: the V1 layer filters were
711 Gabors at four orientations and eight sizes (3-10 pixels wide, or 0.1-0.4° wide given our monitor
712 distance). Subsequent layer filters were imprinted using random samples of activity from the
713 preceding layer. To train these filters, we randomly selected images from the Caltech-256
714 database and passed them through the V1 layer: the resulting activity patterns were sampled
715 randomly to imprint filter shapes for the V2 layer²². Another set of images was passed through the
716 V1→V2 layers, and that set of activity patterns was used to imprint filter shapes for the V4 layer.
717 We did this for every layer in the long and short pathways. After repeating this process hundreds
718 of times, this resulted in a model with different filter shapes at each layer. Filter sizes doubled at
719 every hierarchical step, with the exception of the bypass pathways, where filter sizes quadrupled
720 in width at the one skip level (e.g. V4 filters in the V1→V4 pathway were four times the size of V1
721 filters; PIT filters in the V1→V2|3 pathway were four times the size of V2|3 filters). To make sure
722 that the RF shapes would match the statistics of natural images presented close to and far from
723 the fovea, we also imprinted using differently sized variants of the same images (1°-, 2°- and 4°-
724 wide versions of the same image).

725 Experimental image testing. We began with 293 different images for our experimental
726 dataset. In the electrophysiology experiments, we presented each image multiple times and
727 obtained a distribution of correlated but non-identical response vectors for each individual image.
728 To mimic this trial-by-trial variability in the model, we created six variations of all 293 images,
729 adding pixel-by-pixel noise and random changes in position, simulating differences in fixational
730 eye movements. We then transformed our 293x6 experimental images into simulated PIT
731 responses using the fully assembled network, and used SVMs to measure the classification
732 accuracy from the model PIT units, SVMs were used in an all-vs.-all approach, with leave-one-

733 out cross-validation and shuffled-label randomization control. The key theoretical contrasts were
734 the relative performances between PIT units receiving inputs from different pathways: the
735 $V1 \rightarrow V2|3 \rightarrow V4 \rightarrow \text{PIT}$ pathway, the $V1 \rightarrow V4 \rightarrow \text{PIT}$ pathway and the $V1 \rightarrow V2|3 \rightarrow \text{PIT}$ pathway. We
736 considered the long-pathway PIT units to represent our control temperature population, the latter
737 the deactivated state units.

738 **Local vs. global feature analysis.** The purpose of this analysis was to determine whether
739 closely represented images suffered the same proportional reduction in discriminability as
740 distantly represented images. Distance was determined using a Euclidean approach, measured
741 in activity space; coordinates within this activity space were defined by the activity of multiunits
742 within each pseudo-population. We used the `pdist.m` function in Matlab to obtain the distance
743 between each pair of images (in z-score-normalized activity space), using control (warm) data.
744 We then cycled through each image and rank-ordered all the other images by their distance to it:
745 the images with the smallest Euclidean distance were ranked first, those with the longest distance
746 last. We then used this ranking to sort the cooling classification accuracy values for that same
747 pair of images. Note that the accuracy values were computed using different data from the data
748 used to compute the distance ranking values. After cycling through all images, this resulted in a
749 matrix of decoding accuracy values, where each row corresponded to an image, and each column
750 corresponded to the accuracy value for its near to far neighbors. We then averaged this matrix to
751 obtain the accuracy-distance curve. We repeated the same process using the simulation data:
752 instead of using the cooling decoding accuracy values, we used the simulated cooling values.

753 Relevant data and code is available from the authors upon discussion.

754

755 **REFERENCES**

- 756 1. Distler, C., Boussaoud, D., Desimone, R. & Ungerleider, L. G. Cortical connections of
757 inferior temporal area TEO in macaque monkeys. *J. Comp. Neurol.* **334**, 125–50 (1993).
- 758 2. Serre T, Kouh, M, Cadieu C, Knoblich U, Kreiman G, P. T. MIT AI Memo 2005–
759 036/CBCL Memo 259. [ftp://publications.ai.mit.edu/ai-publications/2005/ AIM-2005-](ftp://publications.ai.mit.edu/ai-publications/2005/ AIM-2005-036.pdf)
760 [036.pdf](ftp://publications.ai.mit.edu/ai-publications/2005/036.pdf) (2005). at <<ftp://publications.ai.mit.edu/ai-publications/2005/>>
- 761 3. Markov, N. T. *et al.* Weight consistency specifies regularities of macaque cortical
762 networks. *Cereb. Cortex* **21**, 1254–72 (2011).
- 763 4. Markov, N. T. *et al.* A Weighted and Directed Interareal Connectivity Matrix for Macaque
764 Cerebral Cortex. *Cereb. Cortex* (2012). doi:10.1093/cercor/bhs270
- 765 5. Serre, T., Oliva, A. & Poggio, T. A feedforward architecture accounts for rapid
766 categorization. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 6424–9 (2007).
- 767 6. Gattass, R., Gross, C. G. & Sandell, J. H. Visual topography of V2 in the macaque. *J.*
768 *Comp. Neurol.* **201**, 519–539 (1981).
- 769 7. Gattass, R., Sousa, A. P. B. & Gross, C. G. Visuotopic organization and extent of V3 and
770 V4 of the macaque. *J. Neurosci.* **8**, 1831–1845 (1988).
- 771 8. Baldassi, C. *et al.* Shape similarity, better than semantic membership, accounts for the
772 structure of visual object representations in a population of monkey inferotemporal
773 neurons. *PLoS Comput. Biol.* **9**, e1003167 (2013).
- 774 9. Brincat, S. L. & Connor, C. E. Underlying principles of visual shape selectivity in posterior
775 inferotemporal cortex. *Nat. Neurosci.* **7**, 880–6 (2004).

- 776 10. Fukushima, K. Cognitron: a self-organizing multilayered neural network. *Biol. Cybern.* **20**,
777 121–36 (1975).
- 778 11. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional
779 architecture in the cat's visual cortex. *J. Physiol.* **160**, 106–54 (1962).
- 780 12. Kobatake, E. & Tanaka, K. Neuronal selectivities to complex object features in the ventral
781 visual pathway of the macaque cerebral cortex. *J. Neurophysiol.* **71**, 856–67 (1994).
- 782 13. Hegd , J. & Van Essen, D. C. A comparative study of shape representation in macaque
783 visual areas v2 and v4. *Cereb. Cortex* **17**, 1100–16 (2007).
- 784 14. Buffalo, E. A., Bertini, G., Ungerleider, L. G. & Desimone, R. Impaired filtering of
785 distracter stimuli by TE neurons following V4 and TEO lesions in macaques. *Cereb.*
786 *Cortex* **15**, 141–51 (2005).
- 787 15. Kiani, R., Esteky, H., Mirpour, K. & Tanaka, K. Object category structure in response
788 patterns of neuronal population in monkey inferior temporal cortex. *J. Neurophysiol.* **97**,
789 4296–309 (2007).
- 790 16. Kriegeskorte, N. *et al.* Matching categorical object representations in inferior temporal
791 cortex of man and monkey. *Neuron* **60**, 1126–41 (2008).
- 792 17. Portilla, J. & Simoncelli, E. P. A Parametric Texture Model Based on Joint Statistics of
793 Complex Wavelet Coefficients. *Int. J. Comput. Vis.* **40**, 49–70 (2000).
- 794 18. Lomber, S. G., Payne, B. R. & Horel, J. A. The cryoloop: an adaptable reversible cooling
795 deactivation method for behavioral or electrophysiological assessment of neural function.
796 *J. Neurosci Methods* **86**, 179 – 194 (1999).
- 797 19. Carrasco, A. *et al.* Influence of core auditory cortical areas on acoustically evoked activity

- 798 in contralateral primary auditory cortex. *J. Neurosci.* **33**, 776–89 (2013).
- 799 20. Kriegeskorte, N. & Kreiman, G. *Visual Population Codes: Toward a Common Multivariate*
800 *Framework for Cell Recording and Functional Imaging*. (MIT Press, 2012). at
801 <<https://books.google.com/books?id=GoI5hxBEjooC&pgis=1>>
- 802 21. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nat.*
803 *Neurosci.* **2**, 1019–25 (1999).
- 804 22. Serre, T. *et al.* A quantitative theory of immediate visual recognition. *Prog. Brain Res.*
805 **165**, 33–56 (2007).
- 806 23. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M. & Poggio, T. Robust object recognition
807 with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**, 411–26 (2007).
808
809

810 **ACKNOWLEDGEMENTS**

811 We thank Gabriel Kreiman for his helpful comments in an earlier draft. This work was
812 funded by NEI grant EY016187 (to MSL) and the Burroughs Wellcome Postdoctoral Enrichment
813 Award (to CRP). Part of this work was realized with assistance from the Core Grant for Vision
814 Research EY12196. Portions of this research were conducted on the Orchestra High
815 Performance Compute Cluster at Harvard Medical School. This NIH-supported shared facility
816 consists of thousands of processing cores and terabytes of associated storage and is partially
817 provided through grant NCRR 1S10RR028832-01. See <http://rc.hms.harvard.edu> for more
818 information. Part of this work was also conducted with support from Harvard Catalyst | The
819 Harvard Clinical and Translational Science Center (National Center for Research Resources
820 and the National Center for Advancing Translational Sciences, National Institutes of Health
821 Award UL1 TR001102) and financial contributions from Harvard University and its affiliated
822 academic healthcare centers. The content is solely the responsibility of the authors and does
823 not necessarily represent the official views of Harvard Catalyst, Harvard University and its
824 affiliated academic healthcare centers, or the National Institutes of Health.

825 **AUTHOR CONTRIBUTIONS**

826 CRP and MSL designed the experiments. CRP, SGL and MSL performed the surgical
827 implantations and intraoperative cooling experiments. CRP conducted all following experiments
828 and analyzed the data. CRP wrote the initial manuscript which was then edited by MSL and also
829 approved by SGL.

830 **COMPETING FINANCIAL INTERESTS**

831 The authors have no competing financial interests.

832