Intrinsic and Extrinsic Effects on Image Memorability
Supplemental Material

Zoya Bylinskii[a,b], Phillip Isola[b,c], Constance Bainbridge[b], Antonio Torralba[a,b], Aude Oliva[b]

[a] *Department of Electrical Engineering and Computer Science, MIT, Cambridge, USA 02141*
[b] *Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, USA 02141*
[c] *Department of Brain and Cognitive Sciences, MIT, Cambridge, USA 02141*

---

1. Memorability Measurements

Here we include a superset of the results presented in section 5 of the main paper. We present multiple memorability measurements, defined as follows:

$$\begin{aligned}
\text{HR(I)} &= \frac{\text{hits(I)}}{\text{hits(I)} + \text{misses(I)}} \times 100\% \\
\text{FAR(I)} &= \frac{\text{false alarms(I)}}{\text{false alarms(I)} + \text{correct rejections(I)}} \times 100\% \\
\text{ACC(I)} &= \frac{\text{hits(I)} + \text{correct rejections(I)}}{\text{total(I)}} \times 100\% \\
\text{DPRIME(I)} &= Z(\text{HR}) - Z(\text{FAR})
\end{aligned}$$

where Z is the inverse of the cumulative Gaussian distribution and:

$$\text{total} = \text{hits(I)} + \text{misses(I)} + \text{false alarms(I)} + \text{correct rejections(I)}$$

Additionally, given the following 2×2 matrix:

$\frac{\text{hits(I)}}{\text{total(I)}}$	$\frac{\text{misses(I)}}{\text{total(I)}}$
$\frac{\text{false alarms(I)}}{\text{total(I)}}$	$\frac{\text{correct rejections(I)}}{\text{total(I)}}$

Mutual information (between a response and whether an image was a repeat) is calculated as:

$$\text{MI(I)} = \sum_i \sum_j p(i,j) \log \frac{p(i,j)}{p(i)p(j)}$$

where i and j index into the matrix above.

Is rank maintained?

Suppose image A is more memorable than image B. Can a change in context reverse the order, making image B more memorable than A? To test this possibility we correlated change in contextual distinctiveness with change in rank memorability. In the rightmost column of Table 4, we show that context changes can indeed change which images are most memorable. However, note that the correlations are weaker for change in rank than for change in metric scores, like HR. This indicates that rank memorability is relatively stable to context changes, while raw recognition rates are affected more.

2. Features

In the following tables (1-5), *scene-based CNN features* come from the Places-CNN from [1] trained to classify scene categories (these are the features for which results are reported in the main paper). The *object-based CNN features* come from a pre-trained model from Caffe[1] tuned to perform object classification [2]. From both CNN models, we took the 4096-dimensional features from the response of the Fully Connected Layer 7 ($fc7$) of the CNNs, which is the final fully connected layer before producing the class predictions. The Gist features, as defined in [3], are calculated using the *LabelMe Toolbox*[2][4].

3. Measuring Context Effects

Figure 1 plots the same correlations as Figure 4 in the main paper, but with the Gist features instead of the scene-based CNN features. Corresponding to this figure are the tables presented on the next few pages. Tables 1 and 2 demonstrate that contextually distinct images are more memorable (see section 5.1 in

[1] available at `http://caffe.berkeleyvision.org`

[2] available at `http://labelme.csail.mit.edu/Release3.0/browserTools/php/matlab_toolbox.php`

the main paper), and this holds within and across categories, across memorability measurements, and feature types. Table 3 demonstrates that more varied image contexts are more memorable overall (section 5.2 in the main paper). Tables 4 and 5 show that changing image context can predictably change image memorability, at the image and category levels, correspondingly (section 5.3 in the main paper). Figures 2-8 visually depict which images are memorable across contexts, and which are most affected by context (section 5.4 in the main paper). Figures 9-15 demonstrate that images that are most affected by context are more likely to look like other scene categories (section 5.4 in the main paper contains a discussion). In the tables that follow we define HS to stand for highly-significant correlation ($p < 0.01$), S for a significant correlation ($p < 0.05$), while NS not significant at the $p = 0.05$ level.

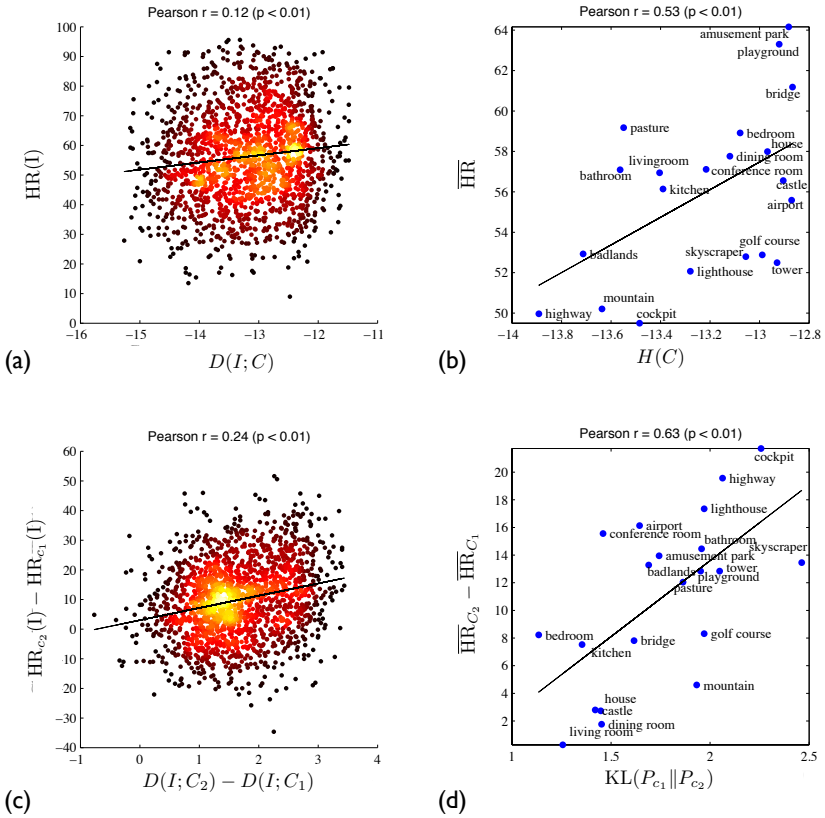

Figure 1: This figure includes the same set of plots as in the main paper, with the difference that kernel densities are estimated over Gist features instead of deep features. We see that the same trends hold: (a) Images are more memorable if they are less likely (more contextually distinct) relative to the other images in the same image context; (b) Image contexts that are more varied (have larger entropy) lead to higher memorability rates overall; (c) Images that become more distinct relative to a new context become more memorable; (d) Scene categories that are more distinct relative to other categories become more memorable in the context of those other categories.

feature space	HR	FAR	ACC	MI	DPRIME
gist features	0.12 (HS)	-0.19 (HS)	0.21 (HS)	0.21 (HS)	0.19 (HS)
object-based CNN features	0.19 (HS)	-0.14 (HS)	0.22 (HS)	0.22 (HS)	0.20 (HS)
scene-based CNN features	0.26 (HS)	-0.26 (HS)	0.35 (HS)	0.36 (HS)	0.34 (HS)

Table 1: $D(I;C)$ of images (C = AMT 1) correlated with different memorability measurements.

feature space	HR	FAR	ACC	MI	DPRIME
gist features	0.12 (HS)	-0.01 (NS)	0.12 (HS)	0.11 (HS)	0.11 (HS)
object-based CNN features	0.13 (HS)	-0.05 (S)	0.15 (HS)	0.12 (HS)	0.12 (HS)
scene-based CNN features	0.24 (HS)	-0.17 (HS)	0.32 (HS)	0.33 (HS)	0.32 (HS)

Table 2: $D(I;C)$ of images (C = AMT 2) correlated with different memorability measurements.

feature space	HR	FAR	ACC	MI	DPRIME
gist features	0.48 (S)	-0.23 (NS)	0.34 (NS)	0.35 (NS)	0.36 (NS)
object-based CNN features	0.51 (S)	-0.29 (NS)	0.34 (NS)	0.38 (NS)	0.32 (NS)
scene-based CNN features	0.52 (S)	-0.16 (NS)	0.28 (NS)	0.37 (NS)	0.34 (NS)

Table 3: $H(C)$ of images in AMT 1 (within-category) correlated with different memorability measurements.

feature space	HR	FAR	ACC	MI	DPRIME	RANK
gist features	0.24 (HS)	0.10 (HS)	0.15 (HS)	0.10 (HS)	0.12 (HS)	0.10 (HS)
object-based CNN features	0.32 (HS)	0.00 (NS)	0.25 (HS)	0.21 (HS)	0.24 (HS)	0.15 (HS)
scene-based CNN features	0.35 (HS)	0.00 (NS)	0.29 (HS)	0.25 (HS)	0.28 (HS)	0.14 (HS)

Table 4: Change in *contextual distinctiveness* between AMT 1 and AMT 2 correlated with different memorability measurements.

feature space	HR	FAR	ACC	MI	DPRIME
gist features	0.64 (HS)	0.41 (NS)	0.47 (S)	0.39 (NS)	0.41 (NS)
object-based CNN features	0.55 (S)	0.24 (NS)	0.35 (NS)	0.46 (S)	0.47 (S)
scene-based CNN features	0.74 (HS)	0.44 (S)	0.51 (S)	0.50 (S)	0.60 (HS)

Table 5: Change in *context entropy* between AMT 1 and AMT 2 correlated with different memorability measurements.

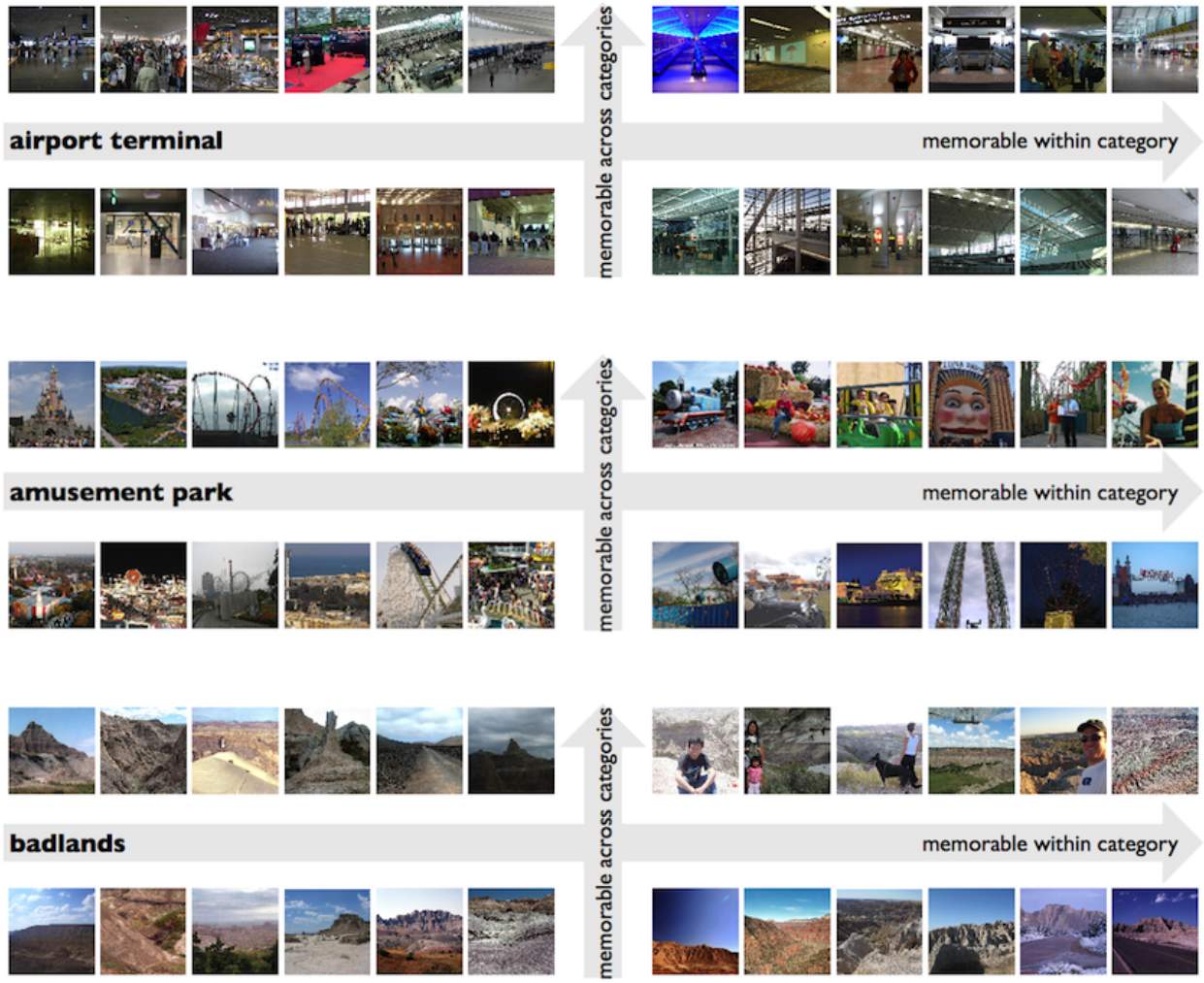

Figure 2: Memorability scores of images in the top right quadrant of each plot are least affected by context whereas the scores of images in the bottom right quadrant are most affected by context. Images in the top right are distinct with respect to both contexts, while images in the bottom right are distinct only with respect to their own category.

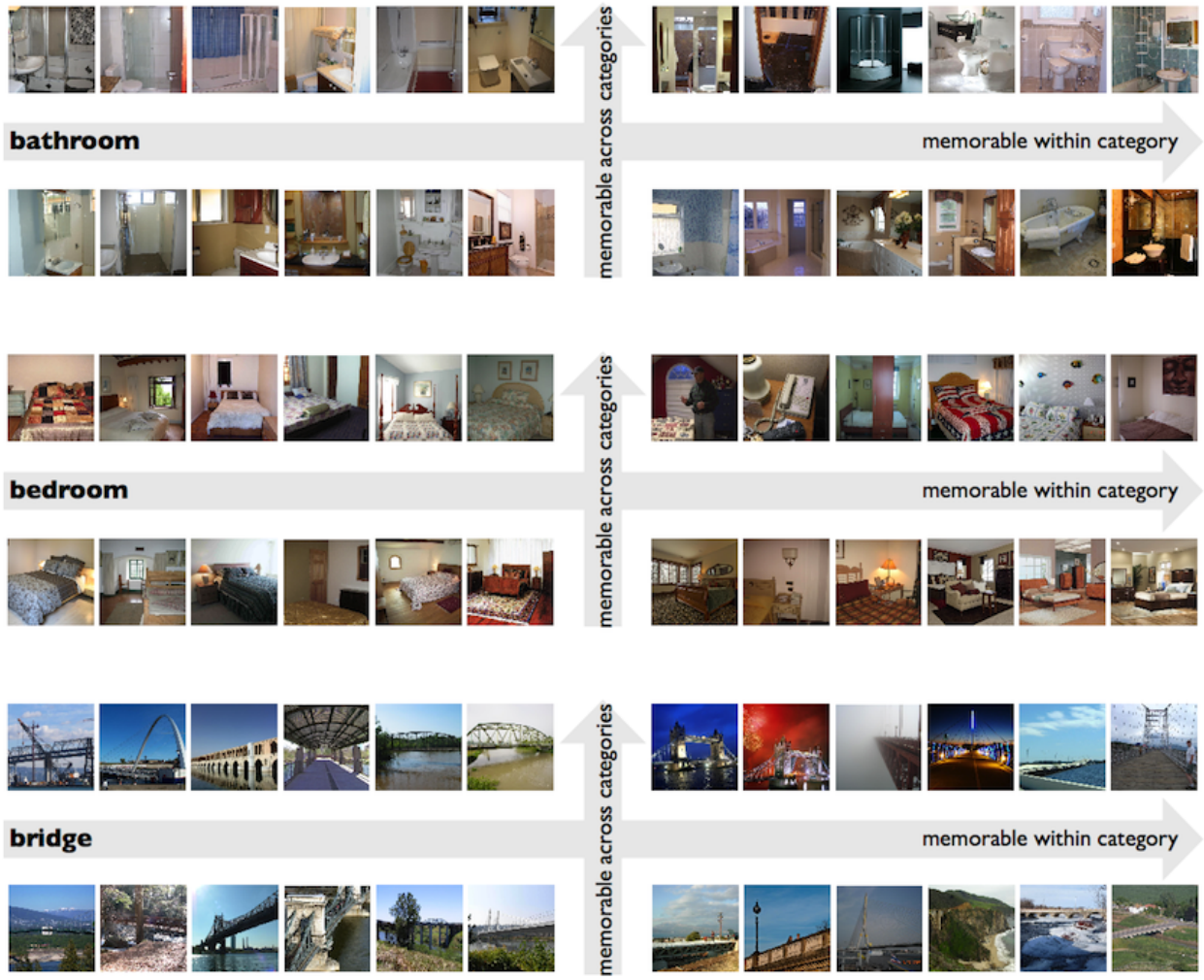

Figure 3: Continued from fig.2.

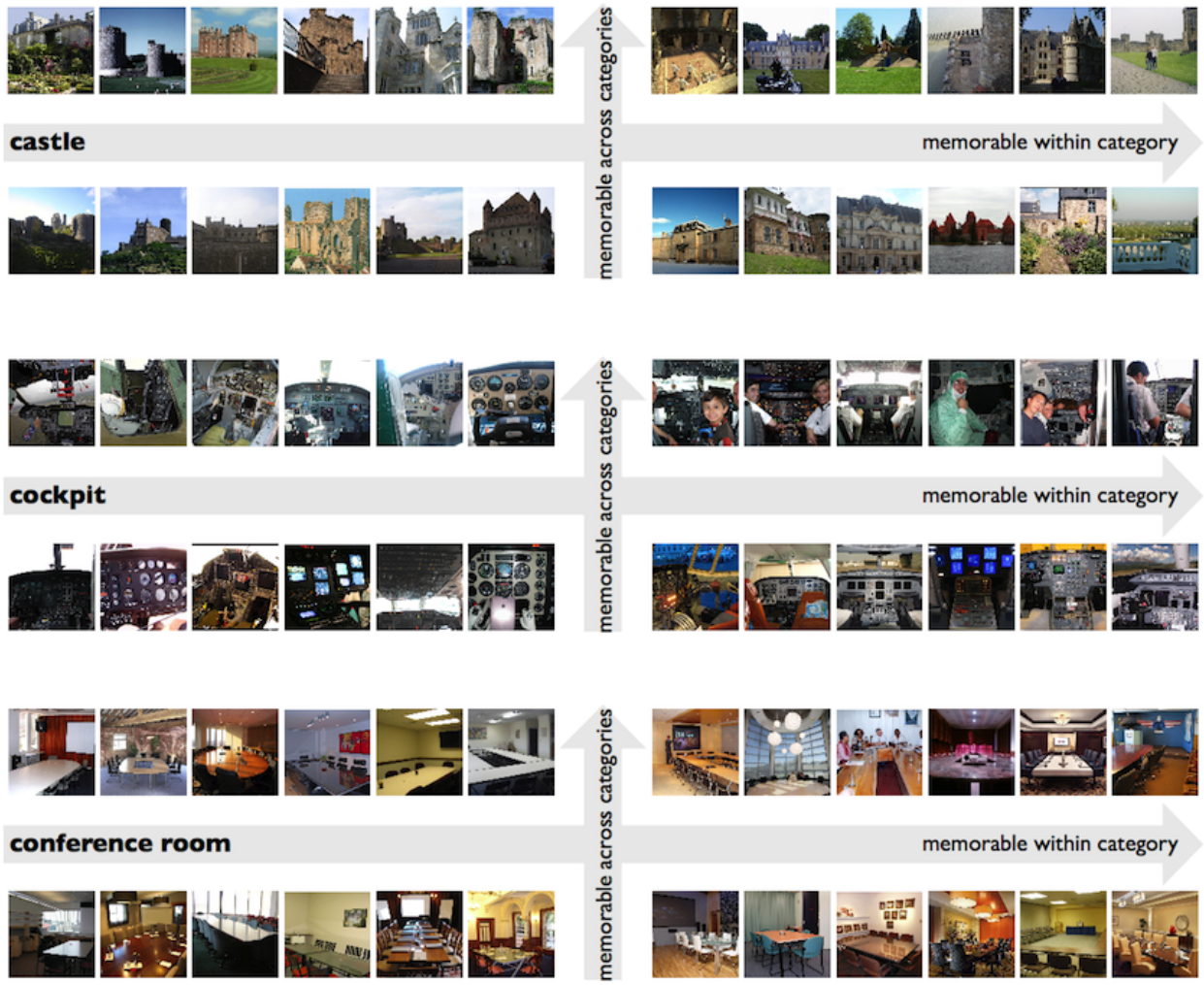

Figure 4: Continued from fig.2.

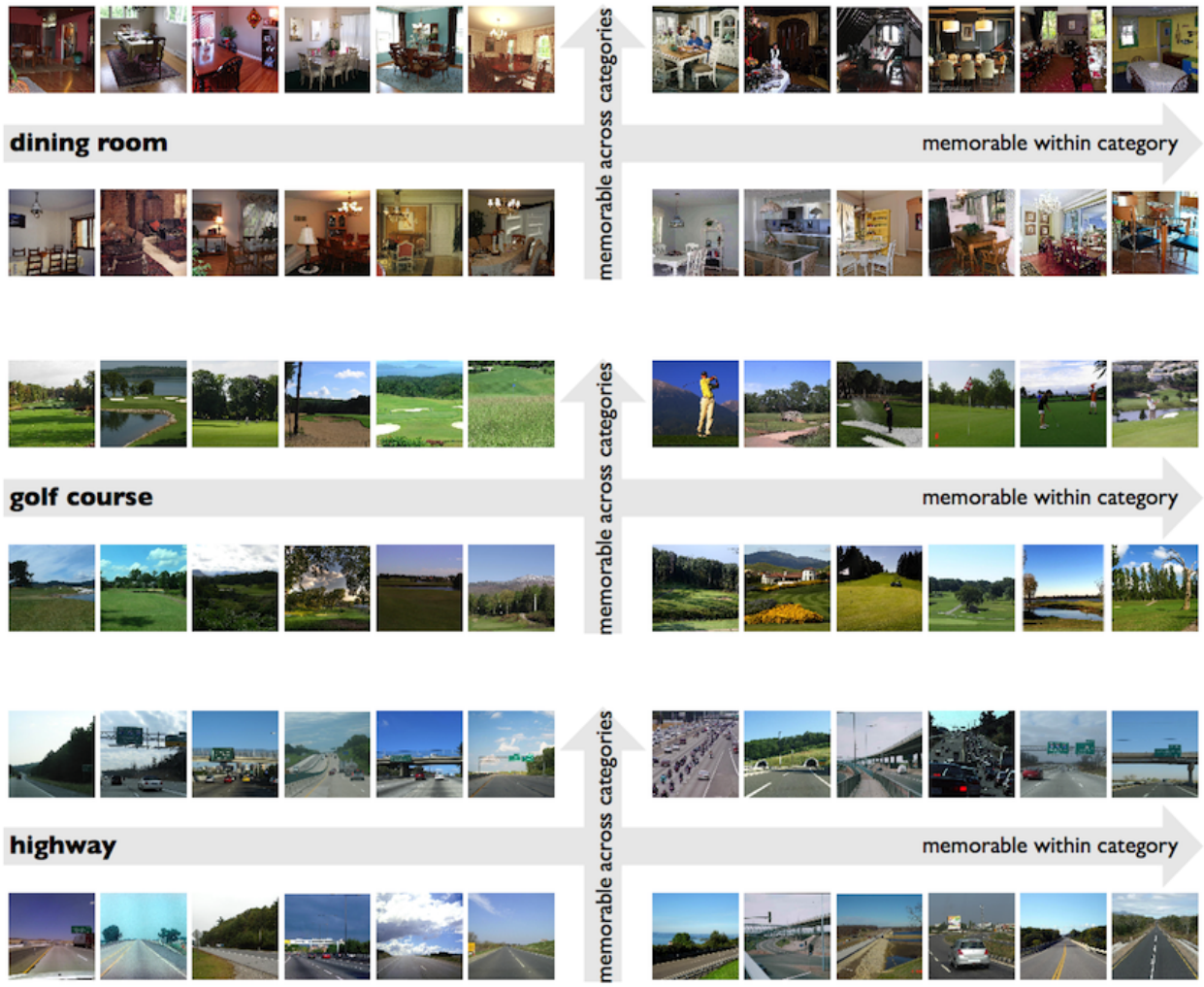

Figure 5: Continued from fig.2.

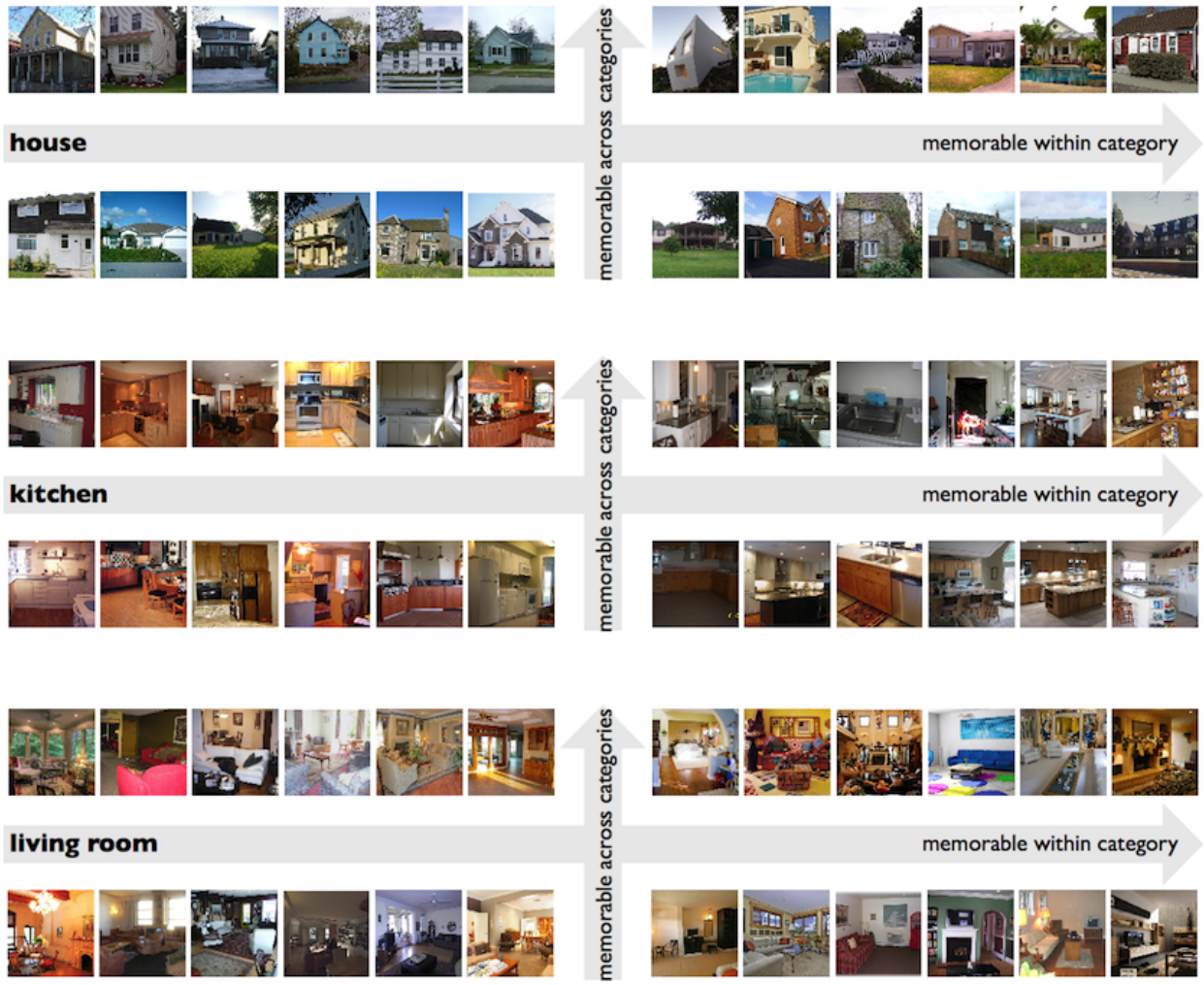

Figure 6: Continued from fig.2.

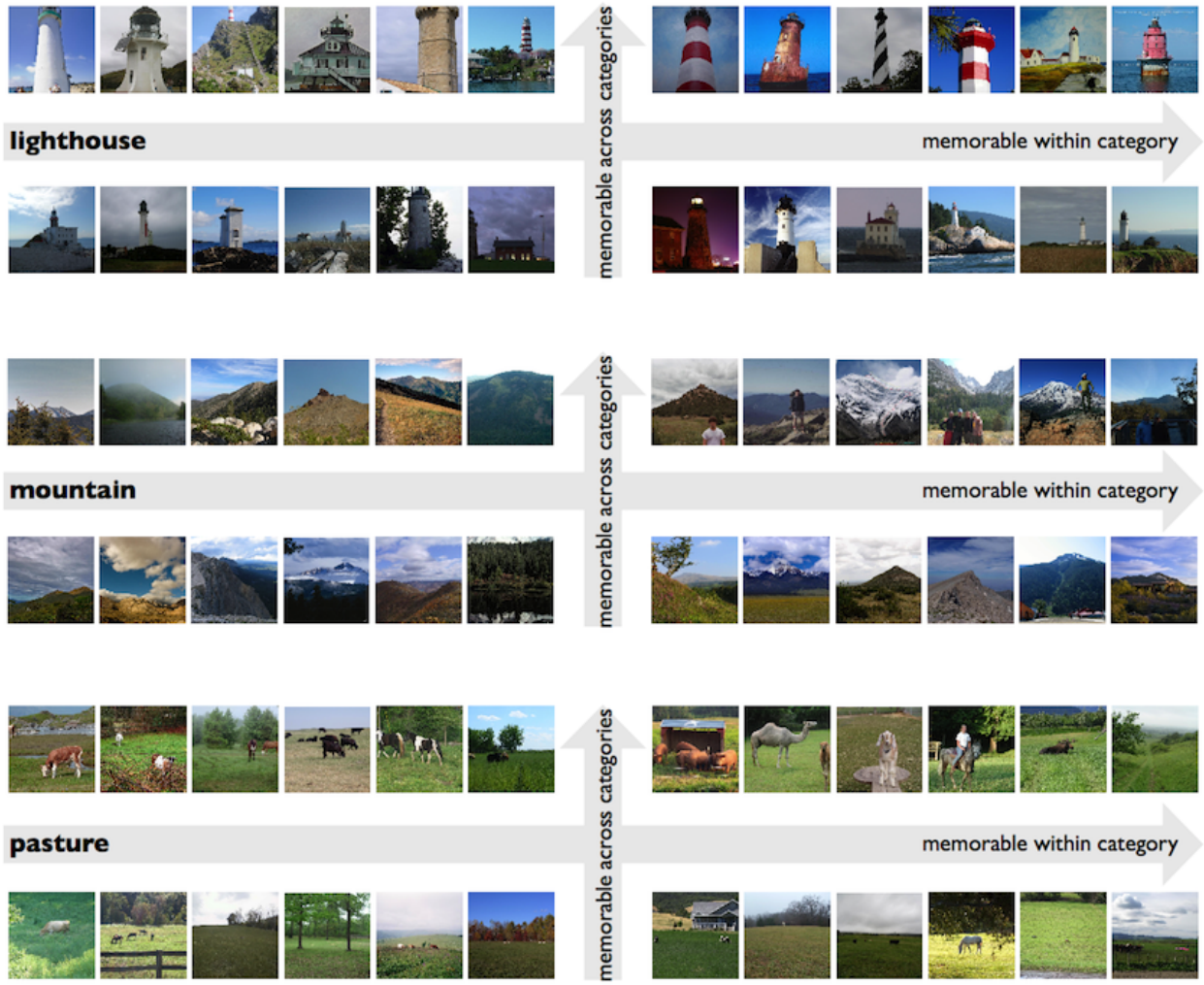

Figure 7: Continued from fig.2.

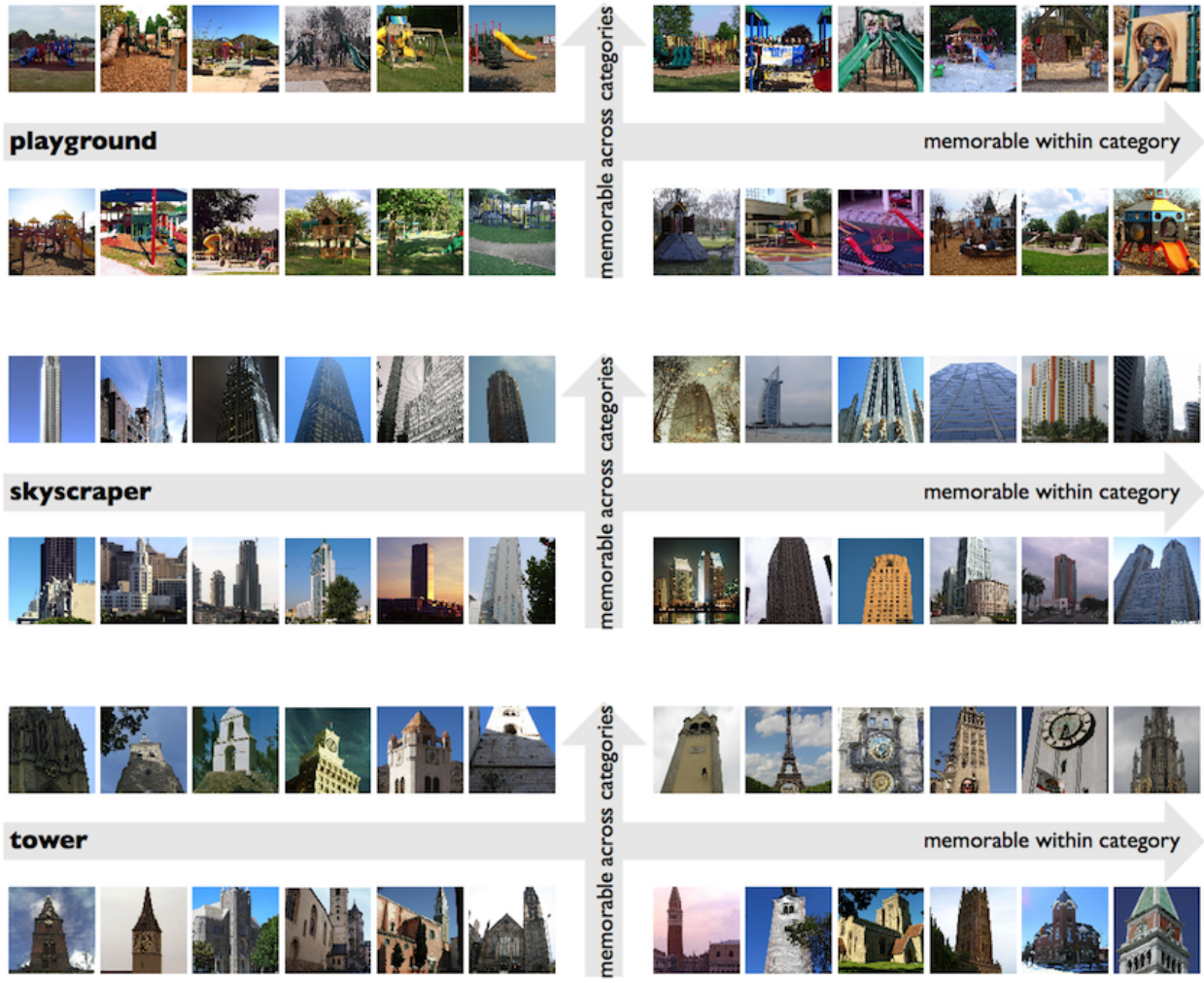

Figure 8: Continued from fig.2.

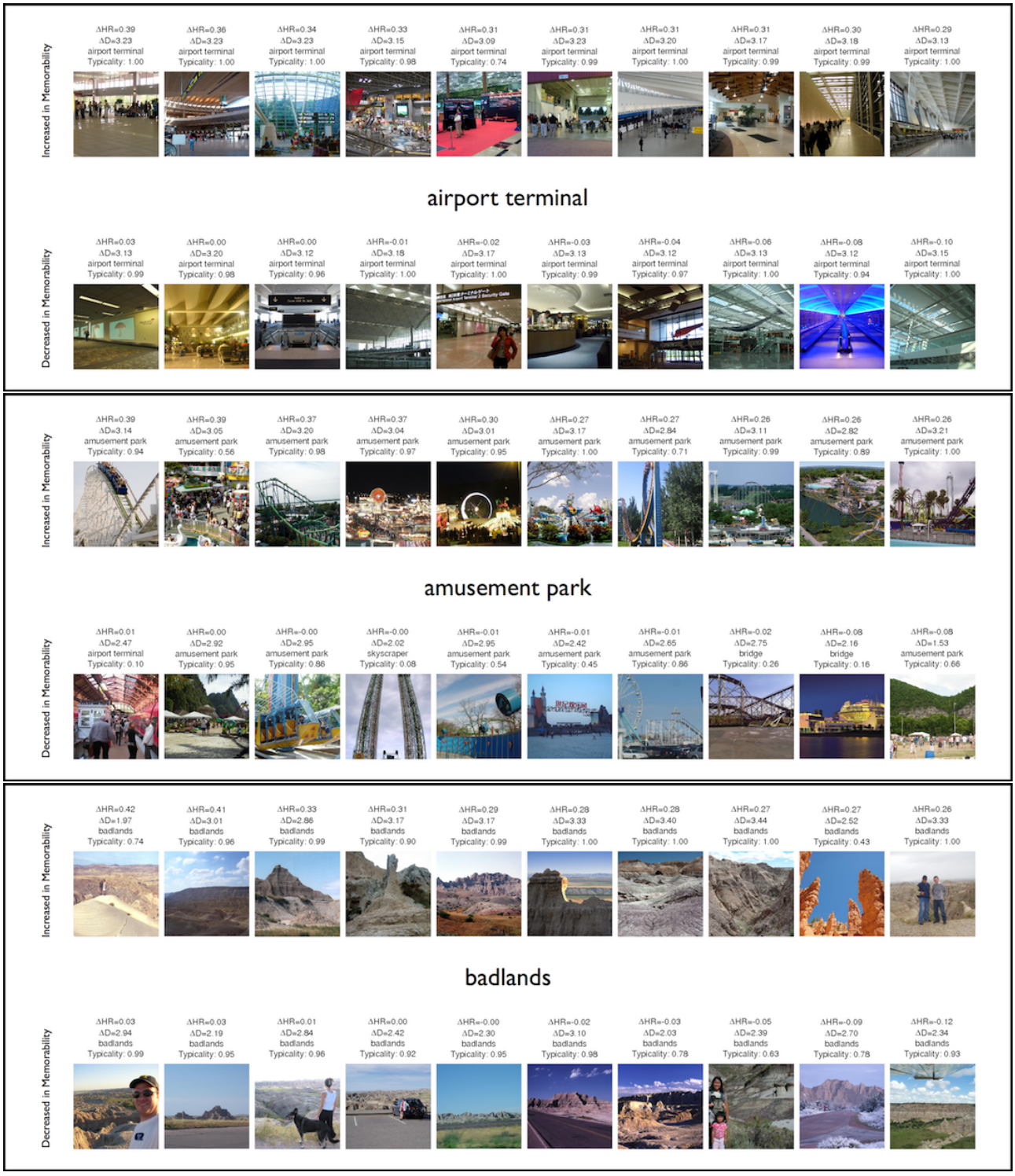

Figure 9: We evaluated a scene classifier on the images in that increased most and those that dropped most in memorability when combined with other categories. For each image, we provide the classifier?s predicted category label, as well as the probability of the correct category label (where * is replaced with the correct category). Images that drop in memorability are more likely to be confused with other categories.

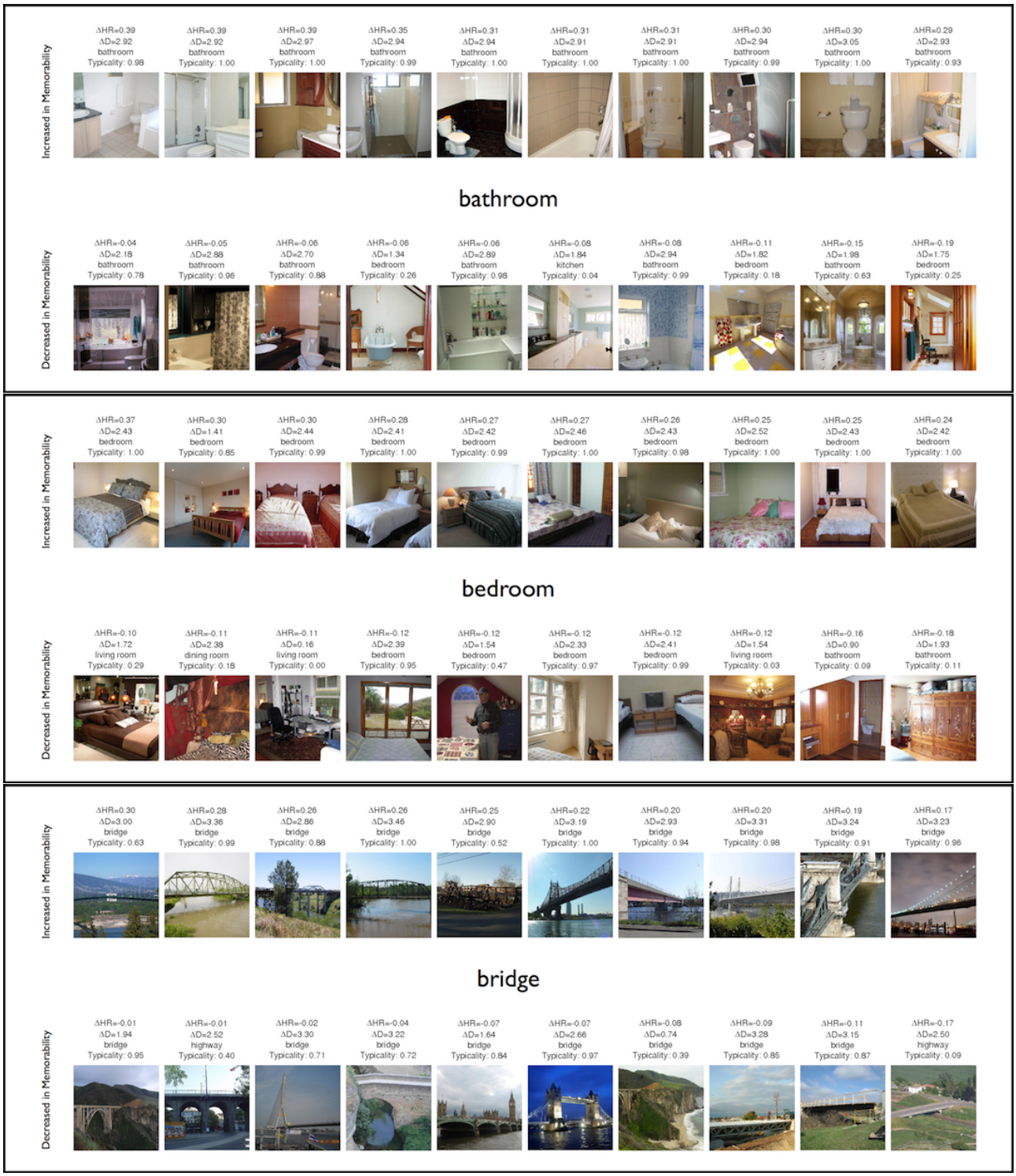

Figure 10: Continued from fig.9.

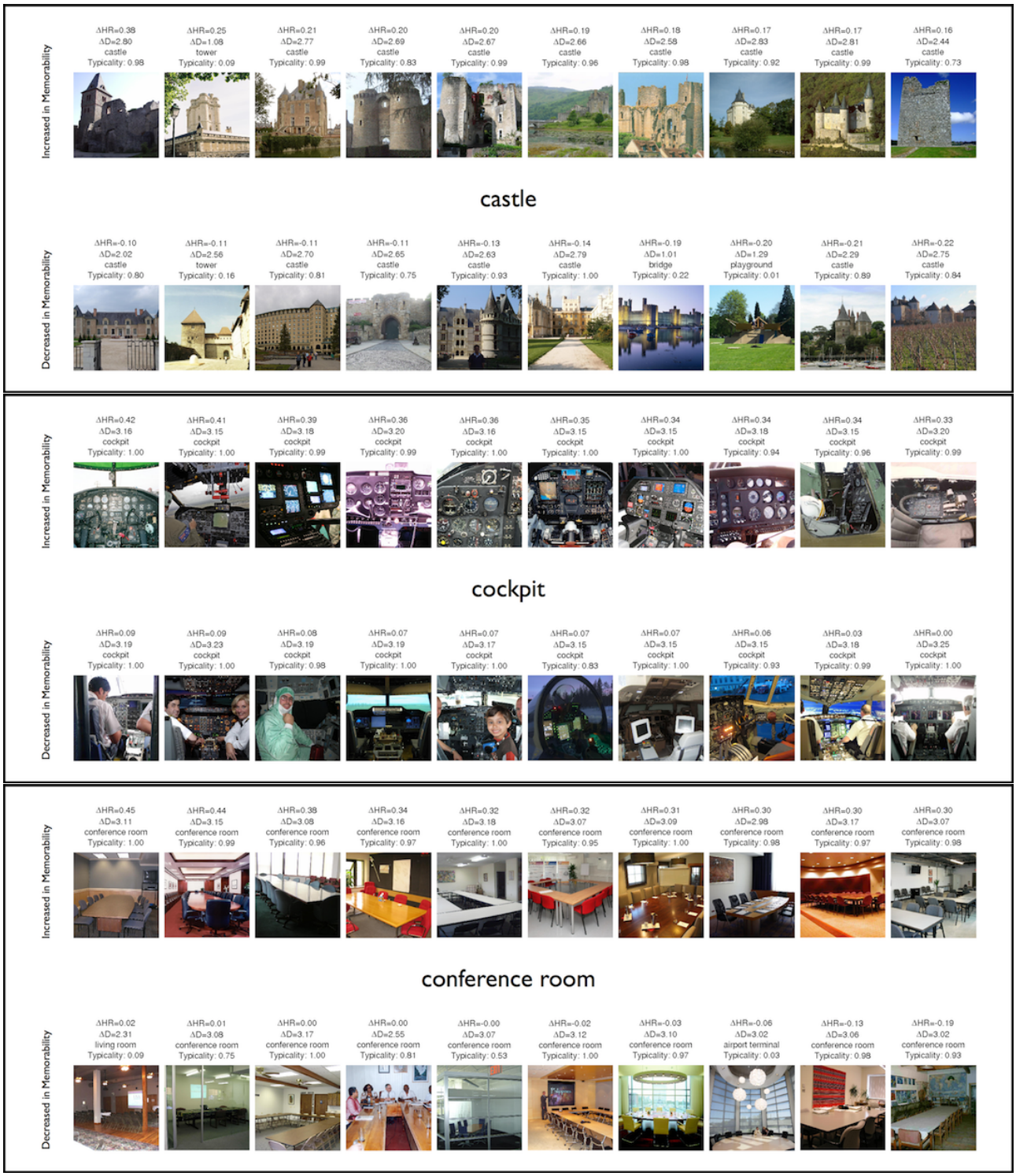

Figure 11: Continued from fig.9.

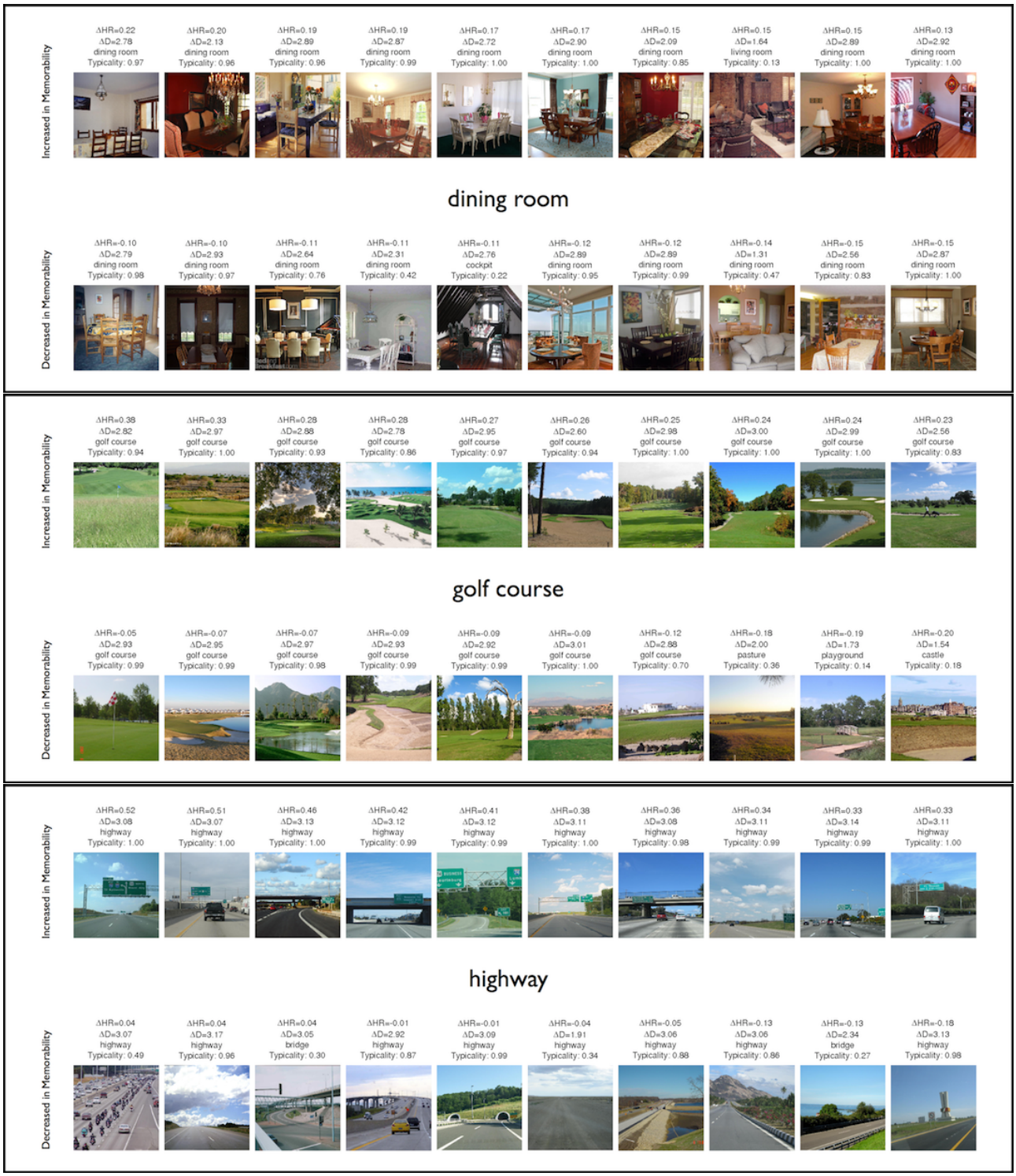

Figure 12: Continued from fig.9.

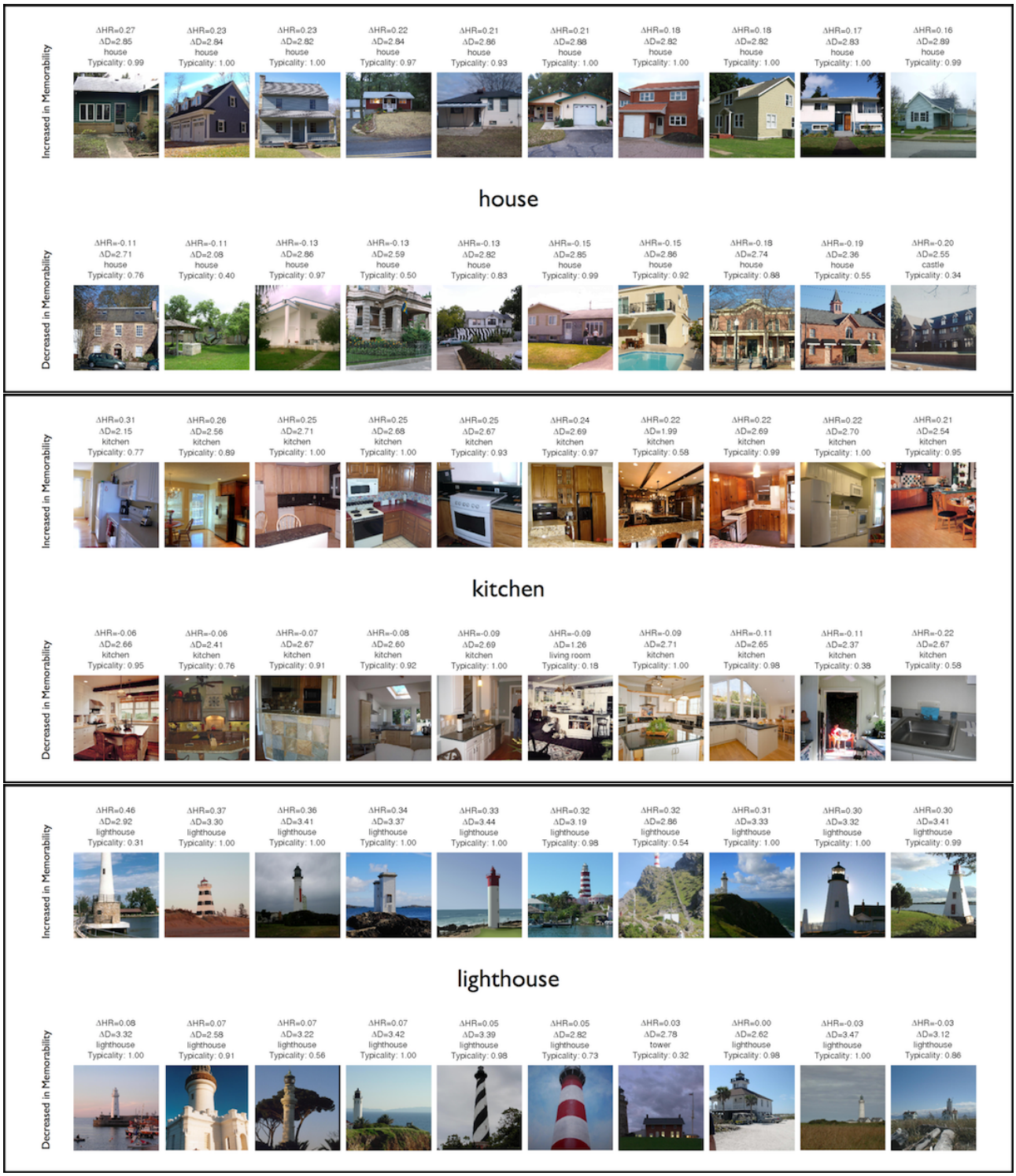

Figure 13: Continued from fig.9.

Figure 14: Continued from fig.9.

Figure 15: Continued from fig.9.

References

[1] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning Deep Features for Scene Recognition using Places Database, in: Conference on Neural Information Processing Systems (NIPS), 2014.

[2] Y. Jia, Caffe: An Open Source Convolutional Architecture for Fast Feature Embedding, `http://caffe.berkeleyvision.org/` (2013).

[3] A. Oliva, A. Torralba, Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope, International Journal of Computer Vision (IJCV) 42 (3) (2001) 145–175.

[4] B. C. Russell, A. Torralba, K. P. Murphy, W. T. Freeman, LabelMe: a database and web-based tool for image annotation, International Journal of Computer Vision (IJCV) 77 (1–3) (2008) 157–173.