RESEARCH ARTICLE

Figures and figure supplements

Emergence of brain-like mirror-symmetric viewpoint tuning in convolutional neural networks

Amirhossein Farzmahdi *et al.*

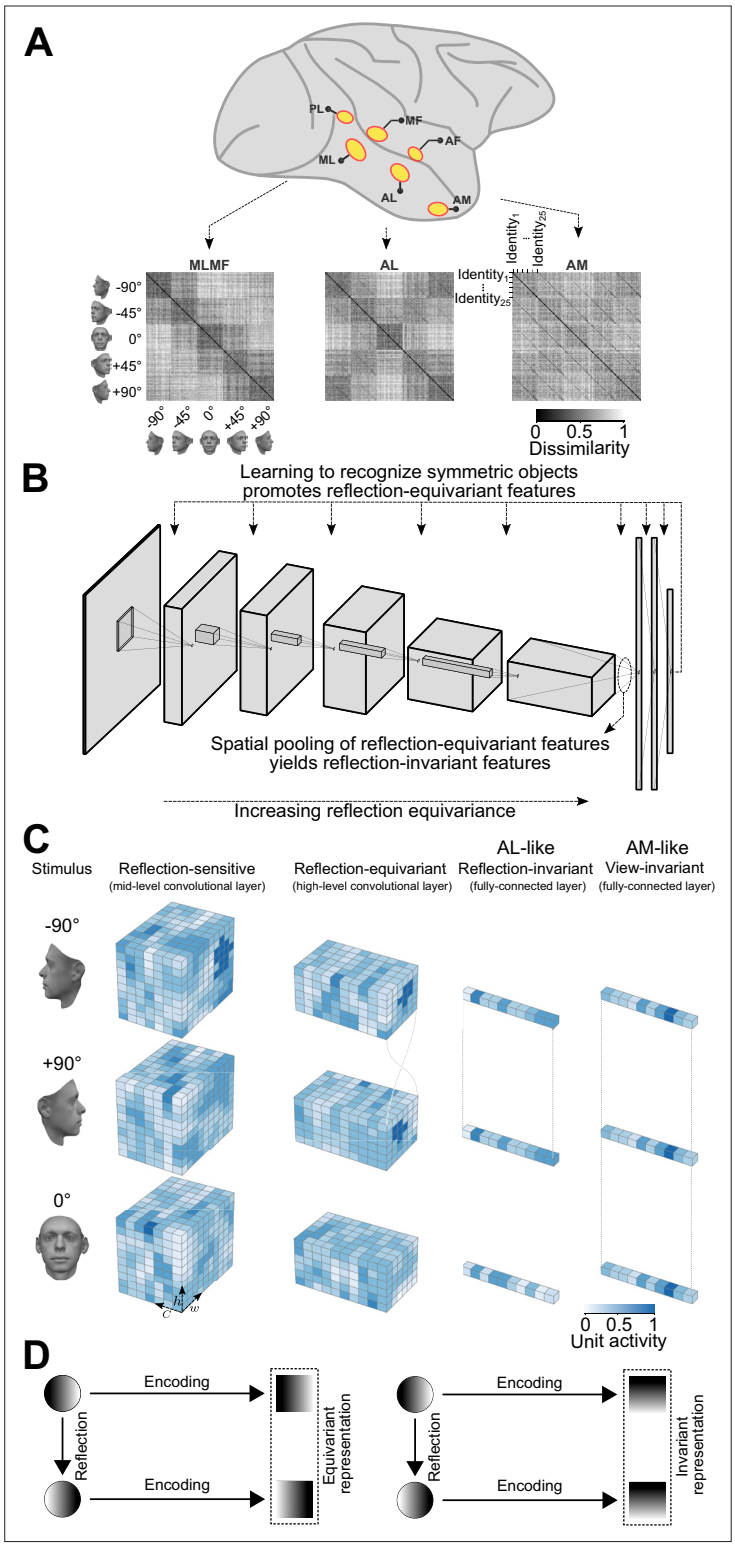

Figure 1. An overview of our claim: convolutional deep neural networks trained on discriminating among bilaterally symmetric object categories provide a parsimonious explanation for the mirror-symmetric viewpoint tuning of the macaque AL face-patch. (**A**) The macaque face-patch system. Face-selective cortical areas are highlighted in yellow. The areas ML, AL, and AM exhibit substantially different tuning proprieties when presented with faces of different head orientations (***Freiwald and Tsao, 2010***). These distinct tuning profiles are evident in population-level representational dissimilarity matrices (RDMs). From posterior to anterior face

Figure 1 continued on next page

Figure 1 continued

areas, invariance to viewpoints gradually increases: from view-tuned in ML, through mirror-symmetric in AL, to view-invariant identity selectivity in AM (neural data from ***Freiwald and Tsao, 2010***). (**B**) Training convolutional deep neural networks on recognizing specific symmetric object categories (e.g. faces, cars, the digit 8) gives rise to AL-like mirror-symmetric tuning. It is due to a cascade of two effects: First, learning to discriminate among symmetric object categories promotes tuning for reflection-equivariant representations throughout the entire processing layers. This reflection equivariance increases with depth. Then, long-range spatial pooling (as in the transformation of the last convolution layer to the first fully connected layer in CNNs) transforms the equivariant representations into reflection-invariant representations. (**C**) Schematic representations of three viewpoints of a face (left profile, frontal view, right profile) are shown in three distinct stages of processing. Each tensor depicts the width (**w**), height (**h**), and depth (**c**) of an activation pattern. Colors indicate channel activity. From left to right: In a mid-level convolutional layer, representations are view-specific. A deeper convolutional layer produces reflection-equivariant representations that are view-specific. Feature vectors of a fully connected layer become invariant to reflection by pooling reflection-equivariant representations from the last convolutional layer. (**D**) A graphical comparison of reflection-equivariance and reflection-invariance. Circles denote input images, and squares denote representations.

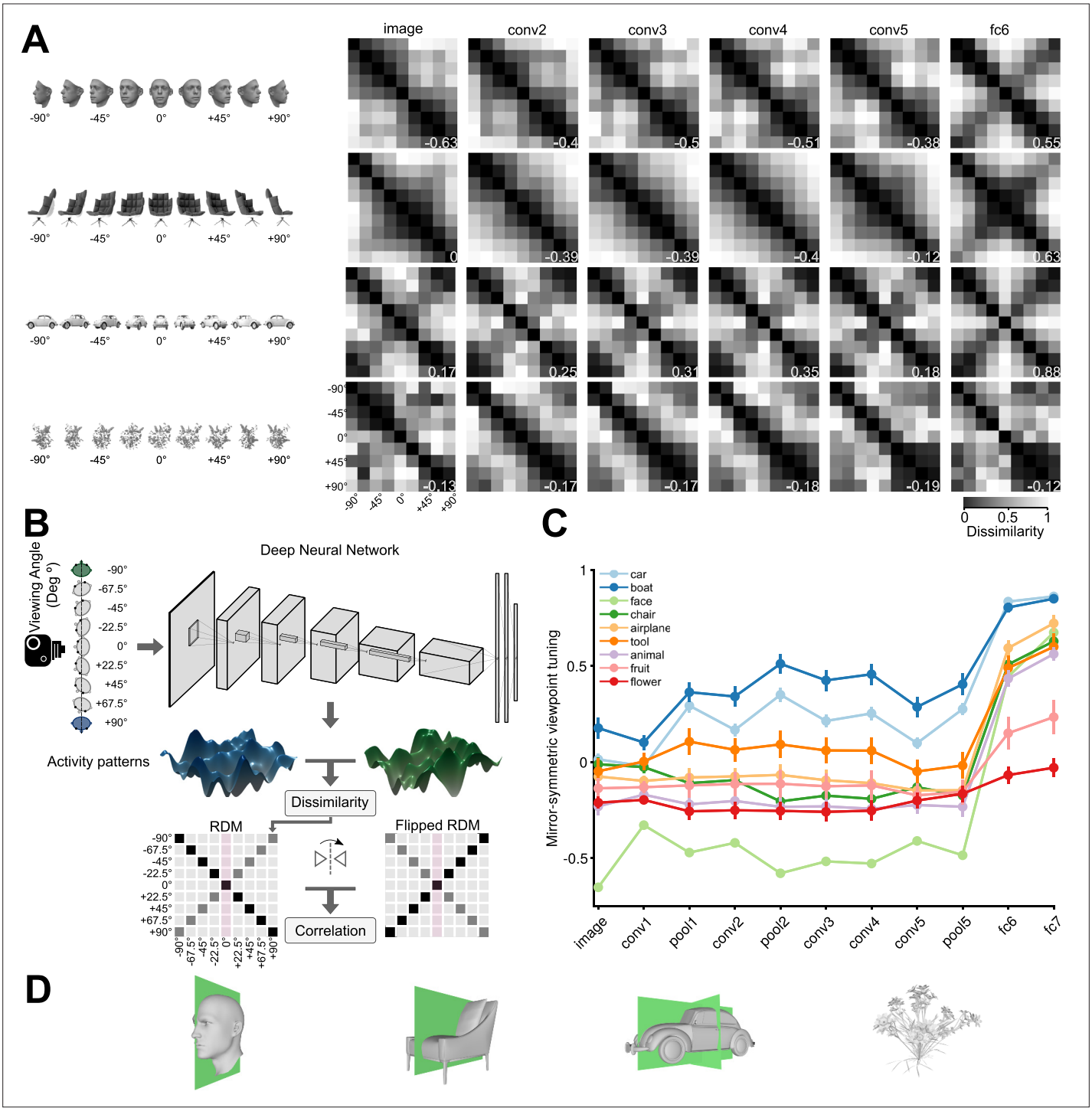

Figure 2. Mirror-symmetric viewpoint tuning of higher level deep neural network representations emerges for multiple object categories. (**A**) Different viewpoint tuning across the layers of AlexNet for four example objects. For each object, the responses to nine views (–90° to +90° in the steps of 22.5°) were measured in six key AlexNet layers, shallow (input, *left*) to deep (fc6, *right*). For each layer, a Representational Dissimilarity Matrix (RDM) depicts how the population activity vector varies across different object views. Each element of the RDM represents the dissimilarity (1 - Pearson correlation coefficient) between a pair of activity vectors evoked in response to two particular views. The symmetry of the RDMs about the major diagonal is inherent to their construction. However, the symmetry about the minor diagonal (for the face and chair, in fc6, and for the car, already in conv2) indicates mirror-symmetric viewpoint tuning. (**B**) The schematic shows how the mirror-symmetric viewpoint tuning index was quantified. We first fed the network with images of each object from nine viewpoints and recorded the activity patterns of its layers. Then, we computed the dissimilarity between activity patterns of different viewpoints to create an RDM. Next, we measured the correlation between the obtained RDM and its horizontally

Figure 2 continued on next page

Figure 2 continued

flipped counterpart, excluding the frontal view (which is unaffected by the reflection). (**C**) The Mirror-symmetric viewpoint tuning index across all AlexNet layers for nine object categories (car, boat, face, chair, airplane, animal, tool, fruit, and flower). Each solid circle denotes the average of the index over 25 exemplars within each object category. Error bars indicate the standard error of the mean. The mirror-symmetric viewpoint tuning index values of the four example objects in panel B are shown at the bottom right of each RDM in panel B. ***Figure 2—figure supplement 4*** shows the same analysis applied to representations of the face stimulus set used in ***Freiwald and Tsao, 2010***, across various neural network models. (**D**) 3D Objects have different numbers of symmetry axes. A face (left column), a non-face object with bilateral symmetry (a chair, second column), an object with quadrilateral symmetry (a car, third column), and an object with no obvious reflective symmetry planes (a flower, right column).

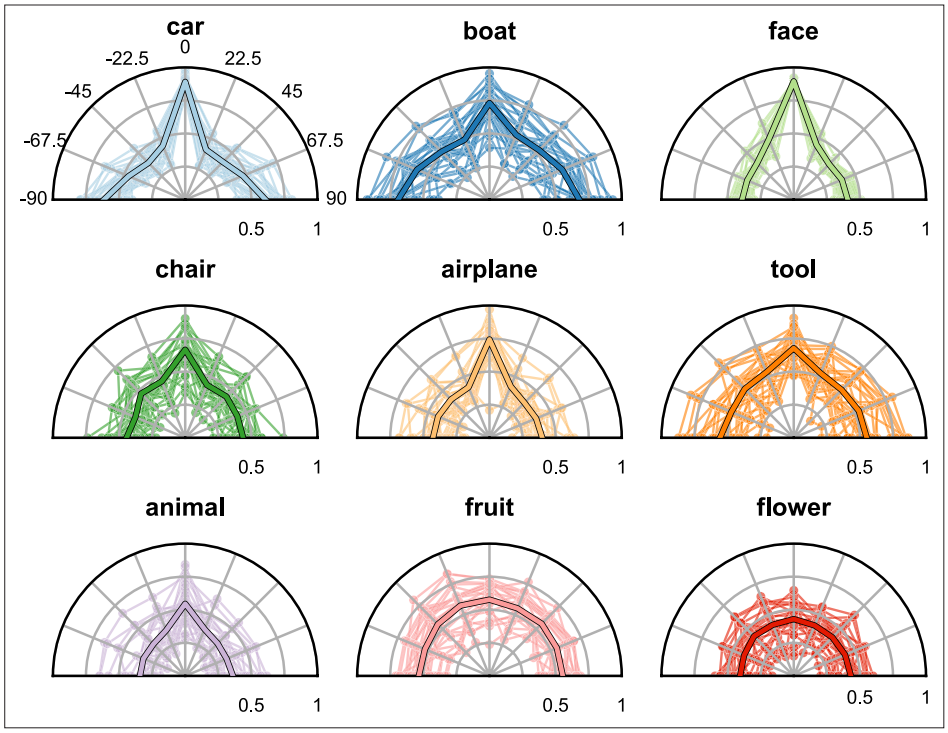

Figure 2—figure supplement 1. Assessment of symmetry planes in 3D renders across viewpoints. For each 3D object (25 exemplars for each of the nine categories) and each rendering viewpoint (nine viewpoints from –90° to 90° at 22.5° intervals) used in the stimulus set, we measured the horizontal symmetry of the resulting 2D render by correlating the left half of the 2D image with a flipped version of its right half. In each such measurement, we systematically shifted the plane of reflection and used the highest correlation across all shifts. The resulting correlation coefficients, representing horizontal symmetry as a function of viewpoint, are displayed on polar plots. In these plots, each depicting a single object category, thin lines indicate individual object exemplars (e.g. a particular face), and bold lines indicate the average correlation coefficients across the 25 exemplars of each category. By setting a threshold at half a standard deviation above the mean correlation, we heuristically counted the number of symmetry axes for each object category. Notably, images of cars and boats have strong image-space symmetry in both frontal and side views, explaining the pronounced mirror-symmetric viewpoint tuning index observed already in early convolutional layers. These two categories exhibit dual symmetry axes—left–right and front–back. In comparison, objects like faces, chairs, airplanes, tools, and animals have a single left-right symmetry plane, expressed in the 2D renders as high horizontal symmetry of the frontal view. Fruits and flowers have relatively uniform correlation values across views, which is indicative of radial symmetry. This radial symmetry translates to a lower mirror-symmetric viewpoint tuning index of the neural network representations of these categories.

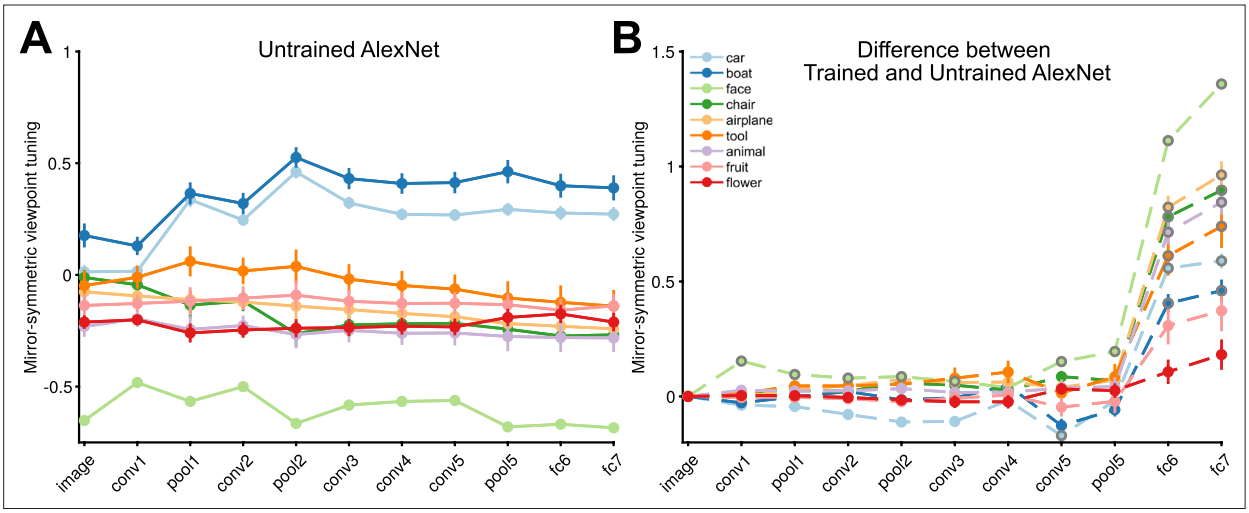

Figure 2—figure supplement 2. The mirror-symmetric viewpoint tuning index remains unchanged as the signal moves into the fully connected layers of the untrained network. (**A**) Each solid circle represents the average index for 25 exemplars within each object category (car, boat, face, chair, airplane, animal, tool, fruit, flower) for the untrained AlexNet network. (**B**) Each solid circle refers to the difference between the mirror-symmetric viewpoint tuning index of the trained versus the untrained AlexNet network. We evaluated the difference using the rank-sum test. We used the ***Benjamini and Hochberg, 1995*** procedure for controlling the False discovery rate (FDR) across 90 comparisons at $q<0.05$ (9 categories and 10 layers, excluding the input layer, as it is the same in both networks). The solid circles with gray outlines indicate where the difference after FDR adjustment is significant. Error bars indicate the standard error of the mean.

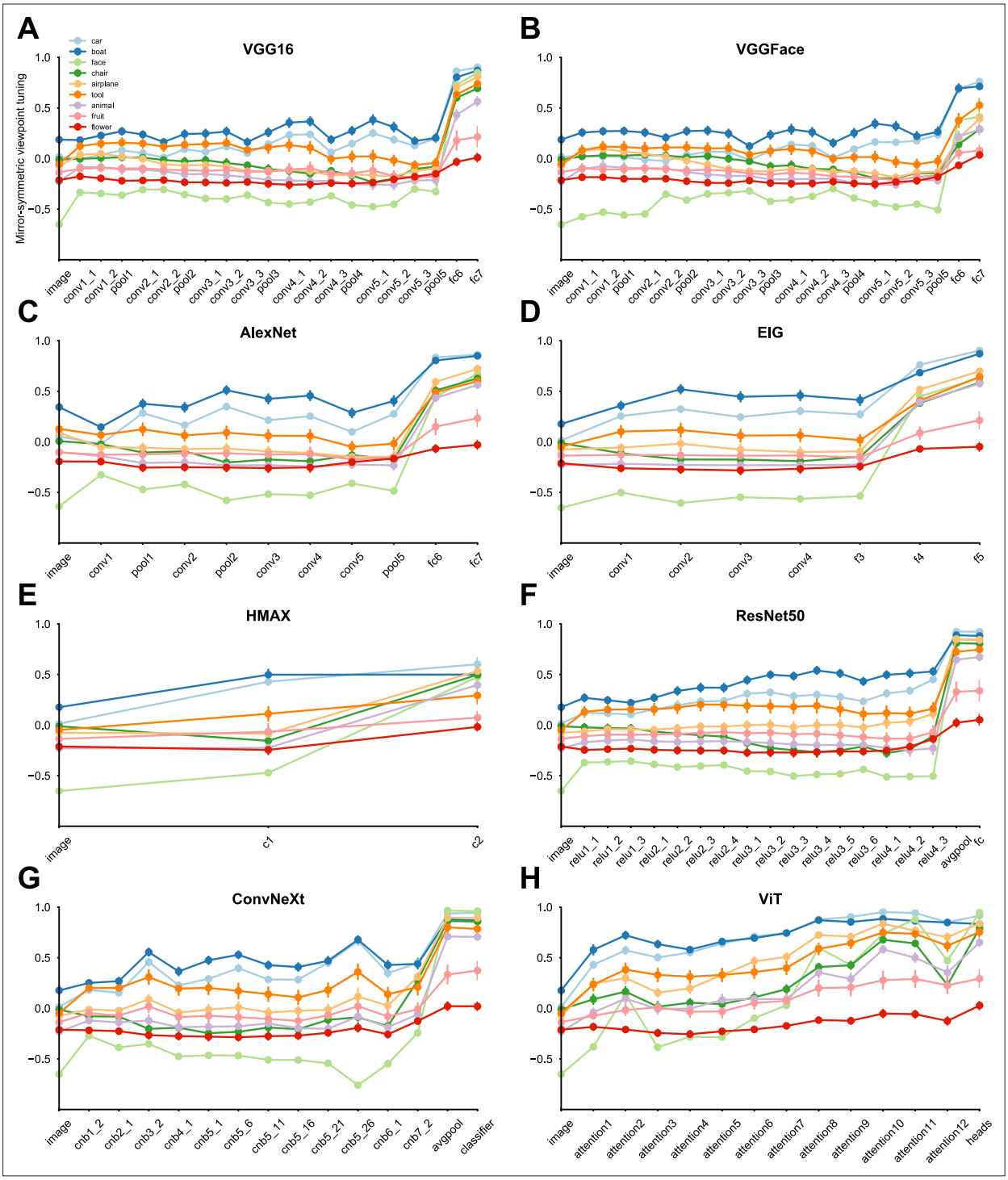

Figure 2—figure supplement 3. Convolutional networks, regardless of their architecture and training objectives, exhibit peak mirror-symmetric viewpoint tuning at the fully-connected and average pooling layers. (**A–H**) The colored curves represent the mirror-symmetric viewpoint tuning indices across nine object categories (car, boat, face, chair, airplane, animal, tool, fruit, and flower) across the neural network layers. Each solid circle indicates the average index value across 25 exemplars within each object category. Error bars denote the standard error of the mean. In all of the convolutional networks, the mirror-symmetric viewpoint tuning index peaks at the fully-connected or average pooling layers. ViT, with its non-convolutional architecture, does not exhibit this tuning profile. For face stimuli, there is a unique progression in mirror-symmetric viewpoint tuning: the index is negative for the convolutional layers, and it abruptly becomes highly positive when transitioning to the first fully connected layer. The negative indices in the convolutional layers can be attributed to the image-space asymmetry of non-frontal faces; compared to other categories, faces demonstrate pronounced front-back asymmetry, which translates to asymmetric images for all but frontal views (***Figure 2—figure supplement 1***). The features

Figure 2—figure supplement 3 continued on next page

Figure 2—figure supplement 3 continued

that drive the highly positive mirror-symmetric viewpoint tuning for faces in the fully connected layers are training-dependent (***Figure 2—figure supplement 2***), and hence, may reflect asymmetric image features that do not elicit equivariant maps in low-level representations; for example, consider a profile view of a nose. Note that cars and boats elicit high mirror-symmetric viewpoint tuning indices already in early processing layers. This early mirror-symmetric tuning is independent of training (***Figure 2—figure supplement 2***), and hence, may be driven by low-level features. Both these object categories show pronounced quadrilateral symmetry, which translates to symmetric images for both frontal and side views (***Figure 2—figure supplement 1***).

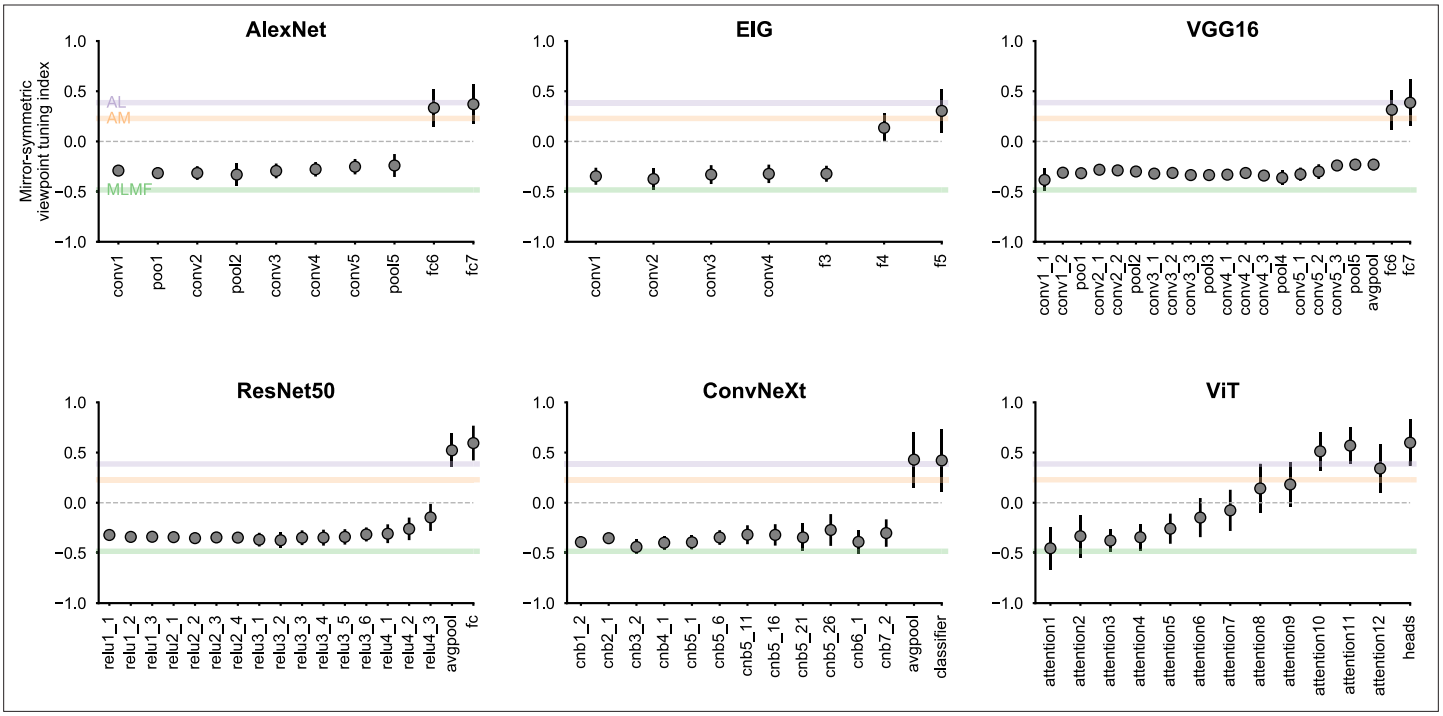

Figure 2—figure supplement 4. Mirror-symmetric viewpoint tuning of various neural network architectures measured with respect to the FIV face stimulus set (***Freiwald and Tsao, 2010***) and compared to the mirror-symmetric viewpoint tuning of three face-patches (MLMF, AL, and AM). This figure contrasts the mirror-symmetric viewpoint tuning index of macaque face patches with equivalent measurements in different neural network layers. Solid circles indicate indices for network layers, averaged across 25 face exemplars of the FIV stimulus set. The error bars show the standard error. The colored horizontal lines represent estimated mirror-symmetric viewpoint indices for three face patches (MLMF, AL, AM). To ensure that neural noise does not attenuate the measured mirror-symmetric viewpoint tuning, we divided the raw index estimated for each patch with a reliability estimate. This estimate was obtained by correlating neural RDMs pertaining to two equally sized disjoint sets of neurons recorded in that patch, averaging the result over 100 random splits, and applying a Spearman-Brown correction. Notably, the AL face patch demonstrates the most pronounced mirror-symmetric viewpoint tuning among the face patches, closely aligning with the measurements in deeper network layers. Conversely, the MLMF patch, characterized by its asymmetric representation, shows a negative index value, similar to the early and mid-level network layers. The positive index of the AM face patch, though lower than that of the AL, is consistent with a view-invariant representation (***Freiwald and Tsao, 2010***). Diverse convolutional architectures mimic the emergence of mirror-symmetric viewpoint tuning between the MLMF and AL face patches.

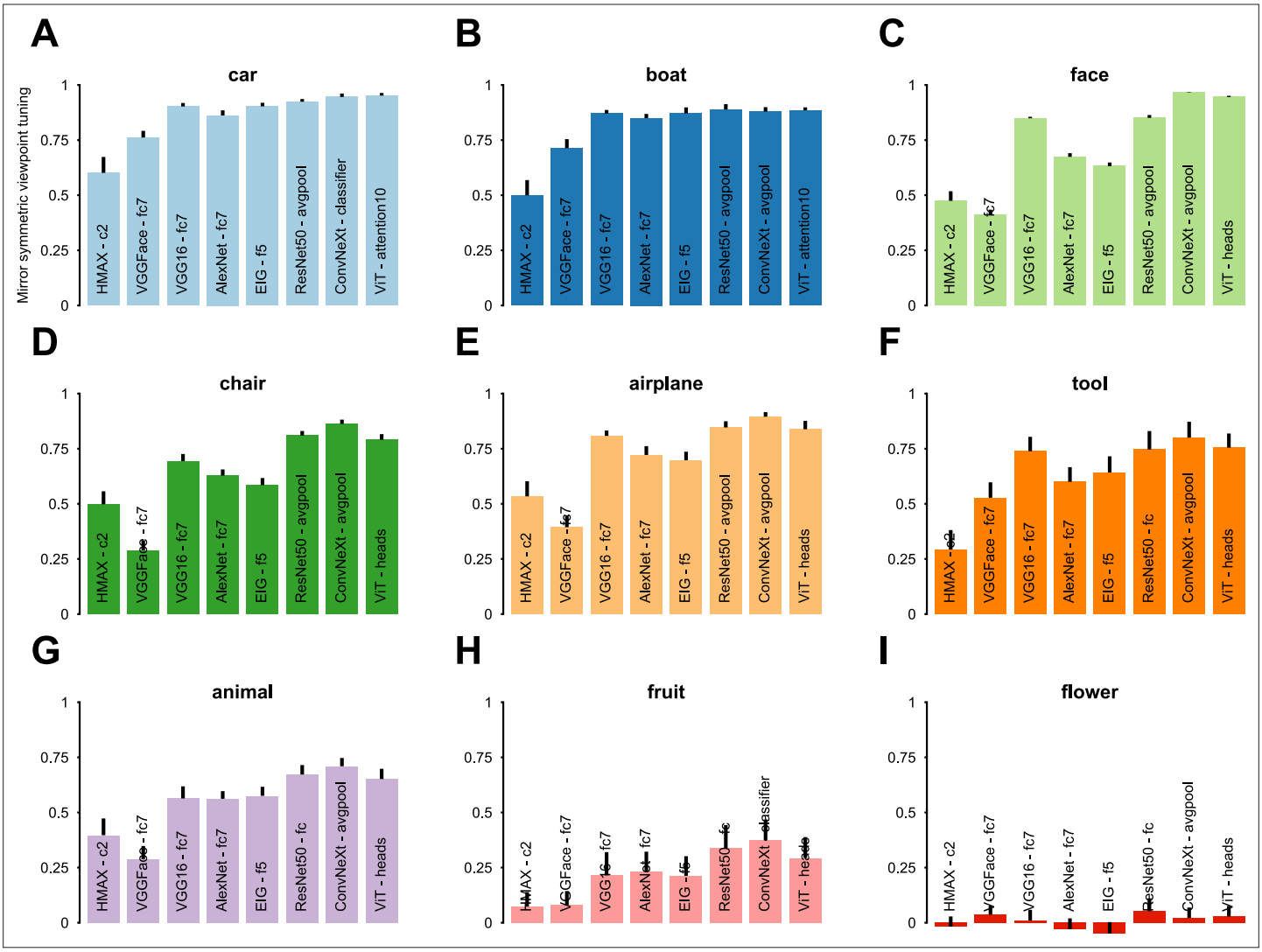

Figure 2—figure supplement 5. The highest mirror-symmetric viewpoint tuning index across all layers of each evaluated neural network model. We evaluated the following networks: HMAX, VGGFace, VGG16, AlexNet, EIG, ResNet50, ConvNeXt, and ViT. Each panel indicates the layer displaying the peak mirror-symmetric viewpoint tuning index for one object category, measured separately for each network. The deepest layers of the ConvNeXt network, especially the average pooling (avgpool) and classifier layers, exhibit the highest indices for nearly all categories. ***Yildirim et al., 2020*** reported that CNNs trained on faces, notably VGGFace, exhibited lower mirror-symmetric viewpoint tuning compared to neural representations in area AL. Consistent with their findings, our results demonstrate that VGGFace, trained on face identification, has a low mirror-symmetric viewpoint tuning index. This is especially notable in comparison to ImageNet-trained models such as VGG16. This difference between VGG16 and VGGFace can be attributed to the distinct characteristics of their training datasets and objective functions. The VGGFace training task consists of mapping frontal face images to identities; this task may exclusively emphasize higher-level physiognomic information. In contrast, training on recognizing objects in natural images may result in a more detailed, view-dependent representation. To test this potential explanation, we measured the average correlation-distance between the fc6 representations of different views of the same face exemplar in VGGFace and VGG16 trained on ImageNet. The average correlation-distance between views is 0.70 ± 0.04 in VGGFace and 0.93 ± 0.04 in VGG16 trained on ImageNet. The converse correlation distance between different exemplars depicted from the same view is 0.84 ± 0.14 in VGGFace and 0.58 ± 0.06 in VGG16 trained on ImageNet. Therefore, as suggested by Yildirim and colleagues, training on face identification alone may result in representations that cannot explain intermediate levels of face processing.

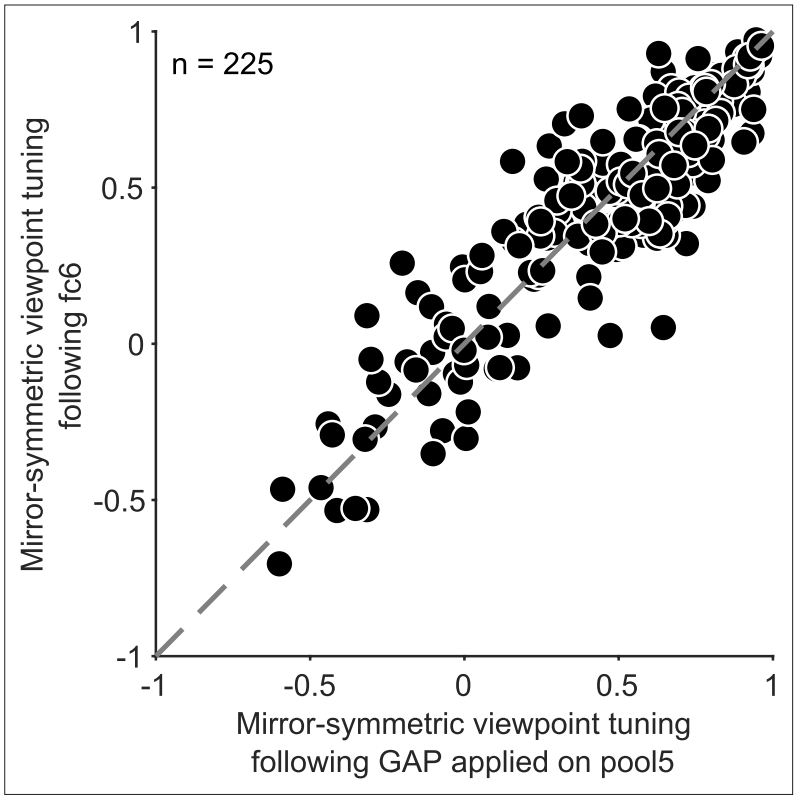

Figure 2—figure supplement 6. One of the key operations in fully-connected layers is spatial pooling. We analyzed the impact of this operation by artificially introducing global average pooling (GAP) instead of the first fully-connected layer (fc6) of ImageNet-trained AlexNet. Each element of the GAP representation refers to a spatial average of unit activations of one pool5 feature map. The scatterplot shows the mirror-symmetric viewpoint tuning index of GAP applied to pool5 (x-axis) relative to an fc6 representation (y-axis). Each circle represents one exemplar object. These results indicate that global spatial pooling introduced instead of fc6 is sufficient for rendering the pool5 representation mirror-symmetric viewpoint selective, reproducing the symmetry levels of the different fc6 view tuning curves across objects.

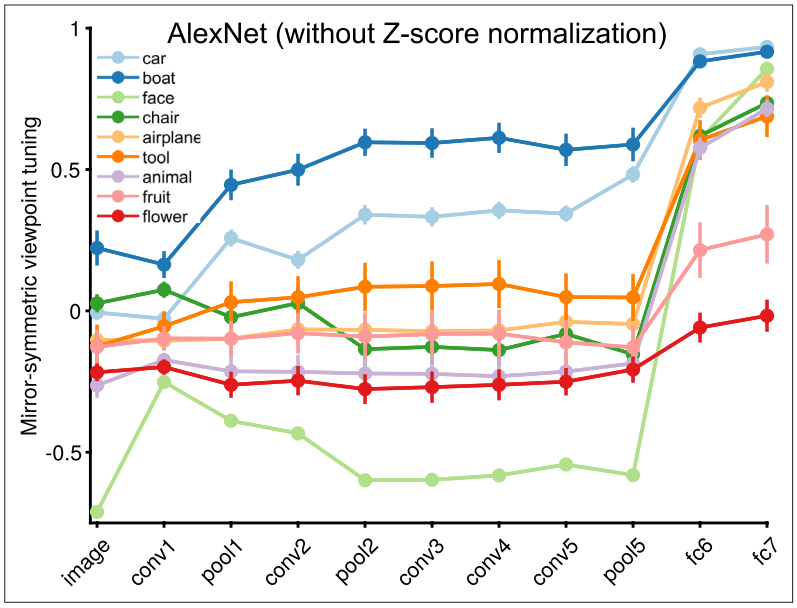

Figure 2—figure supplement 7. Layer-wise mirror-symmetric viewpoint tuning profiles measured by linear correlation without employing unit-specific z-score normalization. As in ***Figure 2***, colored curves show the mirror-symmetric viewpoint tuning indices for nine object categories across AlexNet layers. Each solid circle indicates the average index value derived from 25 exemplars in each object category. Error bars indicate the standard error of the mean. In ***Figure 2***, representational dissimilarities were measured using unit activations first centered and normalized across images (a procedure denoted as $RSA_{CorrDem}$ in ***Revsine et al., 2024***). Here, first-level correlations were calculated using raw activations (a procedure denoted as RSA_{Corr} in ***Revsine et al., 2024***). Revsine and colleagues noted that under linear-system assumptions, RSA_{Corr} yields a representational dissimilarity measure invariant to response gain; response gain might be strongly influenced by low-level factors such as luminance and contrast. The similarity of the tuning profiles observed here and in ***Figure 2*** is consistent with the interpretation of the emergent mirror-symmetric viewpoint tuning in our models as driven by learned equivariant mid-level features rather than low-level stimulus features. This result, however, does not preclude the possibility that other, uncontrolled stimulus sets could elicit viewpoint-tuning profiles that are driven by low-level confounds, as demonstrated by Revsine and colleagues.

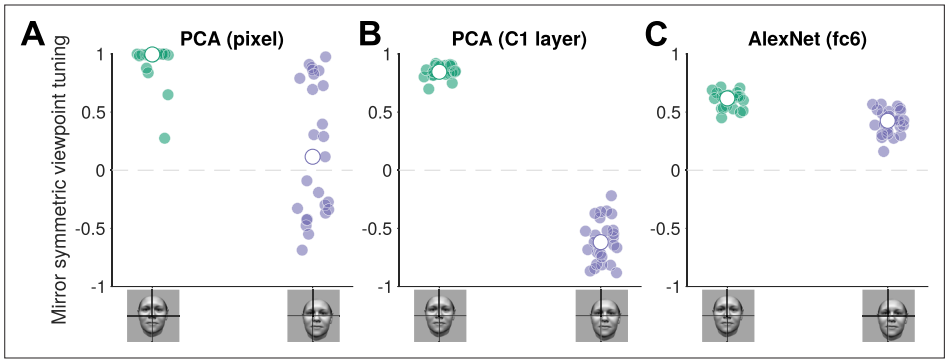

Figure 2—figure supplement 8. Comparison of mirror-symmetric viewpoint tuning in a supervised, PCA-based model (***Leibo et al., 2017***) and a supervised CNN (AlexNet) trained on object recognition. Panels A and B depict how mirror-symmetric viewpoint tuning in a re-implementation of the Leibo and colleagues model (***Leibo et al., 2017***) sharply declines for off-center test stimuli. In contrast, the same shift in center of the test stimuli has only a negligible effect on mirror-symmetric viewpoint tuning in AlexNet (Panel C). Implementation details: To reproduce the model described in ***Leibo et al., 2017***, we generated a training stimulus set using the Basel Face Model. The stimulus set consisted of untextured synthetic faces of 40 identities, each depicted from 39 viewpoints. For panel A, we estimated a PCA of the pixel-space representation of this stimulus set. For panel B, we estimated a PCA of the stimulus set's HMAX C1 layer representation. In both cases, the resulting latent representation had 1560 features (40×39). To test the model, we used the face stimulus set containing 25 exemplars in 9 viewpoints employed in ***Figure 2***. The viewpoints ranged from –90° to 90°, with a step of 22.5°. Mirror-symmetric viewpoint tuning was extracted from a representational dissimilarity matrix (RDM) created per exemplar. Green and purple circles represent mirror-symmetric viewpoint tuning in centered and shifted images (with 15-pixel shifts in the x and y axes), respectively. White circles indicate the mean across all exemplars.

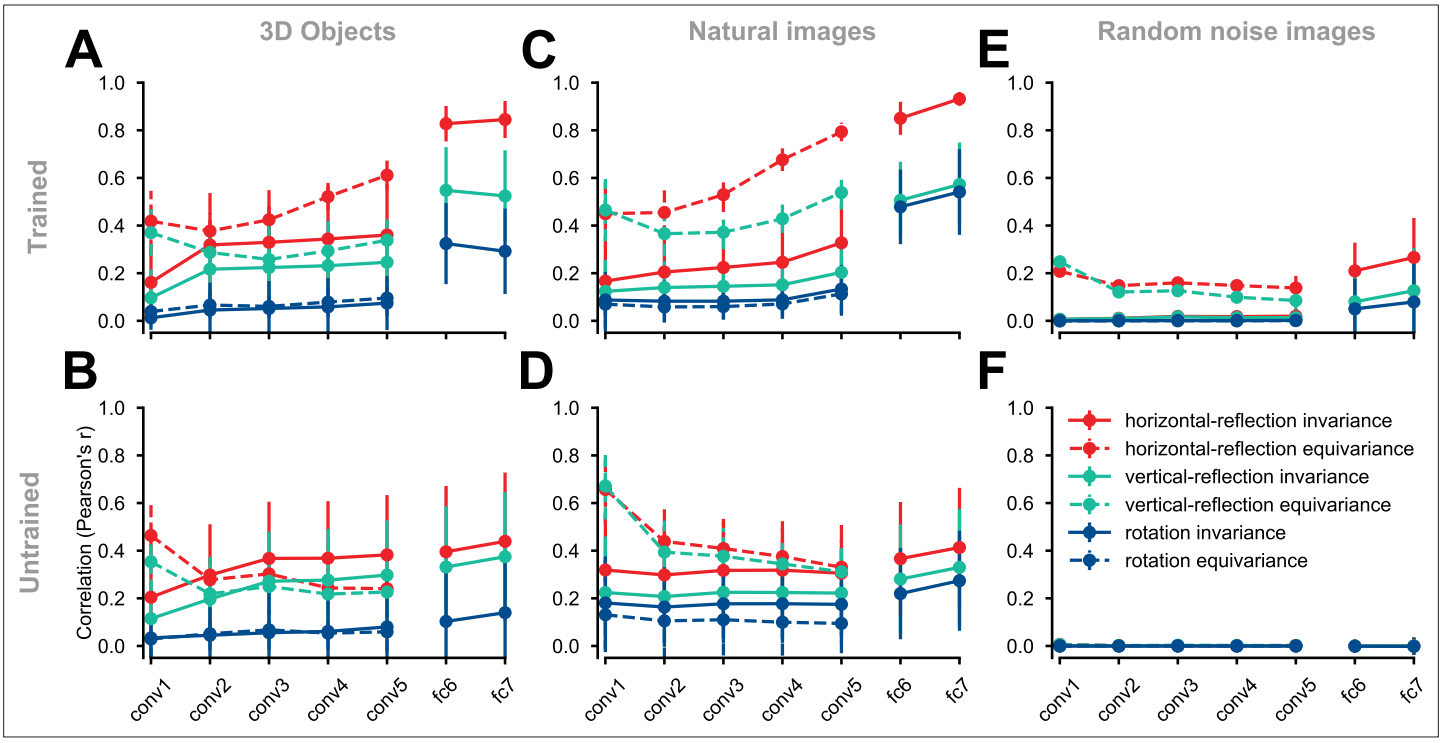

Figure 3. Equivariance and invariance in trained and untrained deep convolutional neural networks. Each solid circle represents an equivariance or invariance measure, averaged across images. Hues denote different transformations (horizontal flipping, vertical flipping, or 90° rotation). Error bars depict the standard deviation across images (each test condition consists of 2025 images). Invariance is a measure of similarity between the activity pattern an image elicits and the activity pattern its transformed (e.g. flipped) counterpart (solid lines) elicits. Equivariance is a measure of the similarity between the activity pattern of a transformed image elicits and the *transformed* version of the activity pattern the untransformed image elicits (dashed lines). In the convolutional layers, both invariance and equivariance can be measured. In the fully connected layers, whose representations have no explicit spatial structure, only invariance is measurable. (**A**) ImageNet-trained AlexNet tested on the rendered 3D objects. (**B**) Untrained AlexNet tested on rendered 3D objects. (**C**) ImageNet-trained AlexNet tested on the natural images (images randomly selected from the test set of ImageNet). (**D**) Untrained AlexNet tested on the natural images. (**E**) ImageNet-trained AlexNet tested on the random noise images. (**F**) Untrained AlexNet tested on the random noise images.

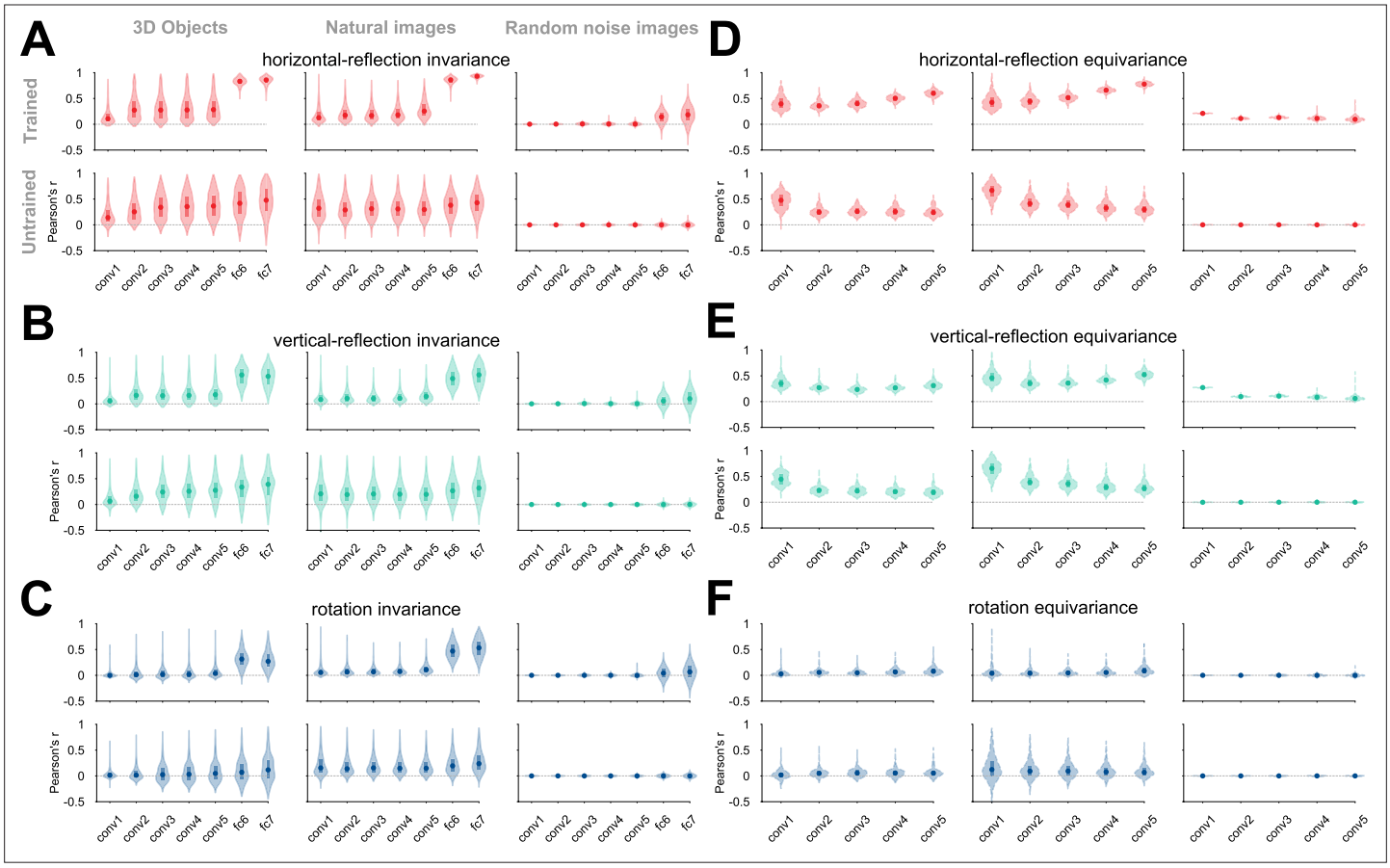

Figure 3—figure supplement 1. Image-specific representational invariance and equivariance across 3D object renders, natural images, and random noise images, measured in a deep convolutional neural network (AlexNet) trained on ImageNet or alternatively, left untrained. Invariance is measured by the linear correlation between the activity pattern elicited by an image and the activity pattern elicited by a transformed version of the image. Equivariance is measured by the linear correlation between the activity pattern elicited by a transformed image and a transformed version of the activity pattern of the untransformed image. Each violin plot depicts the distribution of invariance (panels **A-C**) or equivariance (**D–F**) image-specific measures across 2025 images. The different hues denote the transformations against which the equivariance and invariance were measured: horizontal flipping (red), vertical flipping (green), or 90° rotation (blue). The solid circles denote the median, and the thick bars, the first and third quantiles. Panels A, B, and C show the invariance over horizontally flipped, vertically flipped, and 90° rotated images, respectively. Panels D, E, and F depict the equivariance over the same transformations. ImageNet training induces equivariance (in convolutional layers) and invariance (in fully connected layers) to the horizontal reflection of most natural images and 3D renders. This effect is less pronounced for vertical reflection and 90° rotation.

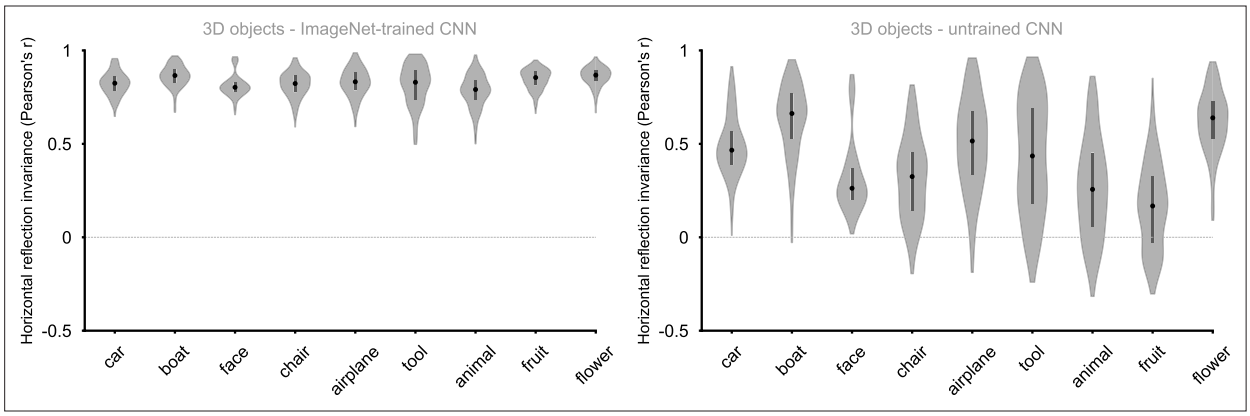

Figure 3—figure supplement 2. Training-induced enhancement of horizontal reflection invariance in the first fully connected layer (fc6), across different object categories. Elaborating on ***Figure 3*** and ***Figure 3—figure supplement 1***, we examined horizontal reflection invariance in each object category in a trained (left panel) and an untrained (right panel) AlexNet network. Reflection invariance was quantified as the correlation between representations of horizontally flipped images. The violin plots show the distribution of these correlation coefficients across views and exemplars for each object category, with vertical bars marking the median and the first and third quartiles. In an untrained network, the differences between object categories primarily reflect pixel-level symmetry. Note that frontal faces, due to their inherent left-right symmetry, elicit a higher correlation compared to other viewpoints (appearing as a positive outlier).

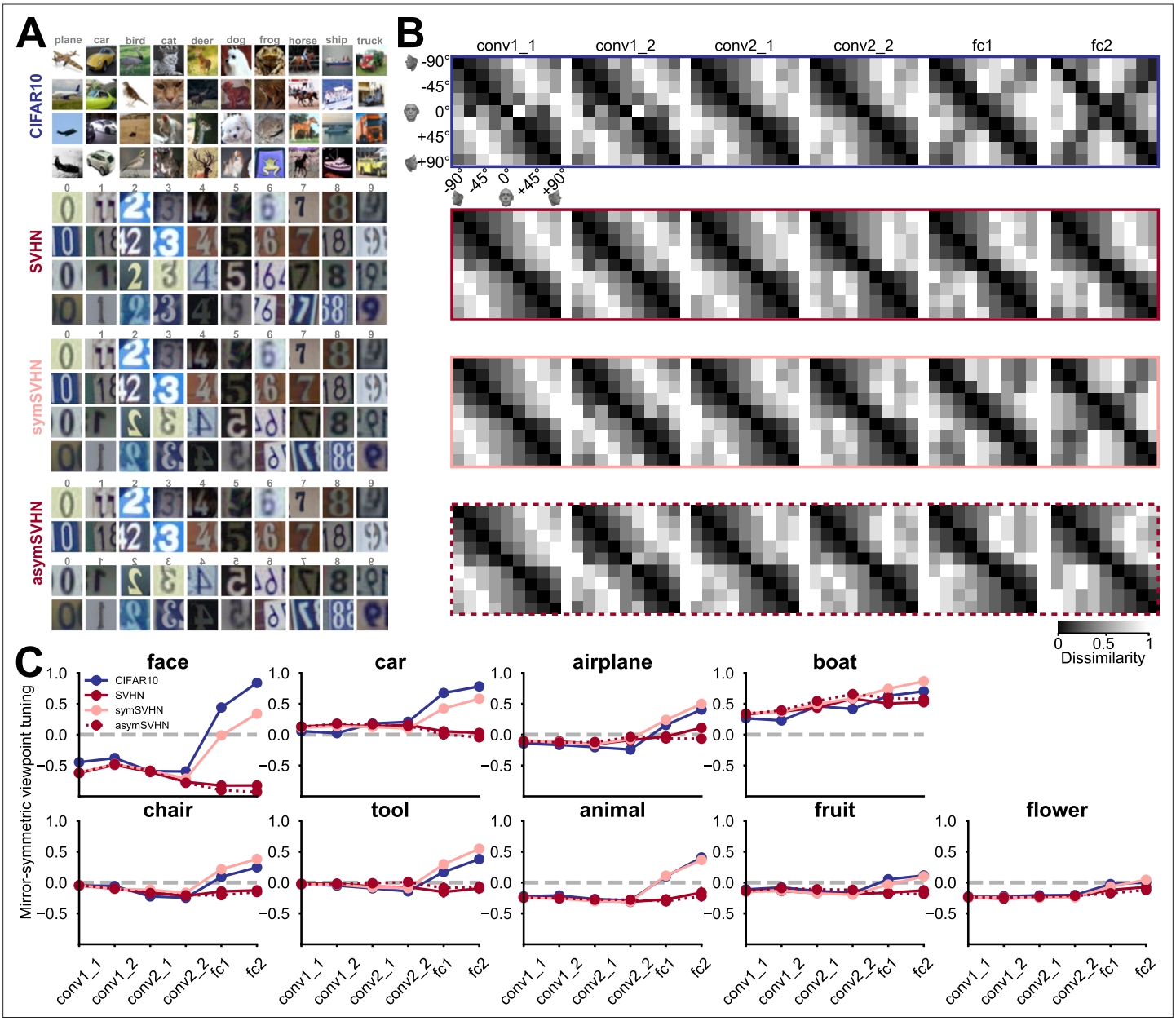

Figure 4. The effect of training task and training dataset on mirror-symmetric viewpoint tuning. (**A**) Four datasets are used to train deep neural networks of the same architecture: CIFAR-10, a natural image dataset with ten bilaterally symmetric object categories; SVHN, a dataset with mostly asymmetric categories (the ten numerical digits); symSVHN, a version of the SVHN dataset in which the categories were made bilaterally symmetric by horizontally reflecting half of the training images (so the digit 7 and its mirrored version count as members of the same category); asymSVHN, the same image set as in symSVHN but with the mirrored images assigned to ten new distinct categories (so the digit 7 and its mirrored version count as members of distinct categories). (**B**) Each row represents the RDMs of the face exemplar images from nine viewpoints for each trained network corresponding to its left side panel. Each entry of the RDM represents the dissimilarity (1 - Pearson's r) between two pairs of image-induced activity vectors in the corresponding layer. The RDMs' order from left to right refers to the depth of layers within the network. As the dissimilarity color bar indicates, the dissimilarity values increase from black to white color. (**C**) Mirror-symmetric viewpoint tuning index values across layers for nine object categories in each of the four networks. The solid circles refer to the average of the index across 25 exemplars within each object category for three networks trained on 10 labels. The red dashed line with open circles belongs to the asymSVHN network trained on 20 labels. The gray dashed lines indicate the index of zero. Error bars represent the standard error of the mean calculated across exemplars.

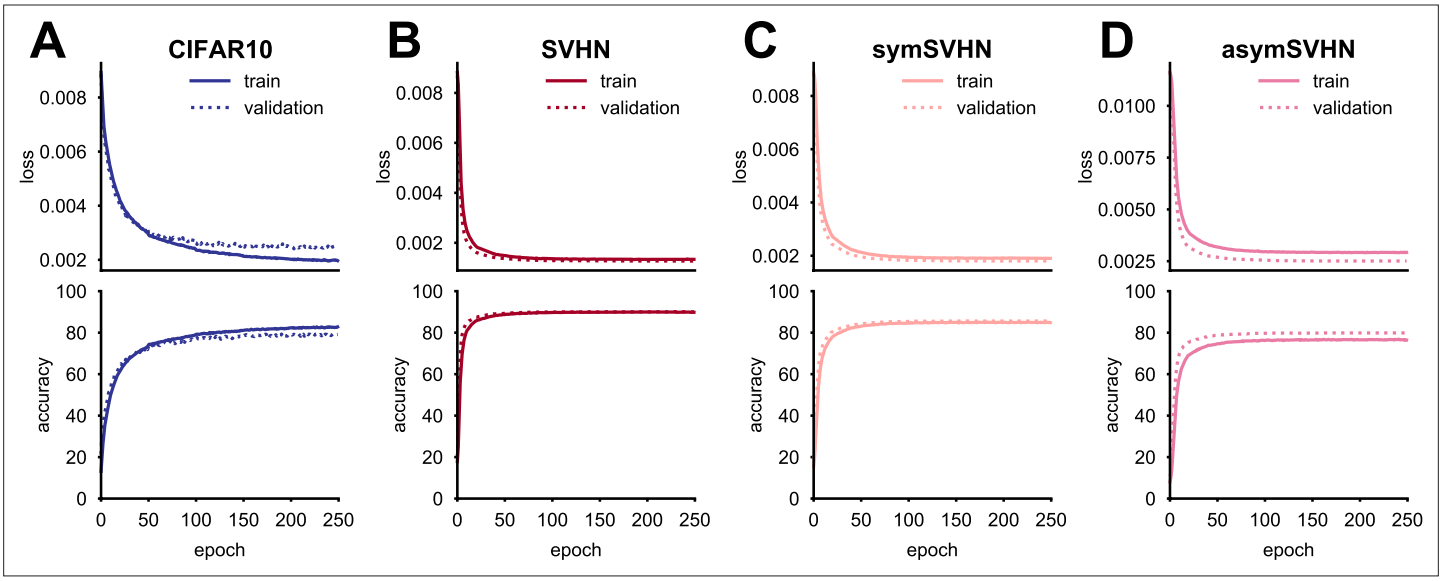

Figure 4—figure supplement 1. Network learning curves. (**A–D**) Loss and accuracy curves for the networks trained by CIFAR-10 (**A**), SVHN (**B**), symSVHN (**C**), asymSVHN (**D**) datasets. The x-axis denotes training epochs. Note that the accuracy of asymSVHN might be negatively affected by the inclusion of relatively symmetric categories such as 0 and 8. We used drop-out during training, which resulted in higher training loss compared to the validation loss.

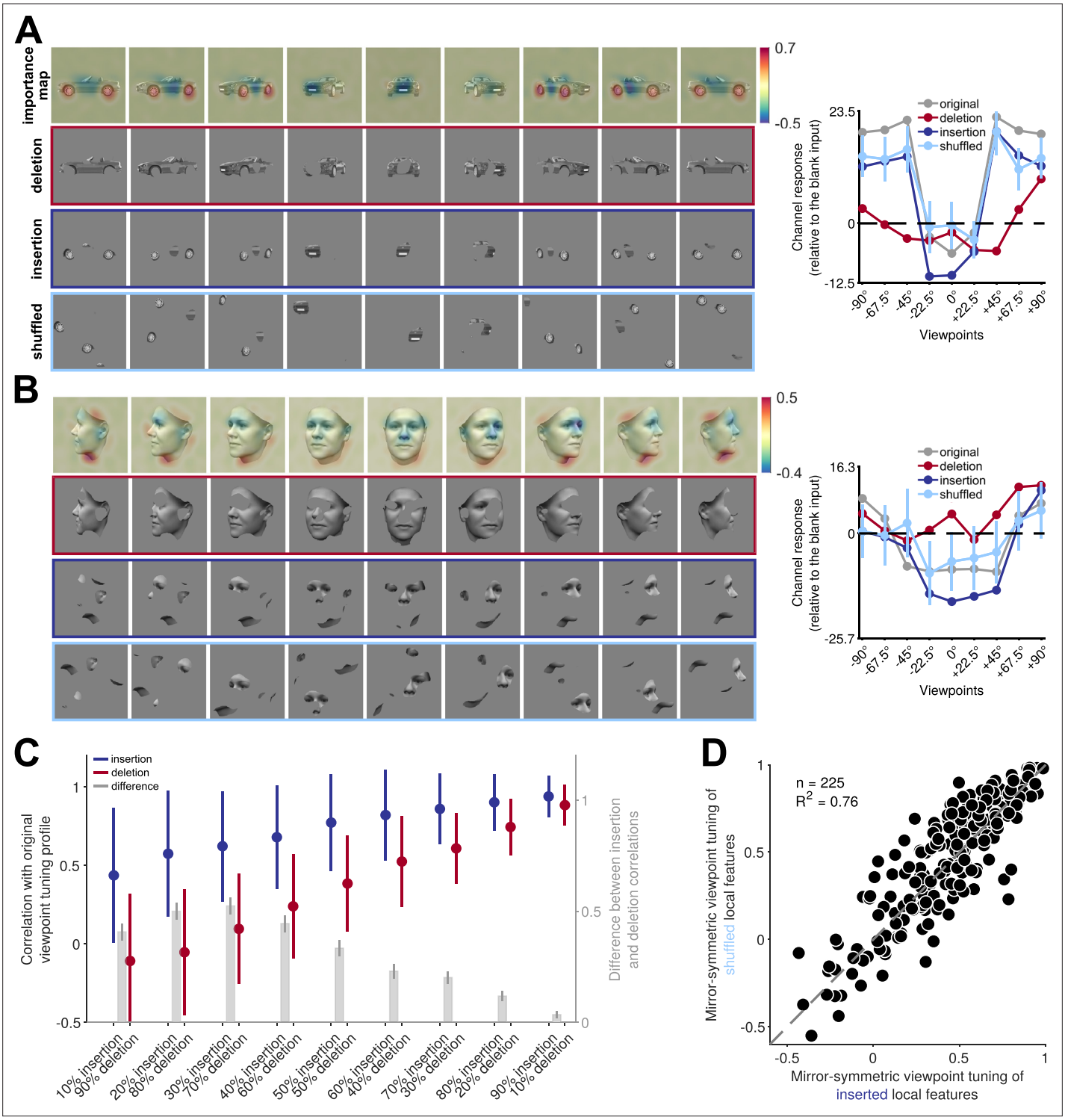

Figure 5. Reflection-invariant viewpoint-specific responses are driven mostly by local features. This figure traces image-level causes for the mirror-symmetric viewpoint tuning using Randomized Input Sampling for Explanation (RISE, ***Petsiuk et al., 2018***). (**A**) Analysis of the features of different views of a car exemplar that drive one particular unit in fully connected layer fc6 of AlexNet. The topmost row in each panel depicts an image-specific *importance map* overlaid to each view of the car, charting the contribution of each pixel to the unit's response. The second row ('deletion') depicts a version of each input image in which the 25% most contributing pixels are masked with the background gray color. The third row ('insertion') depicts a version of the input images in which only the most contributing 25% of pixels appear. The last row represents the shuffled spatial configuration of extracted local features, which maintains their structure and changes their locations. The charts on the right depict the units' responses to the original, deletion, insertion, and shuffled images. The dashed line indicates the units' response to a blank image. The y-axis denotes the unit's responses

Figure 5 continued on next page

Figure 5 continued

compared to its response to a blank image. (**B**) Analogous analysis of the features of different views of a face that drive a different unit in fully connected layer fc6 of AlexNet. (**C**) Testing local contributions to mirror-symmetric viewpoint tuning across all object exemplars and insertion/deletion thresholds. For each object exemplar, we selected a unit with a highly view-dependent but symmetric viewpoint tuning (the unit whose tuning function was maximally correlated with its reflection). We then measured the correlation between this tuning function and the tuning function induced by insertion or deletion images that were generated by a range of thresholding levels (from 10 to 90%). Note that each threshold level consists of images with the same number of non-masked pixels appearing in the insertion and deletion conditions. In the insertion condition, only the most salient pixels are retained, and in the deletion condition, only the least salient pixels are retained. The solid circles and error bars indicate the median and standard deviation over 225 objects, respectively. The right y-axis depicts the difference between insertion and deletion conditions. Error bars represent the SEM. (**D**) For each of 225 objects, we selected units with mirror-symmetric viewpoint tuning above the 95 percentile (≈200 units) and averaged their corresponding importance maps. Next, we extracted the top 25% most contributing pixels from the averaged maps (insertion) and shuffled their spatial configuration (shuffled). We then measured the viewpoint-RDMs for either the inserted or shuffled object image set. The scatterplot compares the mirror-symmetric viewpoint tuning index between insertion and shuffled conditions, calculated across the selected units. Each solid circle represents an exemplar object. The high explained variance indicates that the global configuration does not play a significant role in the emergence of mirror-symmetric viewpoint tuning.

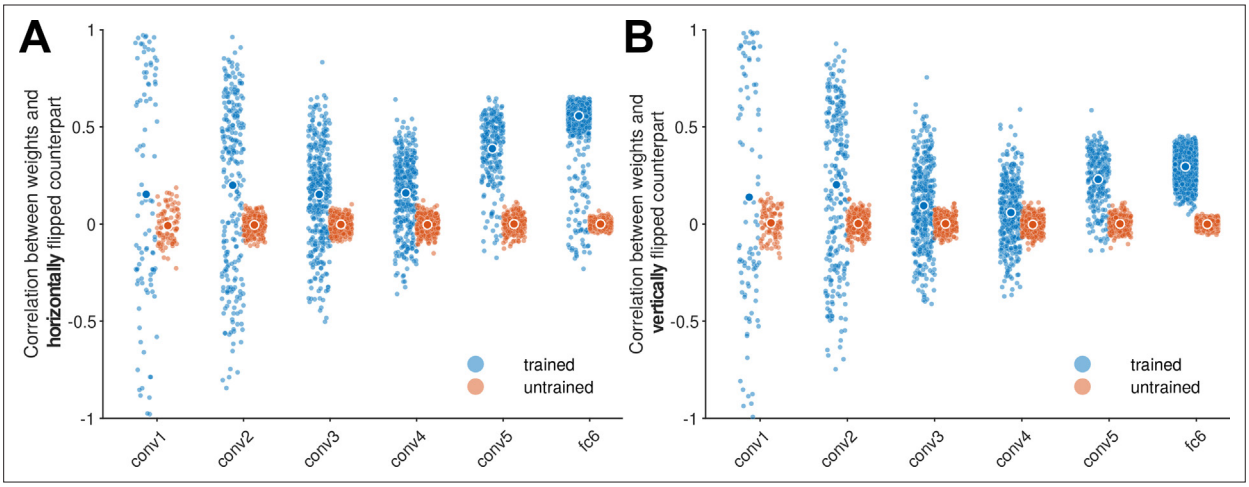

Figure 5—figure supplement 1. The emergence of mirror symmetric weight tensors in AlexNet. In order to examine the symmetry of neural network weights, we measured the linear correlation between each convolutional weight kernel and its horizontally (panel A) or vertically (panel B) flipped counterpart. To avoid replicated observations in the correlation analysis, we considered only the left (or top) half of the matrix, and excluded the central column (or row). Each dot represents one channel. This measurement was done for each convolutional layer in an AlexNet trained on ImageNet, as well as in an untrained AlexNet. The symmetry of the incoming weights to fc6 was evaluated in a similar fashion (note that the weights leading into this layer still have an explicit spatial layout, unlike fc7 and fc8). This analysis demonstrates that in the ImageNet-trained AlexNet network, weight symmetry increases with depth. Note that ImageNet training induces some highly asymmetrical kernels in conv1 and conv2. Together, these results suggest that while asymmetrical filters are useful low-level representations, the trained network incorporates symmetric weight kernels to generate downstream reflection-invariant representations.

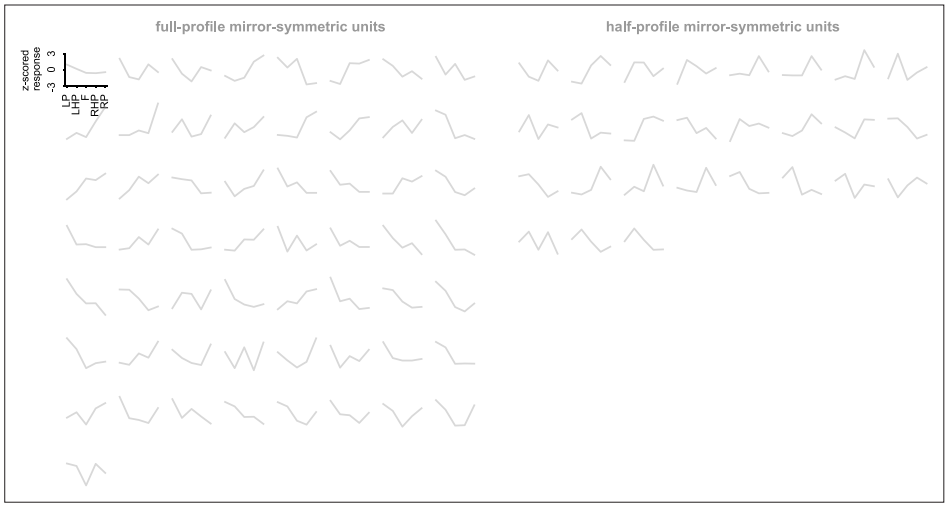

Figure 5—figure supplement 2. Individual neural network units exhibiting mirror-symmetric view tuning according to the criterion employed by ***Baek et al., 2021a***. We screened the units of the deepest convolutional layer of an untrained AlexNet according to the selection criterion proposed by Baek and colleagues (Figure S10 in ***Baek et al., 2021a***), using the official code sharedon https://github.com/vsnnlab/Face (***Baek et al., 2021b***). Each trace represents an individual unit response profile. The x-axis shows the views: left profile (LP), left half-profile (LHP), frontal (**F**), right half-profile (RHP), and right profile (RP). The y-axis depicts the response of an individual unit, z-scored standardized across images. The left panel displays units with full-profile symmetry response tuning, and the right panel displays units with half-profile response tuning. Reproducing Baek and colleagues' findings, we identified many randomly initialized units that met the selection criterion Baek and colleagues proposed. However, as this figure illustrates, a large proportion of these units exhibit markedly asymmetric tuning profiles. Specifically, while the selection criterion requires unit activation to peak at either full-profile or half-profile views, many such units exhibit less pronounced or even minimal responses to opposite views. In our subsequent analyses (***Figure 5—figure supplements 3 and 4***), we applied a stricter selection criterion.

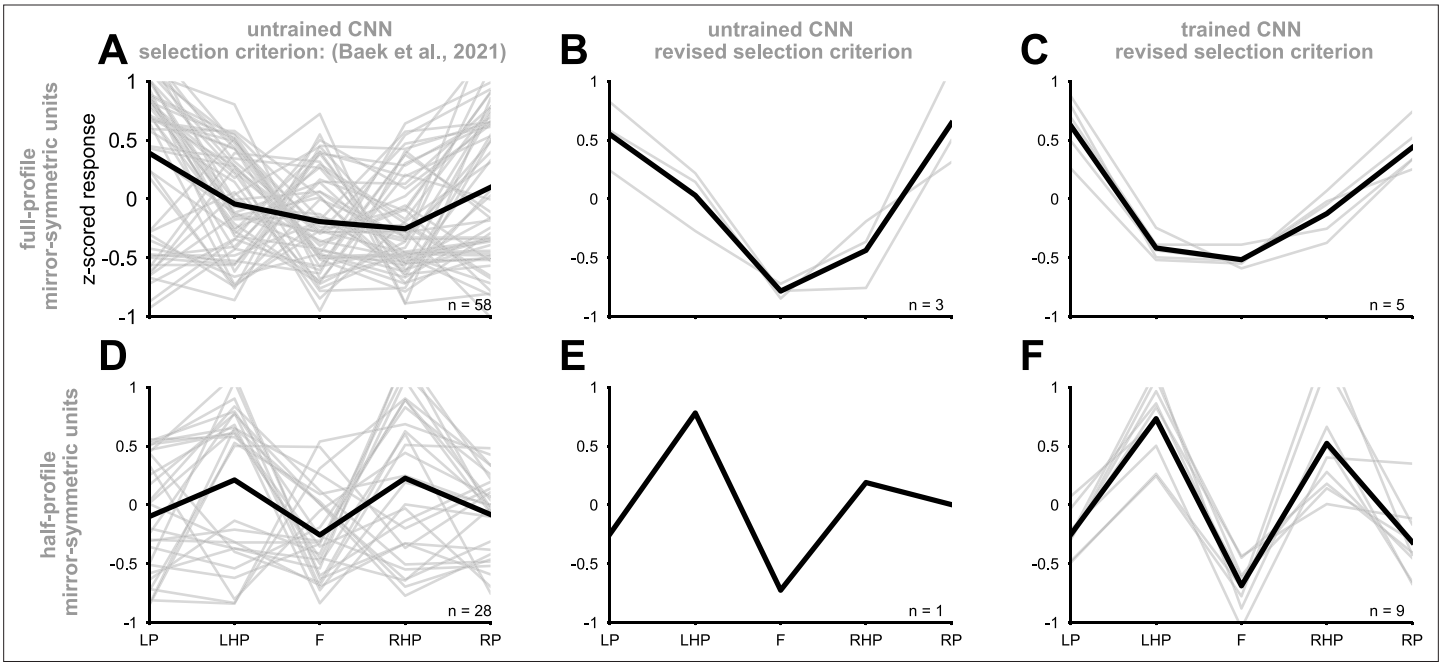

Figure 5—figure supplement 3. Selecting individual units with genuine mirror-symmetric viewpoint tuning. (Left column) Aggregated full-profile (panel A) and half-profile (panel D) mirror-symmetric units (detailed individually in ***Figure 5—figure supplement 2***), accompanied by their average tuning curves (represented as thick lines). Note that the average viewpoint tuning profile demonstrates strong mirror symmetry, yet this profile is unrepresentative of the individual units. (Middle column) The tuning profiles of units selected using a revised selection criterion. Specifically, we required the second peak to occur in response to the view opposite the first peak and ensured that the frontal view elicited the lowest response. This criterion led to fewer units being selected yet ensured each unit individually exhibited mirror-symmetric viewpoint tuning. (Right column) Units meeting the revised criterion in a trained network. Training increased the number of units individually exhibiting mirror-symmetry tuning profiles, as quantified further in ***Figure 5—figure supplement 4***.

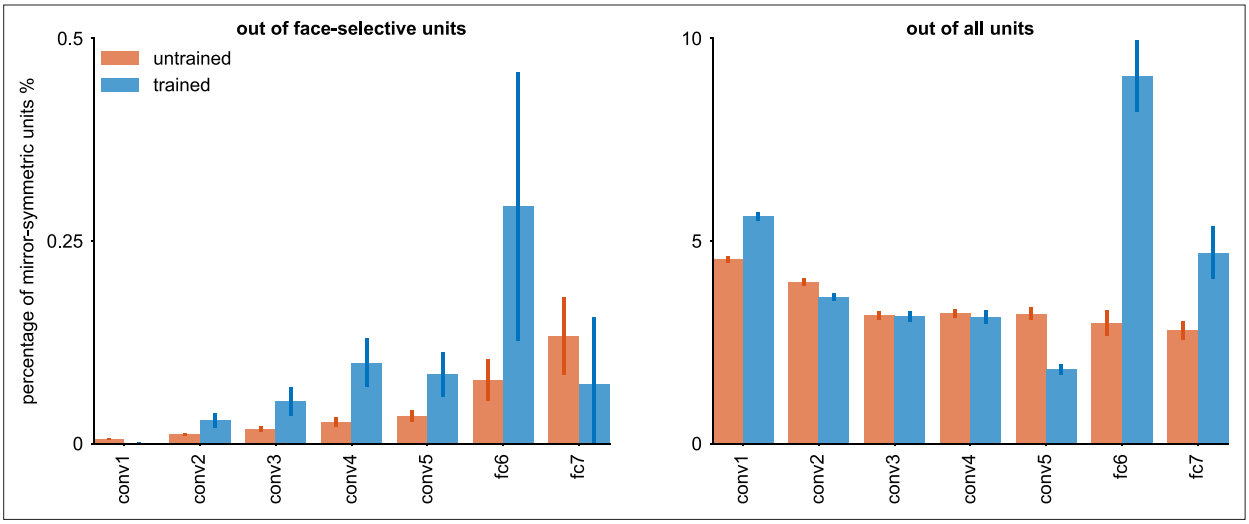

Figure 5—figure supplement 4. Training-dependent emergence of units with mirror-symmetric viewpoint tuning across neural network layers. Using our revised criterion for identifying units with mirror-symmetric tuning, we estimated the percentage of such units in each layer of an AlexNet network (Torchvision implementation), before and after training on ImageNet. (Left panel) The percentage of units with mirror-symmetric tuning out of units defined as 'face-selective' according to the face-selectivity criterion proposed by ***Baek et al., 2021a***. (Right panel) The percentage of units with mirror-symmetric viewpoint tuning, out of all of the units in each layer. Note that the latter measurement aligns more closely with the population RSA analyses in the main text, which likewise consider all units rather than just a face-selective sub-population. For each layer, the orange bars indicate the average percentage of mirror-symmetric units observed across 10 random network initializations, with the orange error bars denoting a 95% confidence interval for this proportion. The blue bars indicate the percentage of such units post-training. Since we used a single trained network for this analysis, the blue error bars denote 95% binomial confidence intervals calculated within each layer rather than across realizations. The first fully connected layer shows the most pronounced training-dependent emergence of mirror-symmetric viewpoint tuning units, consistent with the findings obtained with the population-level RSA findings described in the main text.

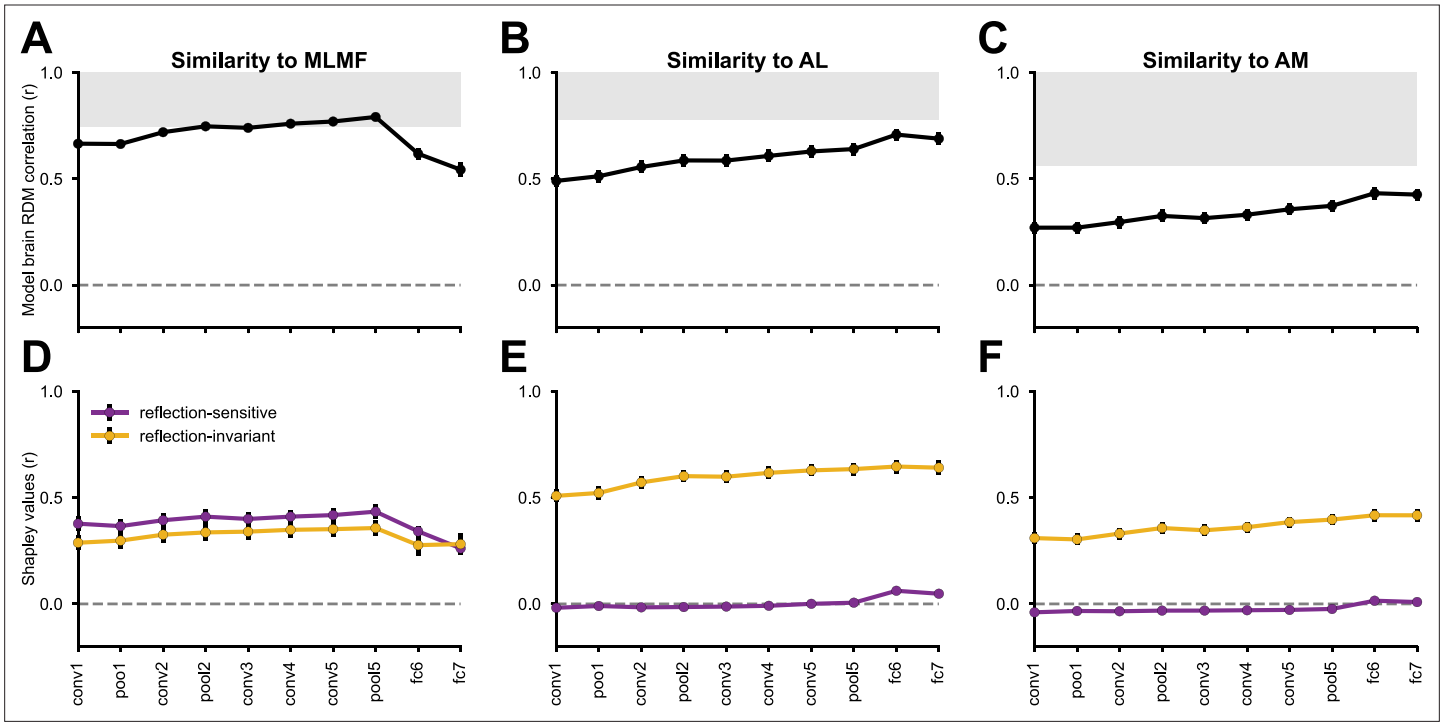

Figure 6. Reflection-invariant and reflection-sensitive contributions to the representational similarity between monkey face patch neurons and AlexNet layers. The neural responses were obtained from ***Freiwald and Tsao, 2010***, where electrophysiological recordings were conducted in three faces patches while the monkeys were presented with human faces of various identities and views. (Top row) linear correlations between RDMs from each network layer and each monkey face patch (MLMF, AL, AM). Error bars represent standard deviations estimated by bootstrapping individual stimuli (see Materials and methods). The gray area represents the neural data's noise ceiling, whose lower bound was determined by Spearman-Brown-corrected split-half reliability, with the splits applied across neurons. (Bottom row) Each model–brain RDM correlation is decomposed into the additive contribution of two feature components: reflection-sensitive (purple) and reflection-invariant (yellow). ***Figure 6—figure supplements 1–3*** present the same analyses applied to a diverse set of neural network models, across the three regions.

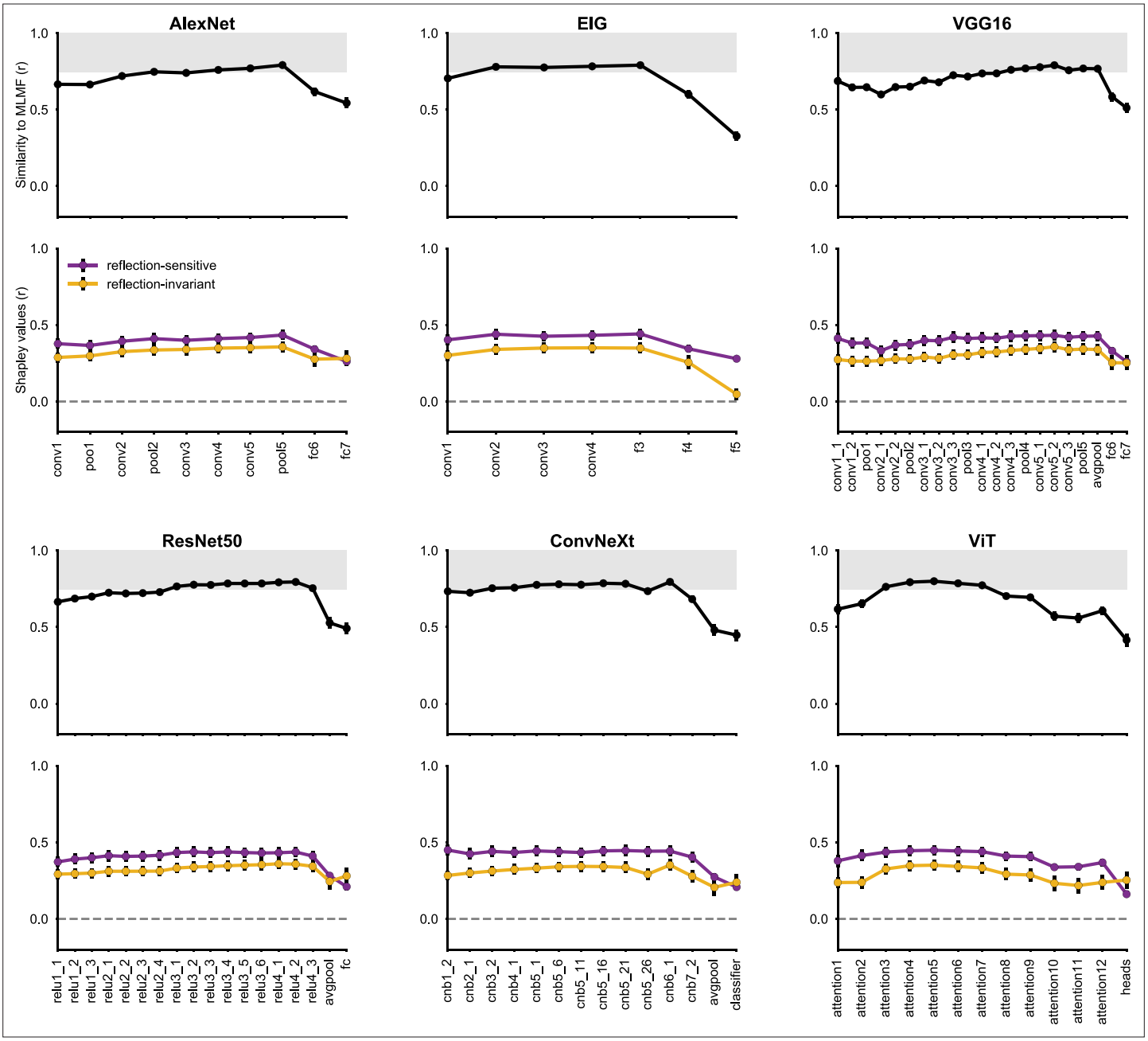

Figure 6—figure supplement 1. Alignment of MLMF and neural network representations across diverse architecures. As in ***Figure 6***, representational alignment was measured with respect to the FIV dataset. Top row depicts the correlation between model RDMs, measured in each individual neural network layer, and a neural population RDM estimated using neural recordings from the MLMF face patch. Black circles represent correlation coefficients averaged across bootstrap simulations (resampling individual stimuli), with error bars denoting standard deviations across bootstrap simulations. The gray area represents the neural RDM's noise ceiling; its lower bound was determined through a Spearman-Brown corrected split-half reliability estimate, splitting the neurons into equally sized random subsets. The bottom row displays Shapley values reflecting the contributions of the reflection-invariant and reflection-sensitive components in the model RDMs. Deeper convolutional layers in various convolutional architectures demonstrated strong alignment with MLMF data; this alignment is primarily explained by reflection-sensitive features.

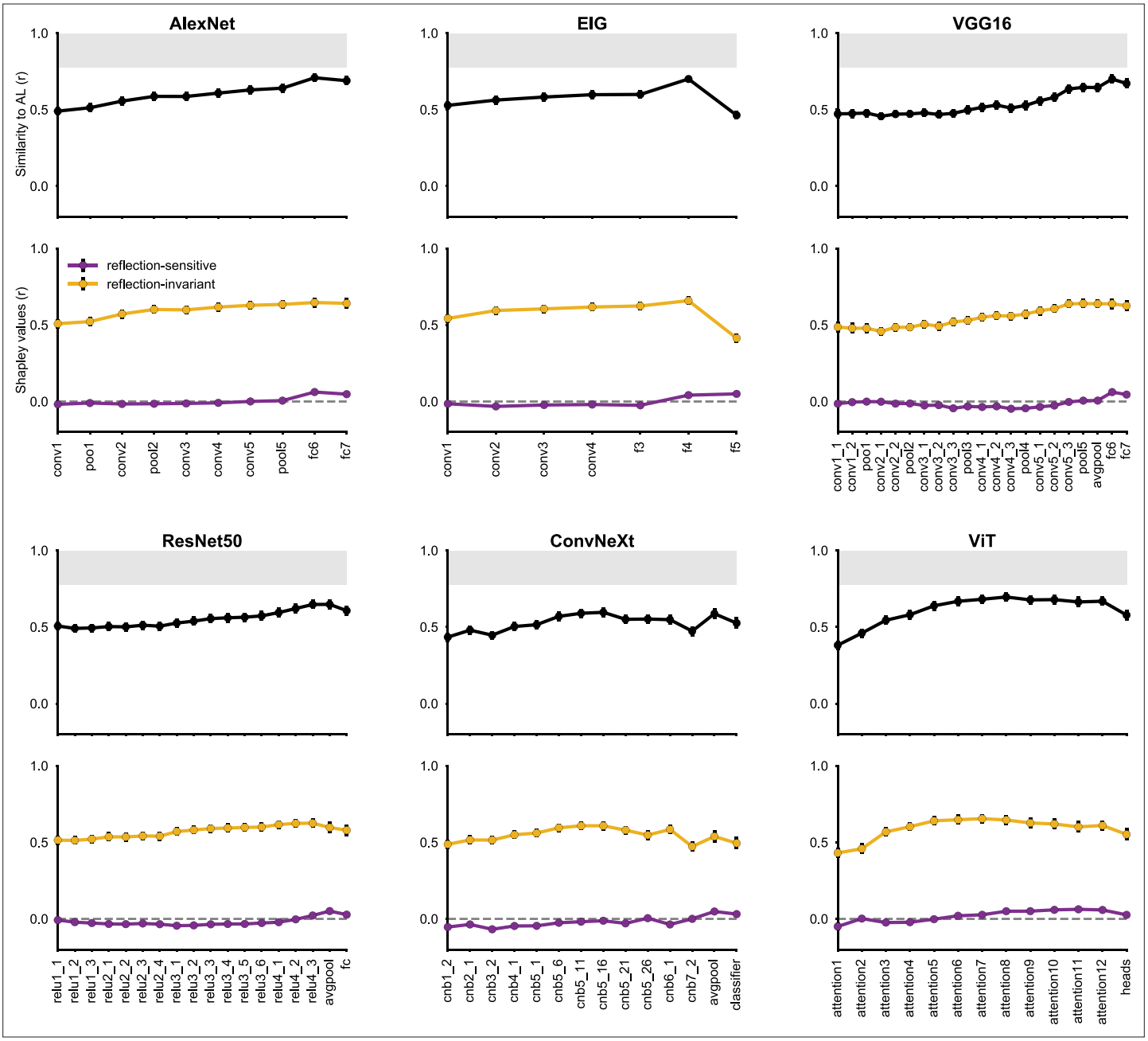

Figure 6—figure supplement 2. Alignment of AL and neural network representations across diverse architecures. The analysis is analogous to what is described in ***Figure 6—figure supplement 1***, but for the AL face patch. In various convolutional architectures, the fully connected and average pooling layers showed notable representational alignment with the AL patch. This alignment is predominantly explained by features that are invariant to reflection, rather than those sensitive to reflection.

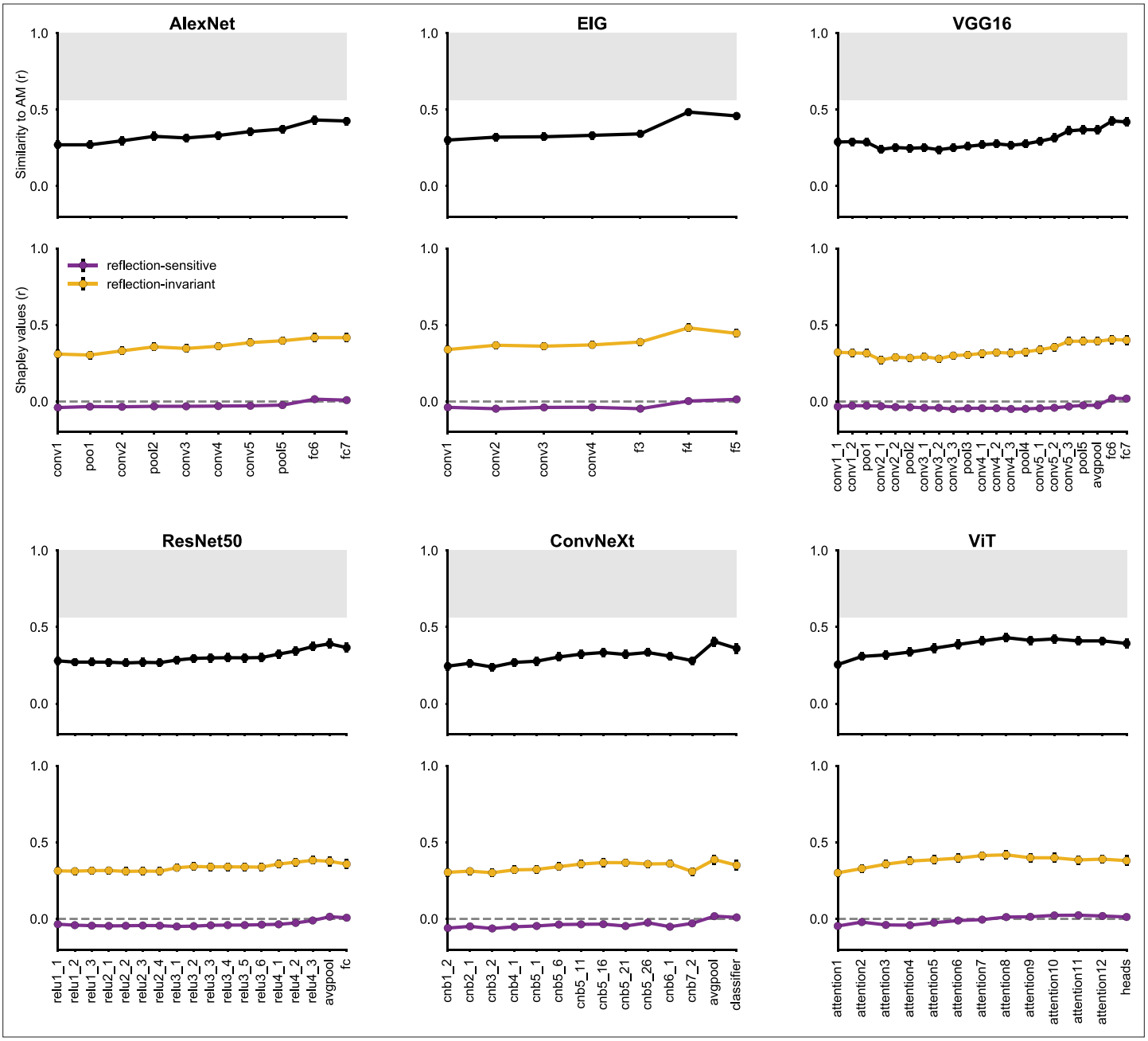

Figure 6—figure supplement 3. Alignment of AM and neural network representations across diverse architecures. The analysis is analogous to what is described in ***Figure 6—figure supplement 1***, but for the AM face patch. The deepest layers in different network architectures, with the exception of ViT, show strong representational alignment with the AM face patch. This alignment is predominantly explained by features that are invariant to reflection, rather than those sensitive to it.