Supplementary Material for HINT: Hierarchical Neuron Concept Explainer

Andong Wang, Wei-Ning Lee, Xiaojuan Qi
The University of Hong Kong
wangad@connect.hku.hk, wnlee@eee.hku.hk, xjqi@eee.hku.hk

In this supplementary file, first, we show the five modified saliency methods and five aggregation approaches with which HINT can be implemented in Section 1 and 2 respectively. Second, we explain the properties that HINT's Shapley value-based neuron contribution scoring approach satisfies in Section 3. Third, we provide detailed descriptions of applications of HINT – saliency method evaluation, explaining adversarial attack, and evaluation of COVID19 classification models – in Section 4. Next, we demonstrate more neuron-concept associations and the activation maps of multimodal neurons in Section 5. Then, we show more quantitative analysis and illustrations of the results of applying HINT for Weakly Supervised Object Localization tasks in Section 6. Finally, we provide more illustrations of ablation studies on modified saliency methods and Shapley value-based scoring approach in Section 7.

1. Modified Saliency Methods

Inspired by backpropagation-based saliency methods, we develop a saliency-guided approach to identify responsible regions in feature map $\boldsymbol{z}$. Equation (S.1) shows how the representative backpropagation-based saliency method, Gradient (Vanilla Backpropagation) [18], calculates the contribution of pixel $\boldsymbol{x}_{:,i_0,j_0}$ to a class C_k.

$$\frac{\partial f^{C_k}(\boldsymbol{x})}{\partial \boldsymbol{x}_{:,i_0,j_0}} \tag{S.1}$$

where f is a deep network, $f^{C_k}(\boldsymbol{x})$ is the logit of $\boldsymbol{x}$ to class C_k, and $\boldsymbol{x}_{:,i_0,j_0}$ is a pixel.

We extend the idea of saliency maps to hidden layers. We take concept e and neurons $\mathcal{D}$ on the l^{th} layer as an example. Given an image $\boldsymbol{x}$ with label C_k where C_k is concept e or a subcategory of concept e, the contribution of spatial activation $\boldsymbol{z}_{\mathcal{D},i_l,j_l}$ to class C_k (also to concept e) is shown in Equation (S.2)

$$\boldsymbol{s}_{\mathcal{D},i_l,j_l} = \frac{\partial f^{C_k}(\boldsymbol{z})}{\partial \boldsymbol{z}_{\mathcal{D},i_l,j_l}} \tag{S.2}$$

where $\boldsymbol{s}_{\mathcal{D},i_l,j_l} \in \mathbb{R}^{|\mathcal{D}|}$ is a vector and $\boldsymbol{s}_{\mathcal{D},i_l,j_l}$s for each i_l and j_l form the saliency map $\boldsymbol{s}$.

As shown in Table S.1, we modify five backpropagation-based saliency methods. All of them can be used in HINT.

2. Aggregation Approaches

With saliency map $\boldsymbol{s}$, the next step is to aggregate $\boldsymbol{s}_{\mathcal{D},i_l,j_l}$, and the aggregated value will be used to decide whether $\boldsymbol{z}_{\mathcal{D},i_l,j_l}$s belong to responsible foreground regions or irrelevant background regions. We implement five aggregation approaches shown in Table S.1. All of them can be applied to HINT. Note that the aggregation is only conducted along the first dimension of $\boldsymbol{s}$.

3. Properties of HINT's Shapley Value-based Neuron Contribution Scoring Approach

In the main paper, the Shapley value ϕ of a neuron d to a concept e is calculated as Equation (S.3).

$$\phi = \frac{\sum_{\boldsymbol{r}} \left| \sum_{i=1}^{M} \left(L_e^{\langle \mathcal{S} \cup d \rangle}(r) - L_e^{\langle \mathcal{S} \rangle}(r) \right) \right|}{M |\boldsymbol{r}_{\mathcal{E}} \cup \boldsymbol{r}_{b^*}|} \tag{S.3}$$

where $\mathcal{D}$ is the set of neurons; L_e is the classifier for concept e; $r = z_{\mathcal{D},i,j}$ represents spatial activation; $\boldsymbol{r}_{\mathcal{E}}$ and $\boldsymbol{r}_{b^*}$ are responsible regions of all concept $e \in \mathcal{E}$ and background regions; $\mathcal{S} \subseteq \mathcal{D} \backslash d$ is the neuron subset randomly selected at each iteration; $\langle * \rangle$ is an operator keeping the neurons in the brackets, *i.e.*, $\mathcal{S} \cup d$ or $\mathcal{S}$, unchanged while randomizing others; M is the number of iterations of Monte-Carlo sampling; $L_e^{\langle * \rangle}$ means that the classifier is re-trained with neurons in the brackets unchanged and others being randomized.

The following explains the properties of efficiency, symmetry, dummy, and additivity that Shapley values satisfy [16], *i.e.*, our Shapley value-based scoring approach satisfies.

Efficiency. The sum of neuron contributions should be equal to the difference between the prediction for r and its

Table S.1. Modified saliency methods and aggregation approaches

Modified saliency methods Λ on the l^{th} layer with respect to concept e		Aggregation approaches ζ	
Vanilla Backpropagation [18]	$\frac{\partial f^e(\boldsymbol{x})}{\partial \boldsymbol{z}}$	Norm	$\|\boldsymbol{s}\|$
Gradient x Input [17]	$\boldsymbol{z} \odot \frac{\partial f^e(\boldsymbol{x})}{\partial \boldsymbol{z}}$	Filter norm	$\|\boldsymbol{s} > 0 \odot \boldsymbol{s}\|$
Guided Backpropagation [21]	$\left(\frac{\partial f^e(\boldsymbol{x})}{\partial \boldsymbol{z}}\right)_{l+1} > 0 \odot \frac{\partial f^e(\boldsymbol{x})}{\partial \boldsymbol{z}}$	Max	$\max(\boldsymbol{s})$
Integrated Gradient [22]	$f_l(\boldsymbol{x} - \bar{\boldsymbol{x}}) \odot \int_0^1 \frac{\partial f^e(\boldsymbol{x}+\alpha(\boldsymbol{x}-\bar{\boldsymbol{x}}))}{\partial \boldsymbol{z}}\, d\alpha$	Abs max	$\max(\|\boldsymbol{s}\|)$
SmoothGrad [20]	$\frac{1}{N}\sum_{n=1}^{N} \frac{\partial f^e(\boldsymbol{x}')}{\partial \boldsymbol{z}'}, \boldsymbol{x}' = \boldsymbol{x} + \mathcal{N}(\mu, \sigma_n^2)$	Abs sum	$\sum(\|\boldsymbol{s}\|)$

expectation as shown in Equation (S.4).

$$\sum_{\mathcal{D}} \phi = \frac{\sum_{\boldsymbol{r}} \left(L_e(r) - E(L_e(r))\right)}{|\boldsymbol{r}_{\mathcal{E}} \cup \boldsymbol{r}_{b^*}|} \tag{S.4}$$

Symmetry. The contribution scores of neuron d_n and d_m should be the same if they contribute equally to concept e.

If

$$L_e^{\langle \mathcal{S} \cup d_n \rangle}(r) = L_e^{\langle \mathcal{S} \cup d_m \rangle}(r), \forall \mathcal{S} \subseteq \mathcal{D} \backslash \{d_n, d_m\} \tag{S.5}$$

Then

$$\phi_{d_n} = \phi_{d_m} \tag{S.6}$$

where $\langle * \rangle$ is an operator keeping the neurons in the brackets, *i.e.*, $\mathcal{S} \cup d_n$ or $\mathcal{S} \cup d_m$, unchanged while randomizing others.

Dummy. If a neuron d has no contribution to concept e, which means d's individual contribution is zero and d also has no contribution when it collaborates with other neurons, d's contribution score should be zero.

If

$$L_e^{\langle \mathcal{S} \cup d \rangle}(r) = L_e^{\langle \mathcal{S} \rangle}(r), \forall \mathcal{S} \subseteq \mathcal{D} \backslash d \tag{S.7}$$

Then

$$\phi_d = 0 \tag{S.8}$$

Additivity. If L_e is a random forest including different decision trees, the Shapley value of neuron d of the random forest is the sum of the Shapley value of neuron d of each decision tree.

$$\phi_d = \sum_{t=1}^{T} \phi_d^t \tag{S.9}$$

where there are T decision trees.

4. Other Applications

We demonstrate more applications of HINT as follows.

4.1. Saliency Method Evaluation

With the emergence of various saliency methods, different sanity evaluation approaches have been proposed [1, 9, 28]. However, as most saliency methods only show responsible pixels on the input images, feature maps on hidden layers are not considered, which makes the sanity evaluation not comprehensive enough. For example, [1] proposed a sanity test by comparing the saliency map before and after cascading randomization of model parameters from the top to the bottom layers. Guided Backpropagation failed the test because its results remained invariant.

We propose to apply the concept classifier implemented with the target saliency method to identify the responsible regions on hidden layer feature maps for the sanity test. The target saliency method passes the sanity test if meaningful responsible regions can be observed. As shown in Figure S.1 (a), on the hidden layer features.8, when fewer layers are randomized, the responsible regions are more focused on the key features of the bird – its beak and tail, which means that Guided Backpropagation does reveal the salient region and Guided Backpropagation could pass the sanity test if hidden layer results are considered.

4.2. Explaining Adversarial Attack

Concept classifiers can also be applied to explain how the object in an adversarial attacked image is shifted to be another class for some types of attacks. As shown in Figure S.1 (b), we attack images of various classes to be *bird* using PGD [13] and apply the *bird* classifier to the attacked images' feature maps. The responsible regions for concept *bird* highlighted in those fake *bird* images imply that adversarial attack does not change all the content of the original object to be another class but captures some details of the original image where there exist shapes similar to *bird*. For example, in the image of a coffee mug where most shapes are round, adversarial attack catches the only pointed shape and attacks it to be like *bird*. Additionally, we find the attacked image still preserves features of the original class. In Figure S.1 (b), the result of applying *mammal* classifier

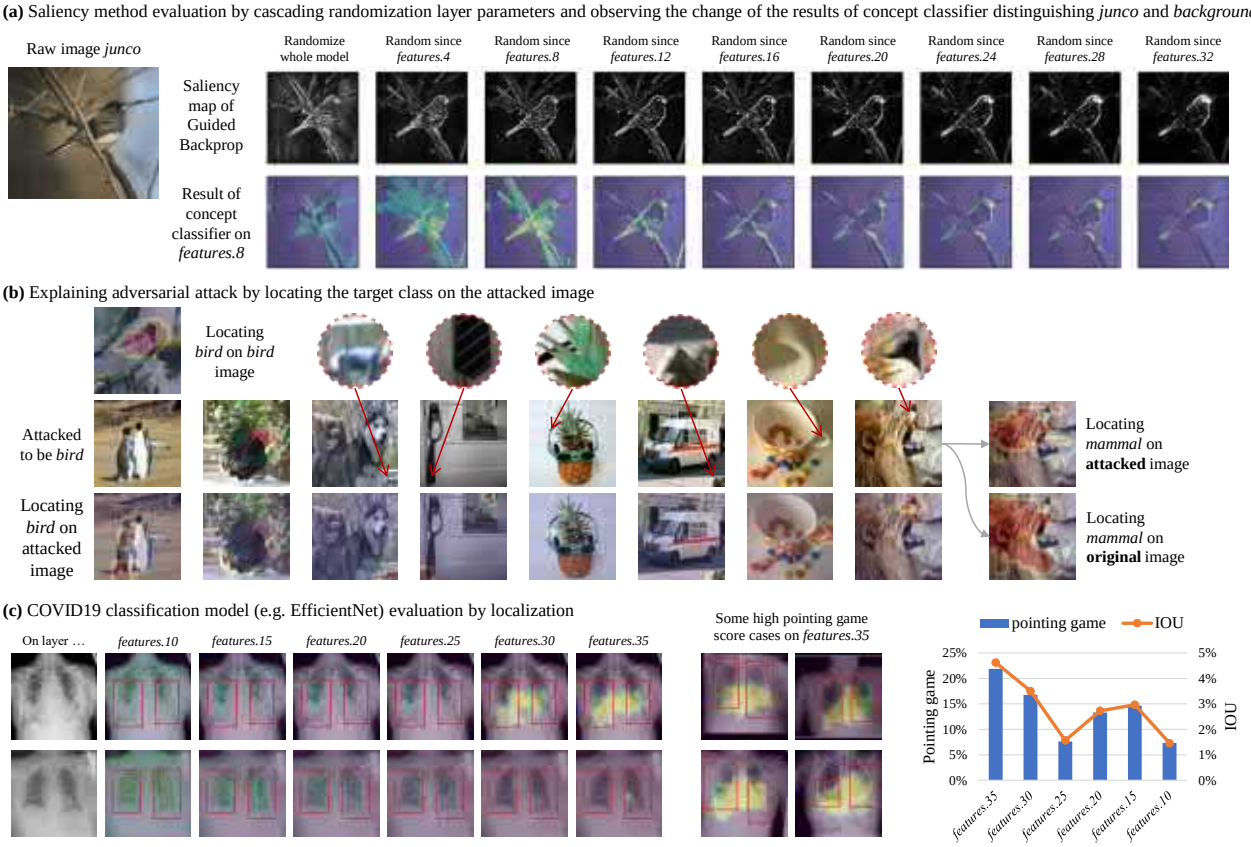

Figure S.1. Other applications of HINT. (a) Saliency method evaluation. See Section 4.1. (b) Explaining adversarial attack. See Section 4.2. (c) Evaluation of COVID19 classification model. See Section 4.3.

on the attacked lion image shows the most parts of the lion face are highlighted, while the result of applying *mammal* classifier on the original lion image shows a similar pattern.

Upon above observations, we design a quantitative evaluation on the faithfulness of our explanations. First, we attack 300 images of other categories excluding *bird* to be *birds* based on VGG19 model. Then, we use a *bird* classifier to find the regions corresponding to the adversarial features of *bird* on the attacked images. By visual inspection, we find most regions contain point shapes. Based on the regions, we train an adversarial attacked "*bird*" classifier ("ad clf"). Finally, we use the "ad clf" to perform the WSOL task on real *bird* images. The accuracy is 64.3% (for true *bird* classifier, it is 70.1%), indicating HINT captures the adversarial *bird* features and validates the explanation: some kind of adversarial attacks may be caused by attacking the similar shapes of the target class.

4.3. COVID19 Classification Model Evaluation

Applying deep learning to the detection of COVID19 in chest radiographs has the potential to provide quick diagnosis and guide management in molecular test resource-limited situations. However, the robustness of those models remains unclear [4, 8]. We do not know whether the model decisions rely on confounding factors or medical pathology in chest radiographs. To tackle the challenge, object localization by HINT can be used to see whether the identified responsible regions overlap with the lesion regions drawn by doctors. With the COVID19 dataset from SIIM-FISABIO-RSNA COVID-19 Detection competition [10],

Table S.2. Pointing game (pointing) and IoU of the localization results of different models on the chest radiographs of COVID19 cases with typical symptoms.

Model	Layer	pointing	IoU
EfficientNet [24]	features.35	21.8%	4.6%
DenseNet161 [6]	denseblock4	94.1%	18.2%
Inception v3 [23]	Mixed_6c	17.3%	3.2%
ResNet50 [5]	layer3.3	15.7%	2.9%
ShuffleNet v2 [12]	stage3.5	22.2%	3.8%
SqueezeNet1 [7]	features.9	0%	0%
VGG19 [19]	features.40	9.9%	1.6%

we trained models used by high-ranking teams and other baseline models for classification. The localization results of COVID19 cases with typical symptoms by EfficientNet [24] are shown in Figure S.1 (c). As you can see, the pointing game and IoU are not high. Many cases having low pointing game and IoU values show that the model does not focus on the lesion region, while for the cases with high pointing game and IoU values, further investigation is still required to see whether they capture the medical pathology features or they just accidentally focus on the area of the stomach.

Figure S.3 illustrates results of other models and Table S.2 quantitatively compares the different models by metrics of pointing game (pointing) and IoU. The accuracy values indicate that the hidden layers of SqueezeNet1 may fail to learn the concept of COVID19 pulmonary lesion. This can also be observed from Figure S.3 that SqueezeNet1 locates background regions. Note that although the pointing game score and IoU of DenseNet161 are very high, it is still possible that DenseNet161 fails to learn the concept of COVID19 pulmonary lesion as it highlights all the regions (see Figure S.3).

5. Identification of Responsible Neurons to Hierarchical Concepts

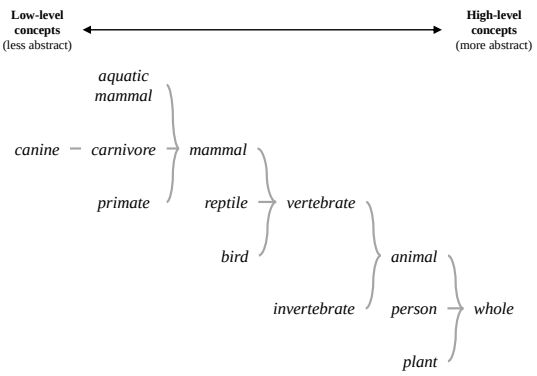

Figure S.2. A hierarchy of concepts.

5.1. More Neuron-concept Associations

This section illustrates more associations between neurons and concepts. The Sankey diagram in Figure S.4 shows the top-10 responsible neurons on layer features.30 of VGG19 to each concept in the hierarchy (see Figure S.2). And the Sankey diagram in Figure S.5 shows the case on layer layer3.5 of ResNet50.

Different layers. Figure S.6 shows the top-10 responsible neurons on different layers on VGG19 to concepts of *mammal*, *bird*, and *reptile*.

Different models. Figure S.7 shows top-10 responsible neurons on layer features.26 of VGG16, layer3.5 of ResNet50, and Mixed_6b of Inception v3 to concepts of *animal*, *person*, and *plant*.

5.2. Contribution Scores (Shapley Values) of Neurons to Concepts.

Concepts of different levels. The bar charts in Figure S.8, S.9, S.10, and S.11 show the contribution scores (Shapley values) of neurons on layer features.30 of VGG19 to concepts of *animal*, *vertebrate*, *mammal*, and *carnivore* respectively. As we can see, the 445^{th} neuron has the highest contribution to all the concepts.

Concepts of the same level. The bar charts in Figure S.14, S.15, S.16 show the contribution scores (Shapley values) of neurons on layer Mixed_6b of Inception v3 to concepts of *animal*, *person*, and *plant* respectively. There are 768 neurons on layer Mixed_6b in total. For *animal*, there are 711 neurons with contribution scores larger than zero. For *person*, the number is 615. And for *plant*, the number is 387. This indicates that there are less neurons responsible for *plant*, which may reflect the bias of the training data that only few categories of *plants* were included and *plant* images only take a small percentage of the whole dataset.

Different models. The bar charts in Figure S.8, S.12, S.13, and S.14 show the the contribution scores (Shapley values) of neurons on different layers of VGG19, VGG16, ResNet50, and Inception v3 to the concept of *animal* respectively. As we can see, the drop of the neurons' contribution scores of ResNet50 is less sharp compared with VGG16 and Inception v3, which means that the neurons of ResNet50 more rely on collaboration to detect *animal*.

5.3. Activation Maps of Multimodal Neurons

As shown in S.4, the 445^{th} neuron on layer features.30 of VGG19 contribute strongly to multiple concepts, indicating it is multimodal. We show the activation maps of the 445^{th} neuron on images of *animal* (see Figure S.17), *mammal* (see Figure S.18), and *canine* (see Figure S.19) respectively.

Also, we show the activation maps of the 199^{th} neuron on layer features.30 of VGG19 which contributes strongly to both *bird* and *car* in Figure S.20 and S.21. The results indicate the 199^{th} neuron activates the head of *bird* while deactivating the wheels of *car*. Therefore, it is multimodal and can detect both *bird* and *car*.

6. Weakly Supervised Object Localization

6.1. Localization Accuracy on CUB-200-2011

As shown in Table S.3, the localization accuracy of HINT is compared with existing methods on the CUB-200-

2011 [25] dataset. We train *animal* classifiers with 10%, 20%, 40%, 80% neurons sorted and selected by Shapley values using different models. Besides, we add a baseline tests of HINT where the neurons are randomly chosen. The results verify that Shapley values are good measurements of neuron contributions and show that different models might have different learning modes: ResNet50 and Inception v3 rely more on neurons' collaboration while neurons in VGG16 work more independently. This can be observed from the Localization Accuracy values. The Localization Accuracy of ResNet50 and Inception v3 increase steadily when more neurons are included in the concept classifier while the Localization Accuracy of VGG16 only has minor increase when more neurons are added.

Table S.3. Comparison of Localization Accuracy on CUB-200-2011. * indicates fine-tuning on CUB-200-2011. "rand" indicates the neurons are randomly selected.

	VGG16	ResNet50	Inception v3
CAM* [33]	34.4%	42.7%	43.7%
ACoL* [31]	45.9%	-	-
SPG* [32]	-	-	46.6%
ADL* [3]	52.4%	62.3%	53.0%
DANet* [27]	52.5%	-	49.5%
EIL* [14]	57.5%	-	-
PSOL* [29]	66.3%	70.7%	65.5%
GCNet* [11]	63.2%	-	-
RCAM* [2]	59.0%	59.5%	-
FAM* [15]	**69.3%**	**73.7%**	**70.7%**
Ours (10%)	**66.6%**	60.2%	49.0%
Ours (10%, rand)	56.2%	4.7%	14.2%
Ours (20%)	65.2%	67.1%	55.8%
Ours (20%, rand)	58.4%	35.9%	34.2%
Ours (40%)	61.3%	77.3%	52.8%
Ours (40%, rand)	60.5%	68.6%	48.1%
Ours (80%)	64.8%	**80.2%**	**56.2%**
Ours (80%, rand)	61.5%	76.5%	53.0%

6.2. Quantitative Results of Applying Concept Classifiers on ImageNet

In this section, because many images in ImageNet only have classification labels, we use the hidden layer saliency map as the mask of the target object. And we apply metrics of pointing game (pointing) [30], Spearman's correlation (spearman cor), and structure similarity index (SSMI) [26] to evaluate concept classifiers' performances on ImageNet. VGG19 is used for testing.

Images of different concepts. As shown in Table S.4, we apply *whole* classifier trained on layer features.30 to images of different concepts. The results indicate that the *whole* classifier can locate all the target objects as the concepts are all subcategories of *whole*. Also, we test the *mammal* classifier to images of other concepts which have no intersection with *mammal*, showing that the *mammal* classifier only responses to image contents of *mammal* (see Table S.5).

Table S.4. Apply *whole* classifier trained on layer features.30 to images of different concepts.

Images of	pointing	spearman cor	SSMI
whole	88.0%	52.2%	34.4%
person	34.0%	32.0%	26.5%
plant	60.4%	37.9%	24.6%
animal	81.9%	62.8%	38.1%
mammal	77.7%	63.4%	43.5%
bird	86.7%	60.3%	44.1%
reptile	68.5%	56.3%	35.8%
carnivore	82.2%	68.3%	42.4%
primate	82.6%	53.7%	36.9%
aquatic mammal	56.9%	57.0%	43.5%

Table S.5. Apply ***mammal* classifier** trained on layer features.30 to ***person*** and ***plant*** images.

Images of	pointing	spearman cor	SSMI
person	8.8%	6.4%	8.6%
plant	3.6%	9.3%	0.9%

Different layers. As shown in Table S.6, we apply *mammal* classifier trained on different layers to *mammal* images. The accuracy values increase as the layer goes higher, indicating the network can learn abstract concepts such as *mammal* on high layers.

Table S.6. Apply ***mammal* classifier** trained on different layers to ***mammal*** images.

Layer	pointing	spearman cor	SSMI
features.2	11.7%	4.9%	3.7%
features.7	13.0%	13.7%	6.1%
features.10	28.7%	30.5%	8.9%
features.14	35.1%	34.5%	9.7%
features.20	58.4%	45.3%	15.4%
features.25	67.8%	51.7%	25.3%
features.30	76.4%	59.8%	37.7%

6.3. Visualizations of Localization Results on ImageNet, CUB-200-2011, and PASCAL VOC

ImageNet. Figure S.22, S.23, S.24, S.25, and S.26 illustrate the localization results of applying *whole* classifier on images containing contents of *whole*, *plant*, *animal*, *bird*,

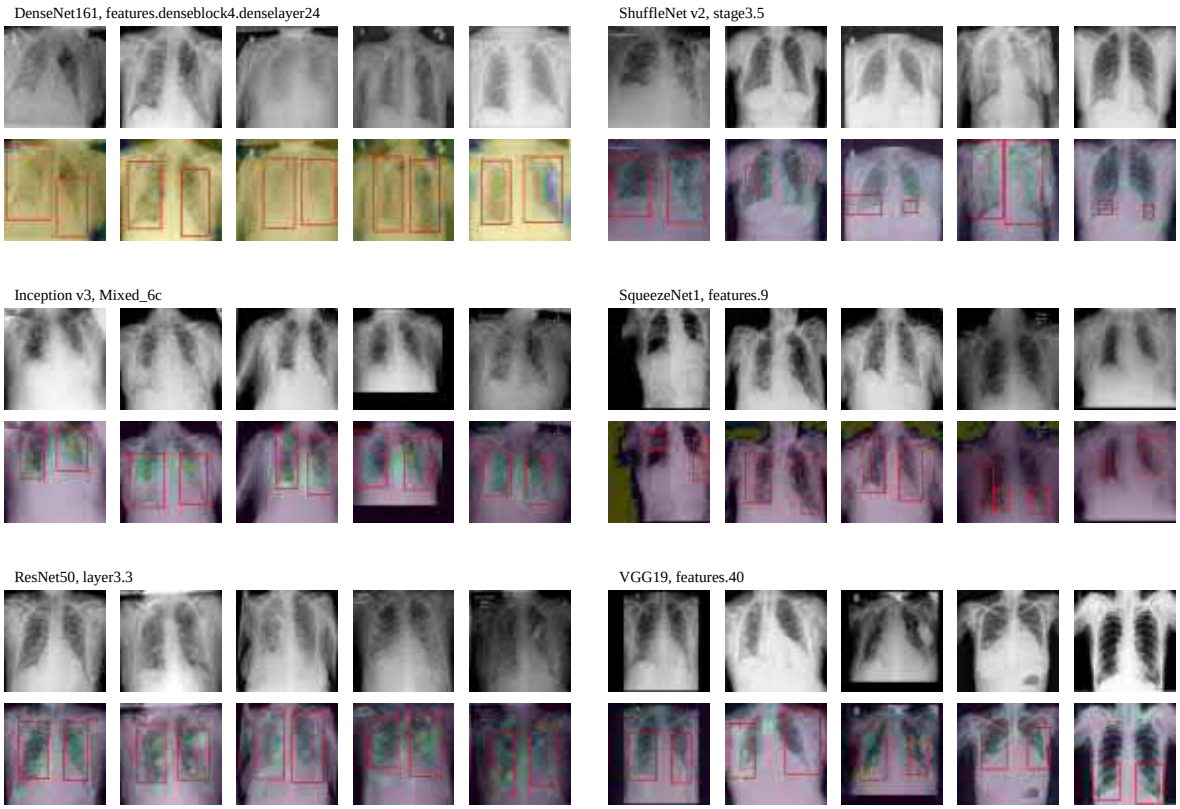

Figure S.3. Localization results of different models on the radiographs of COVID19 cases with typical symptoms. The red bounding boxes are the lesion regions drawn by doctors.

and *canine* respectively. Figure S.27, S.28, and S.29 illustrate the localization results of applying *mammal* classifier on images containing contents of *animal*, *mammal*, and *canine* respectively. Note that some *animals* are not *mammals* and cannot be located. Figure S.26, S.29, and S.30 illustrate the localization results of applying *whole*, *mammal*, and *carnivore* classifiers on images containing contents of *canine* respectively.

CUB-200-2011. Figure S.31, S.32, and S.33 illustrate the localization results of applying *animal* classifier trained on layer Mixed_6b of Inception v3, layer3.5 of ResNet50, and features.26 of VGG16 on the images from CUB-200-2011 respectively.

PASCAL VOC. Figure S.34 shows the sample images from PASCAL VOC used for test with masks indicating the target objects. Figure S.35, S.36, and S.37 illustrate the localization results of applying *whole*, *animal*, and *bird* classifiers on the sample images. The classifiers are all trained on layer features.30 of VGG19 with 20 neurons selected by Shapley values. The results indicate the unique advantage of HINT for object localization: a flexible choice of localization targets.

7. Ablation study

Illustration of the localization results of concept classifiers implemented with different saliency methods. Figure S.38 shows the localization results of concept classifiers using Guided Backpropagation, Vanilla Backpropgation, Gradient x Input, Integrated Gradients, and SmoothGrad on dataset CUB-200-2011. The illustration indicates that HINT is general and can be implemented with different saliency methods.

Illustration of the localization results of concept classifiers trained with neurons chosen by shap, clf_coef, and random Figure S.35, S.39, and S.40 show the localization results of applying *whole* classifiers on the sample images from PASCAL VOC, where the classifiers are trained on layer features.30 of VGG19 with 20 neurons selected by Shapley values (shap), selected by the coefficients of the linear classifier (clf_coef), or randomly selected (random). From observation, "shap" locates more *whole* objects and larger object contents, indicating that Shapley values are good measures of neurons' contributions to concepts.

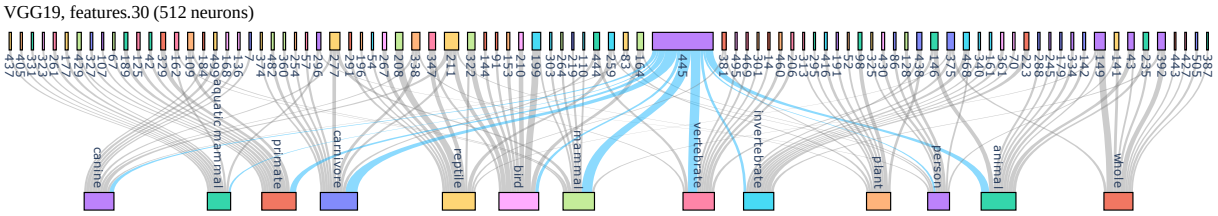

Figure S.4. Top-10 responsible neurons to concepts on layer “features.30” of VGG19.

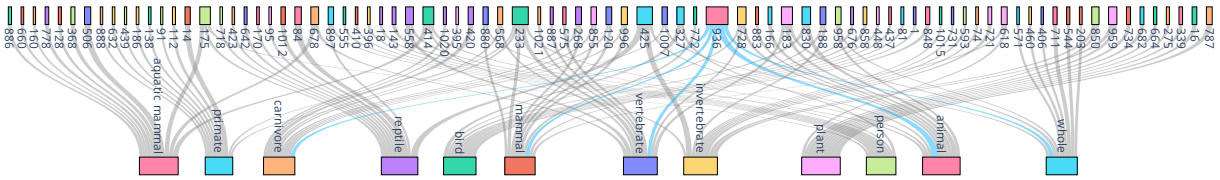

Figure S.5. Top-10 responsible neurons to concepts on layer “layer3.5” of ResNet50.

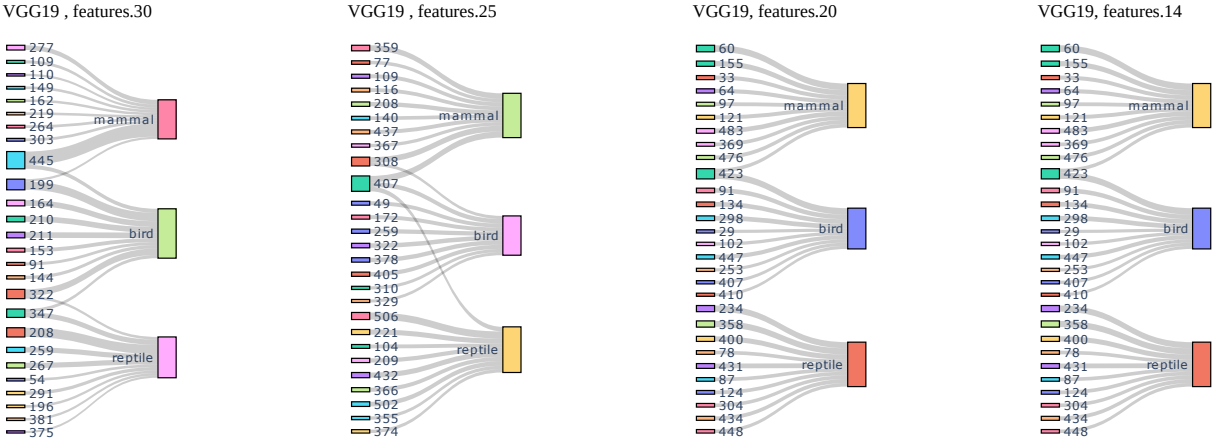

Figure S.6. Top-10 responsible neurons to concepts of *mammal*, *bird*, and *reptile* on different layer of VGG19.

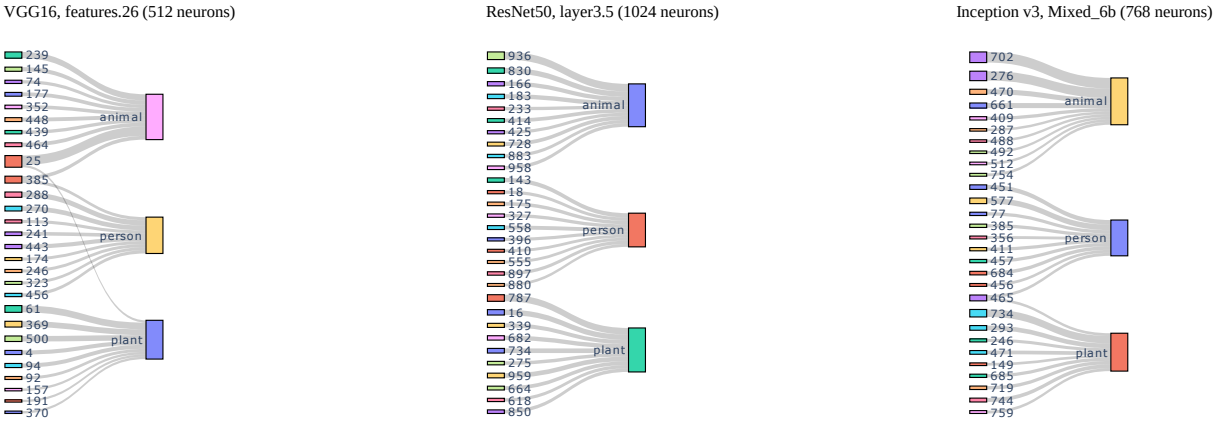

Figure S.7. Top-10 responsible neurons to concepts of *animal*, *person*, and *plant* of other models.

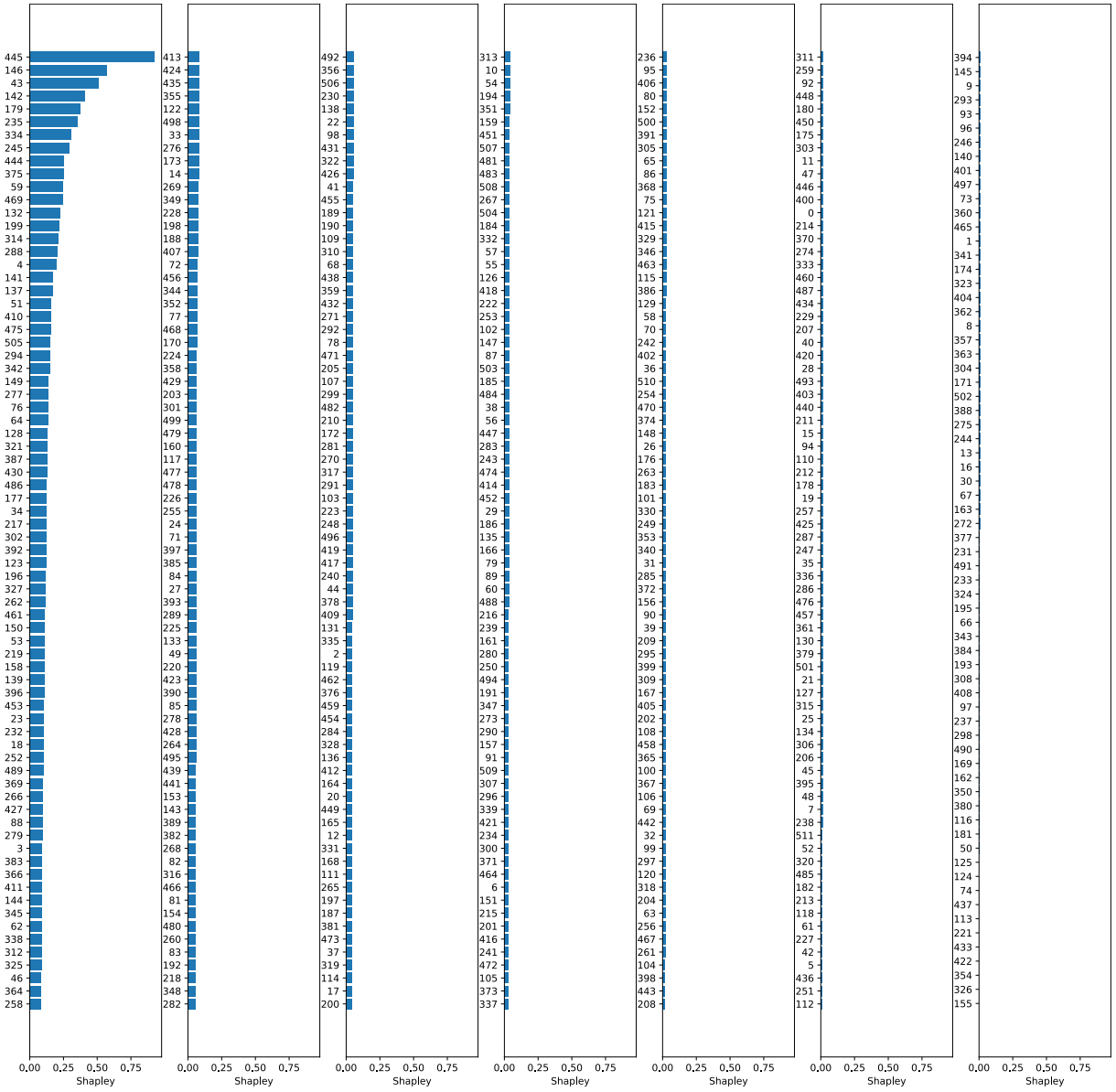

Figure S.8. Contribution scores (Shapley values) of neurons on layer features.30 of VGG19 to the concept of *animal*.

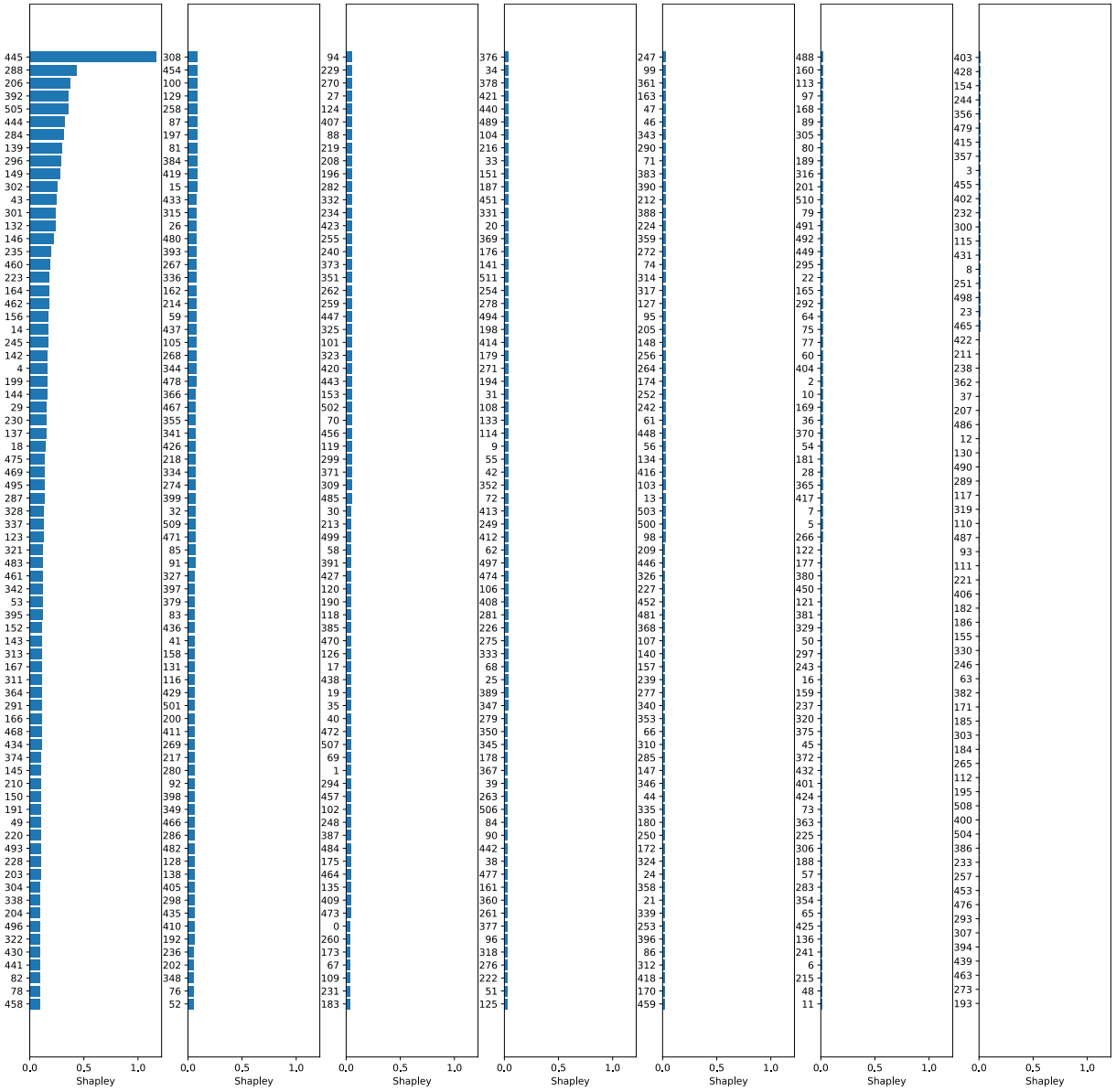

Figure S.9. Contribution scores (Shapley values) of neurons on layer features.30 of VGG19 to the concept of *vertebrate*.

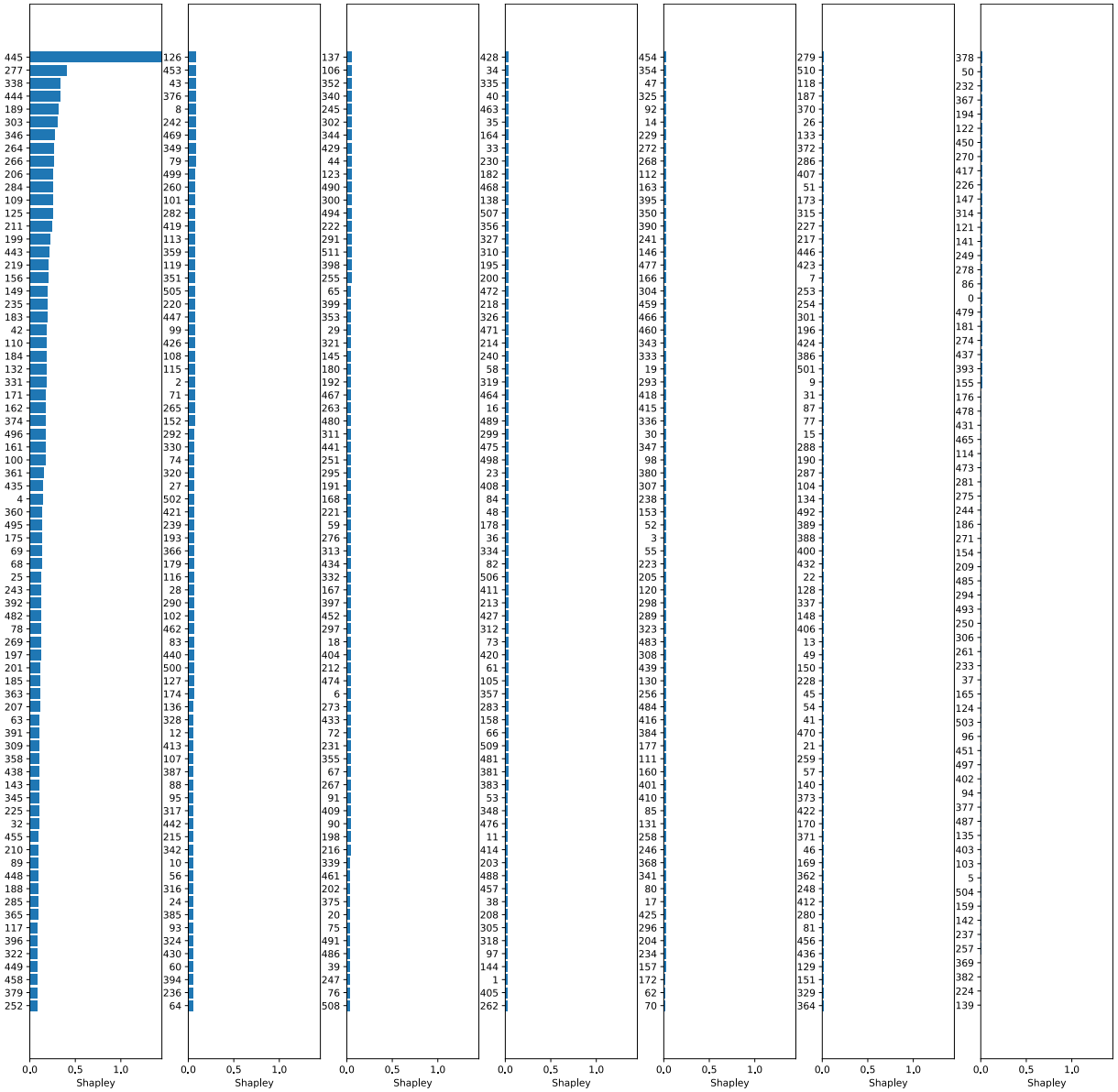

Figure S.10. Contribution scores (Shapley values) of neurons on layer features.30 of VGG19 to the concept of *mammal*.

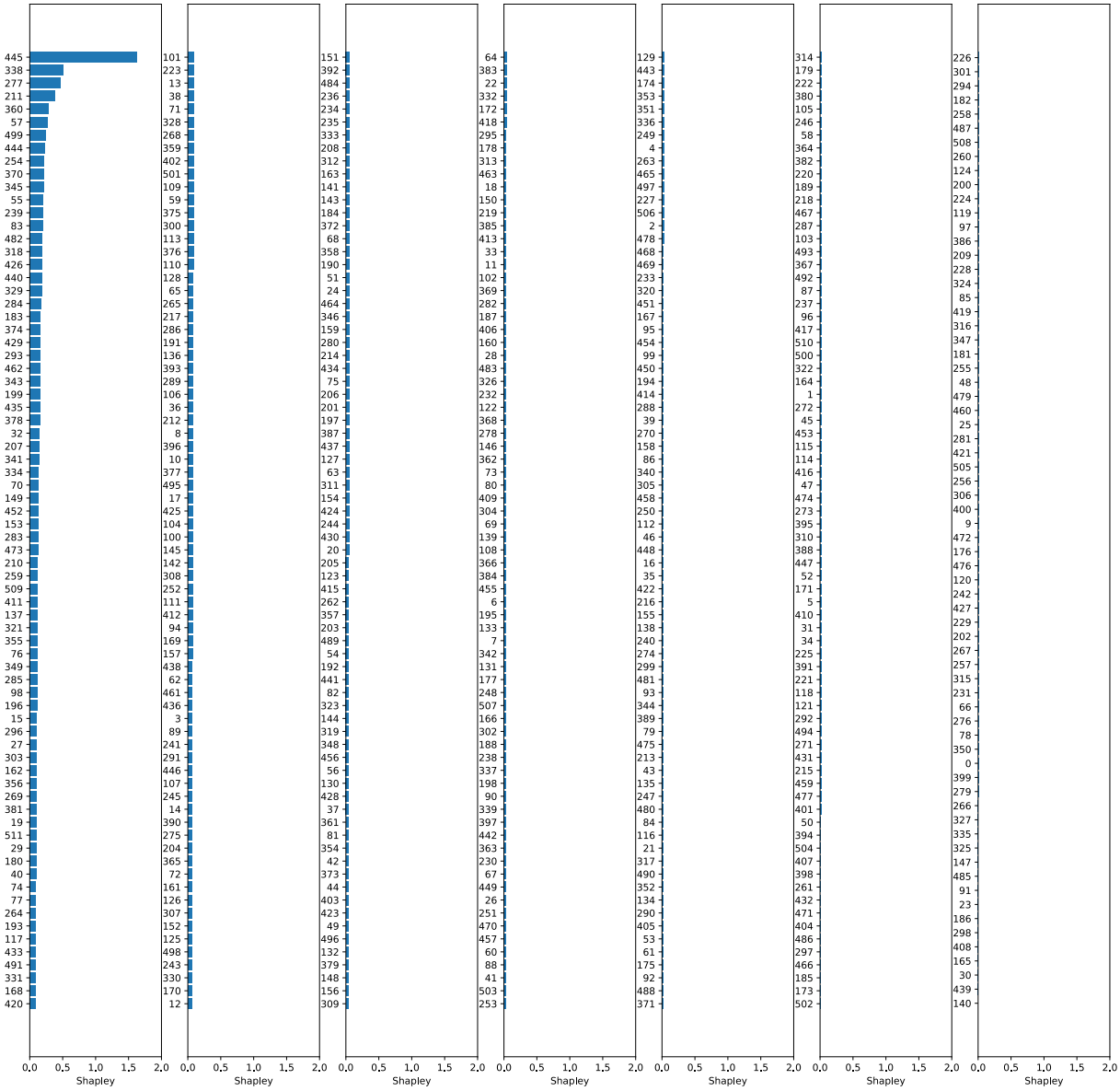

Figure S.11. Contribution scores (Shapley values) of neurons on layer features.30 of VGG19 to the concept of *carnivore*.

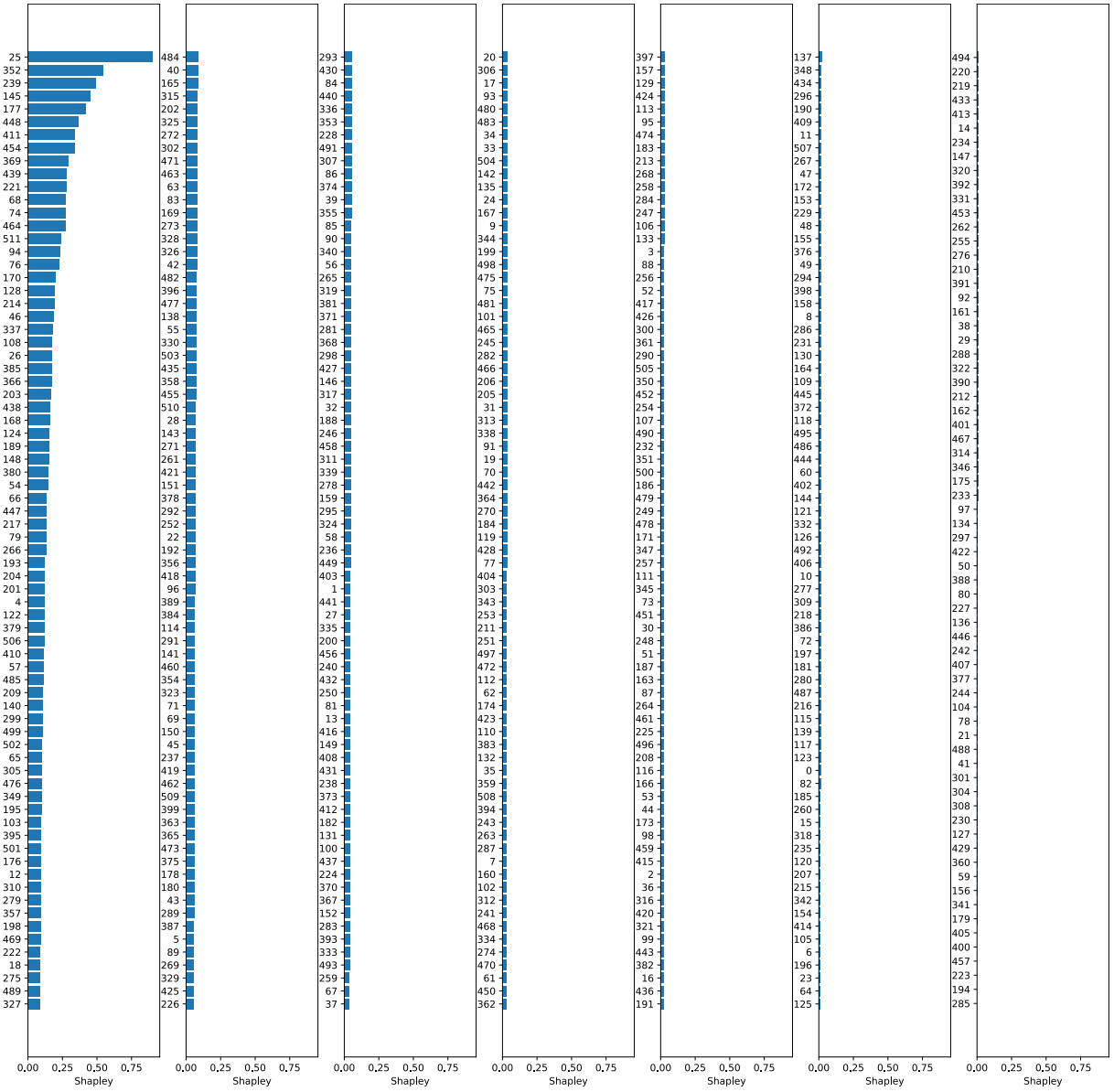

Figure S.12. Contribution scores (Shapley values) of neurons on layer features.26 of VGG16 to the concept of *animal*.

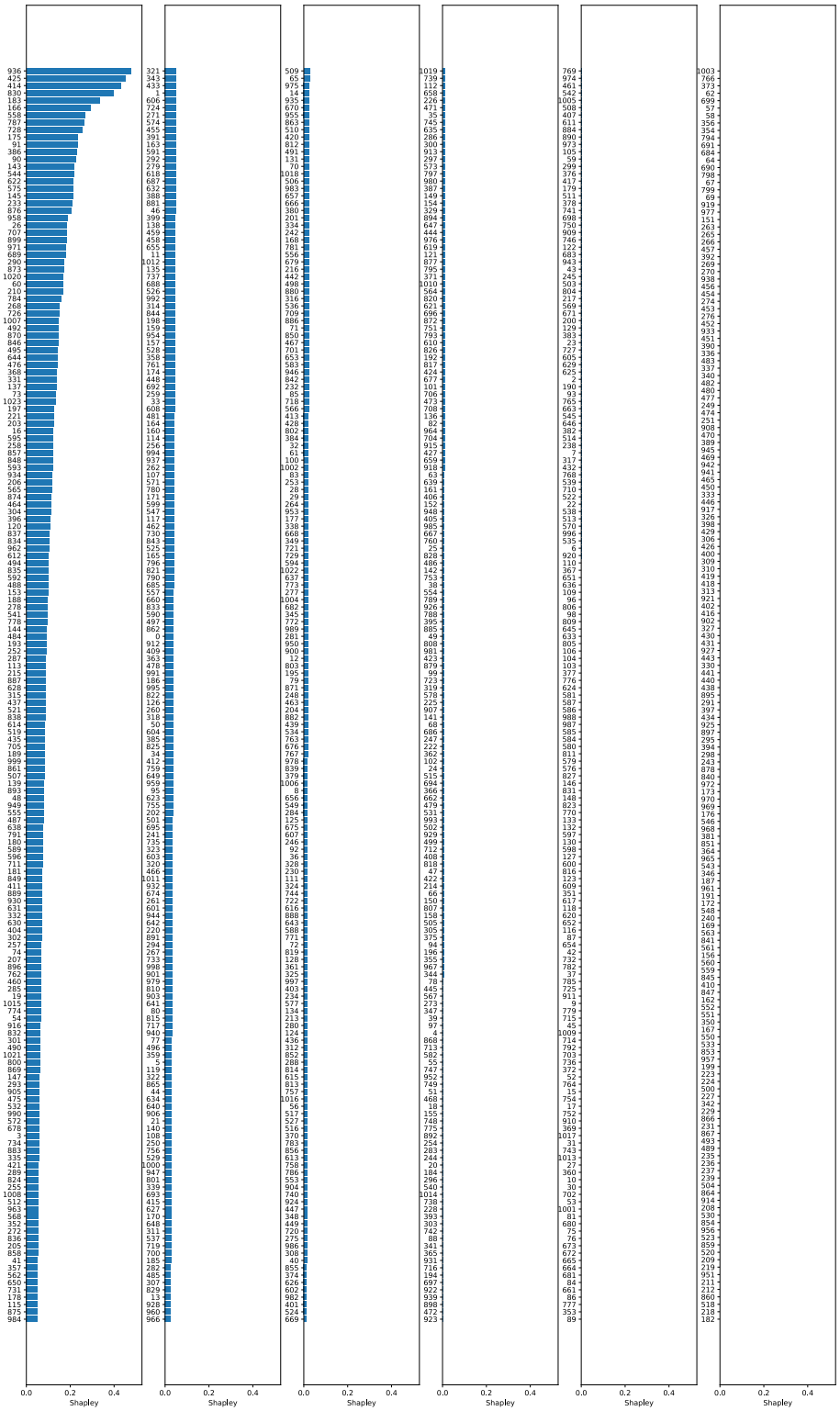

Figure S.13. Contribution scores (Shapley values) of neurons on layer layer3.5 of ResNet50 to the concept of *animal*.

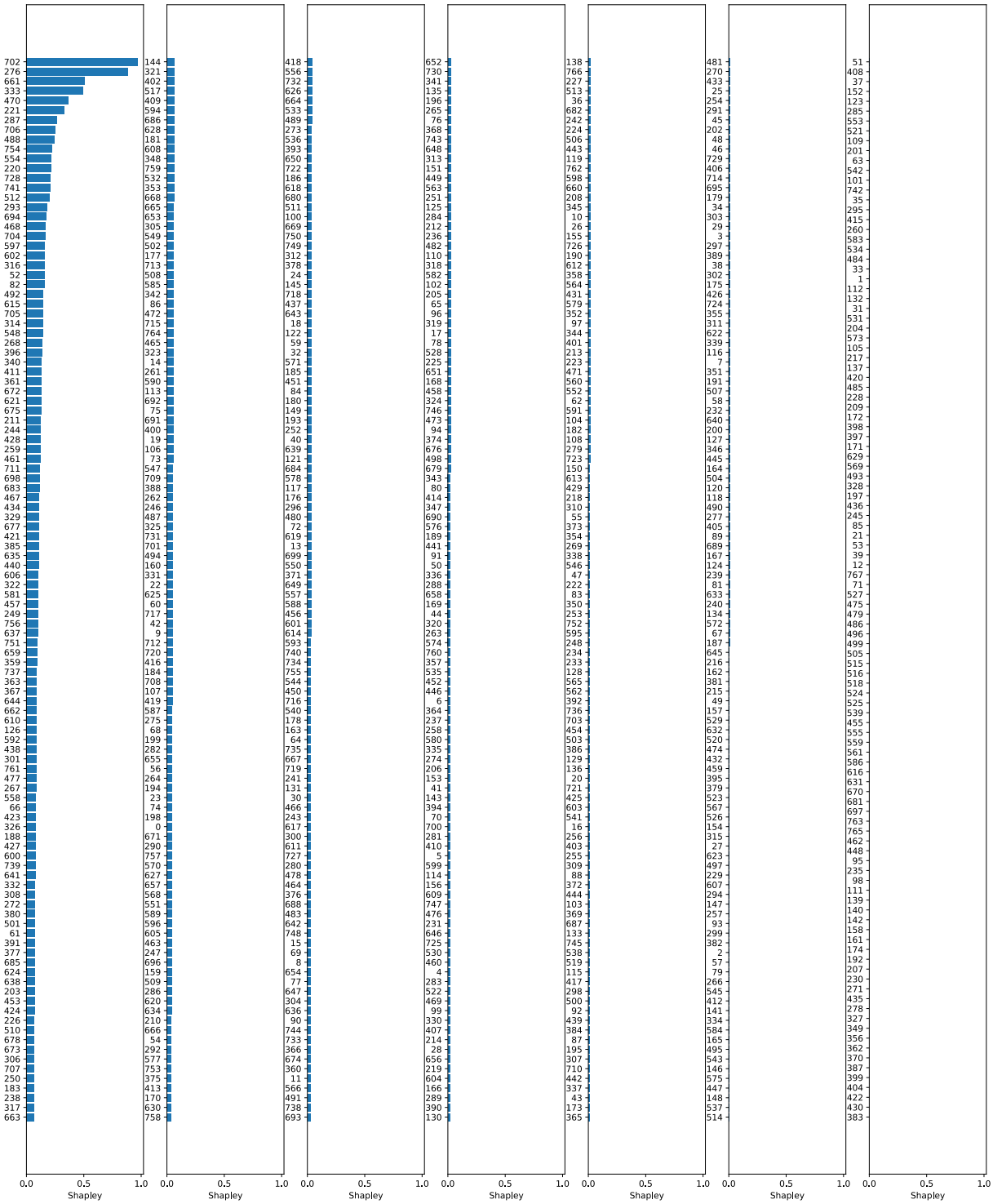

Figure S.14. Contribution scores (Shapley values) of neurons on layer Mixed_6b of Inception v3 to the concept of *animal*.

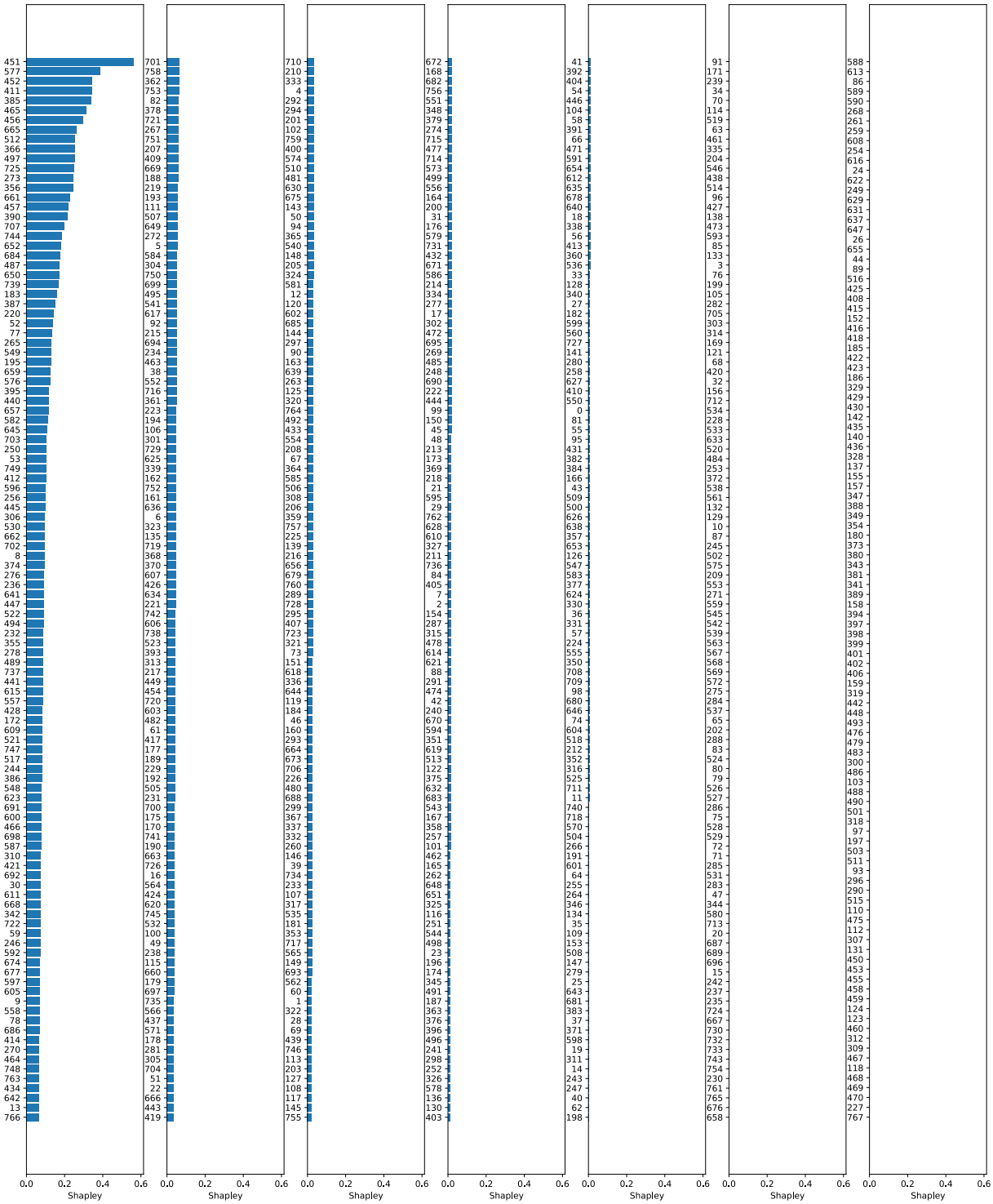

Figure S.15. Contribution scores (Shapley values) of neurons on layer Mixed_6b of Inception v3 to the concept of *person*.

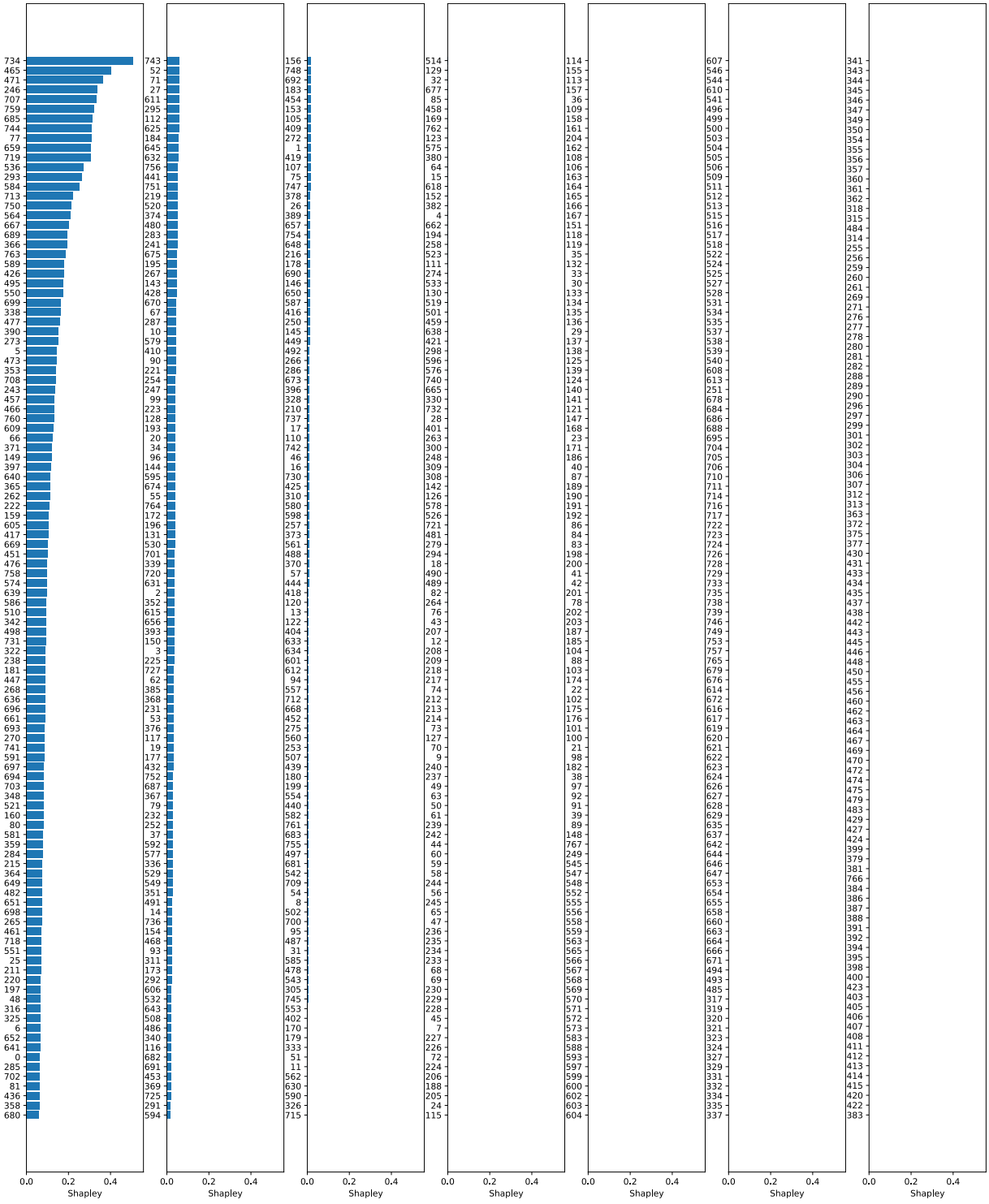

Figure S.16. Contribution scores (Shapley values) of neurons on layer Mixed_6b of Inception v3 to the concept of *plant*.

Figure S.17. Activation map of the 445^{th} neuron on *animal* images.

Figure S.18. Activation map of the 445^{th} neuron on *mammal* images.

Figure S.19. Activation map of the 445^{th} neuron on *canine* images.

Figure S.20. Activation map of the 199^{th} neuron on *bird* images.

Figure S.21. Activation map of the 199^{th} neuron on *car* images.

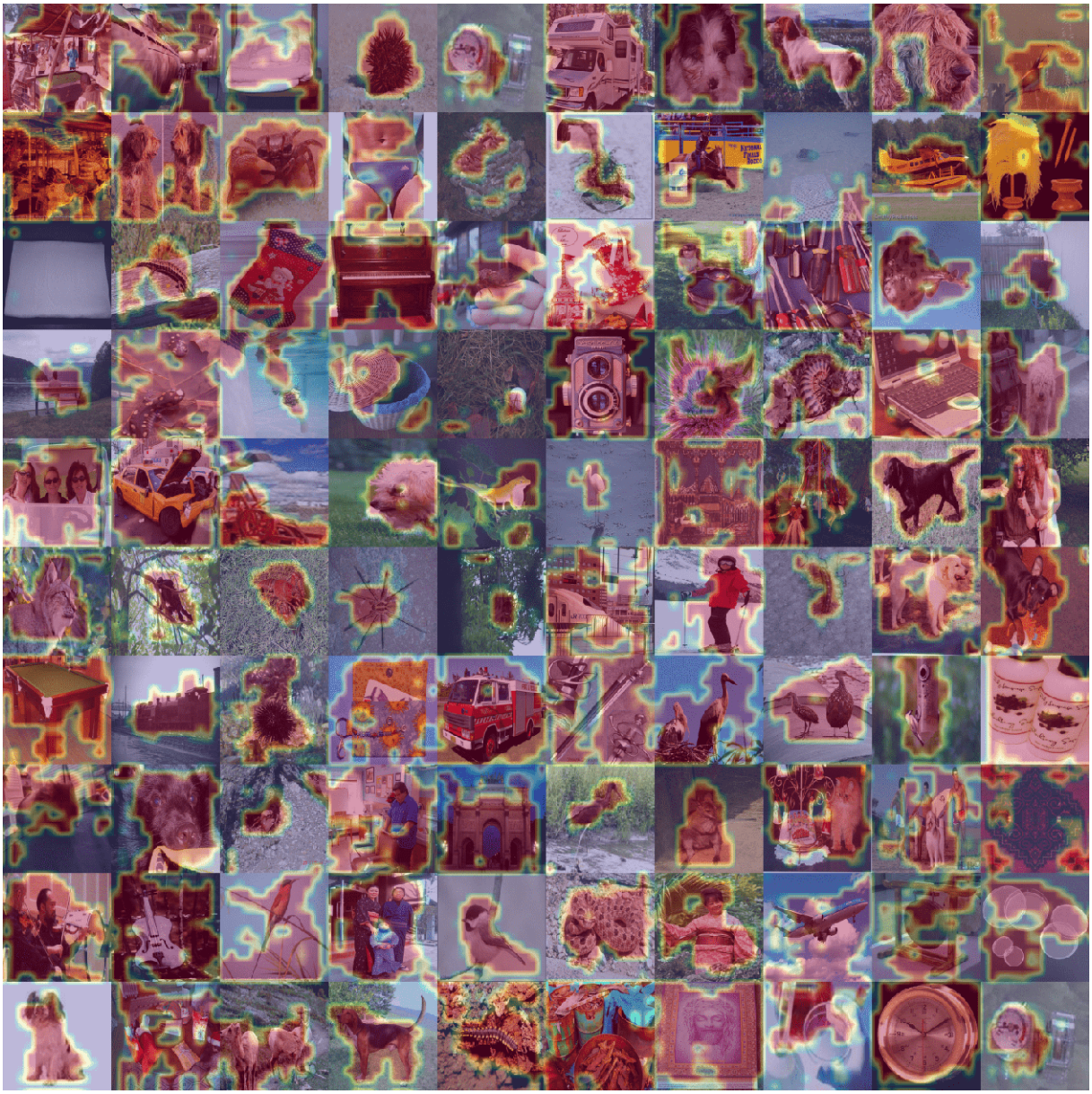

Figure S.22. Localization results of applying *whole* classifier on the images containing the concept of *whole* from ImageNet.

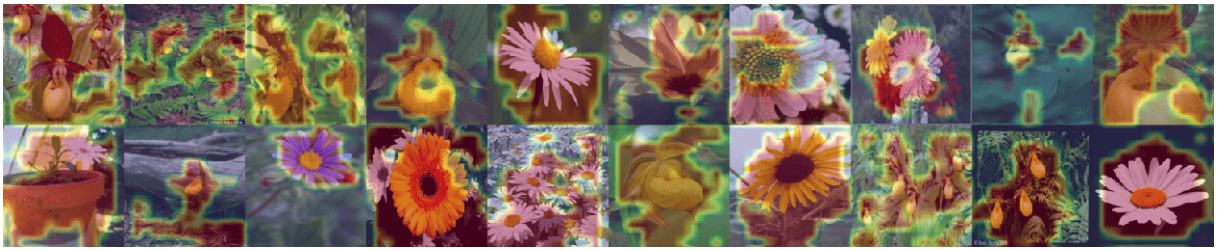

Figure S.23. Localization results of applying *whole* classifier on the images containing the concept of *plant* from ImageNet.

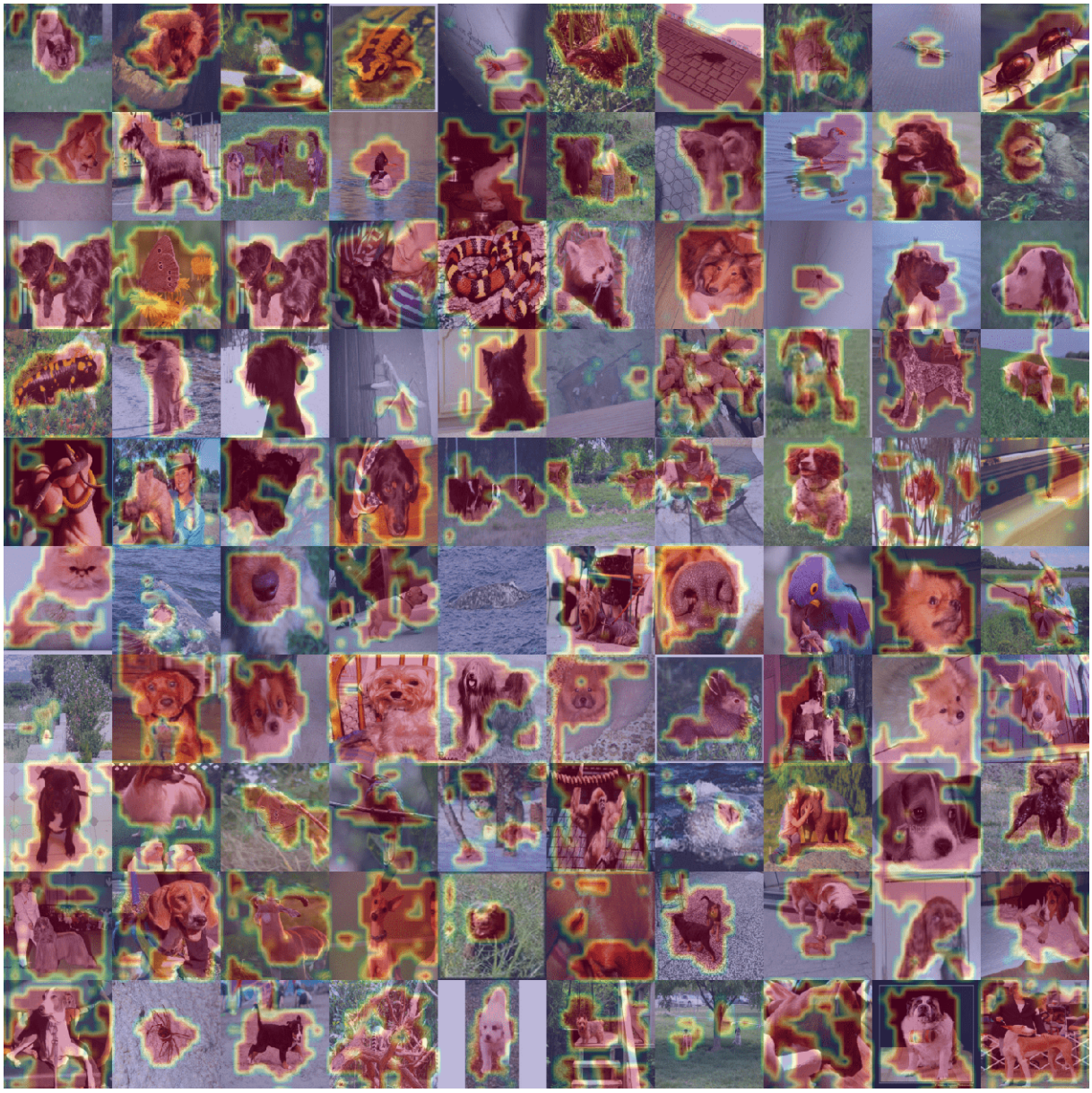

Figure S.24. Localization results of applying *whole* classifier on the images containing the concept of *animal* from ImageNet.

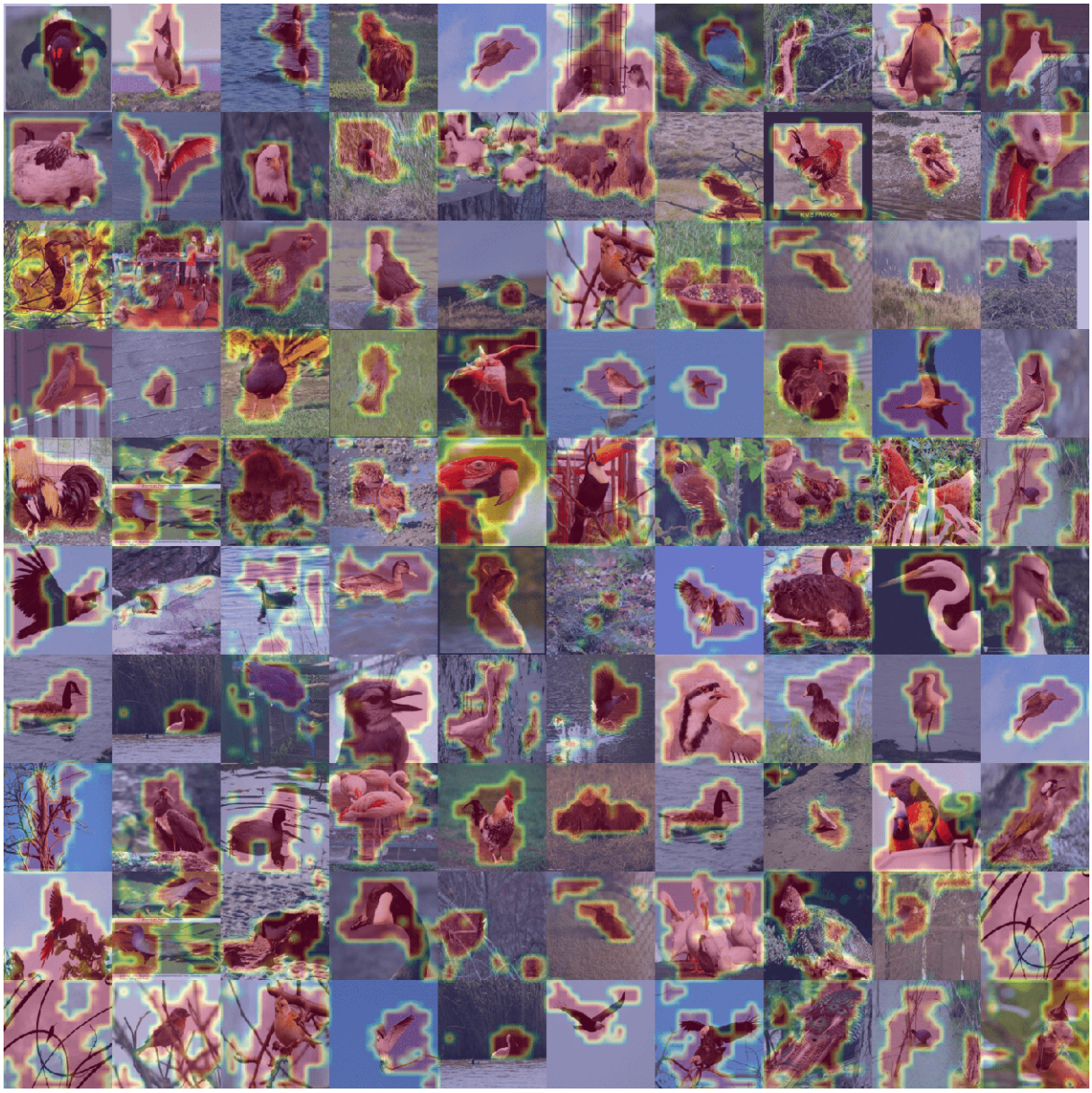

Figure S.25. Localization results of applying *whole* classifier on the images containing the concept of *bird* from ImageNet.

Figure S.26. Localization results of applying *whole* classifier on the images containing the concept of *canine* from ImageNet.

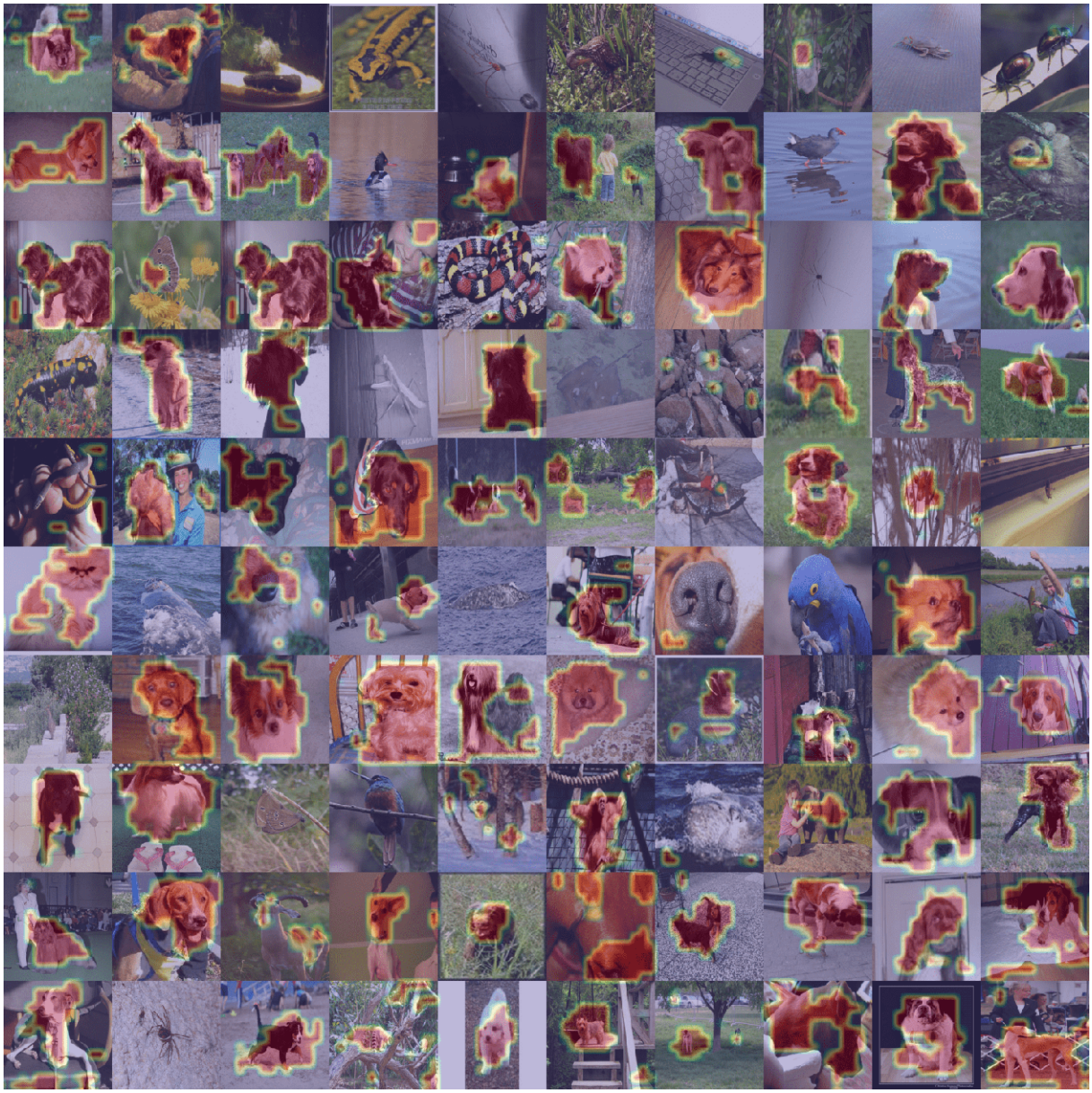

Figure S.27. Localization results of applying *mammal* classifier on the images containing the concept of *animal* from ImageNet. Note that some *animals* are not *mammals* and cannot be located.

Figure S.28. Localization results of applying *mammal* classifier on the images containing the concept of *mammal* from ImageNet.

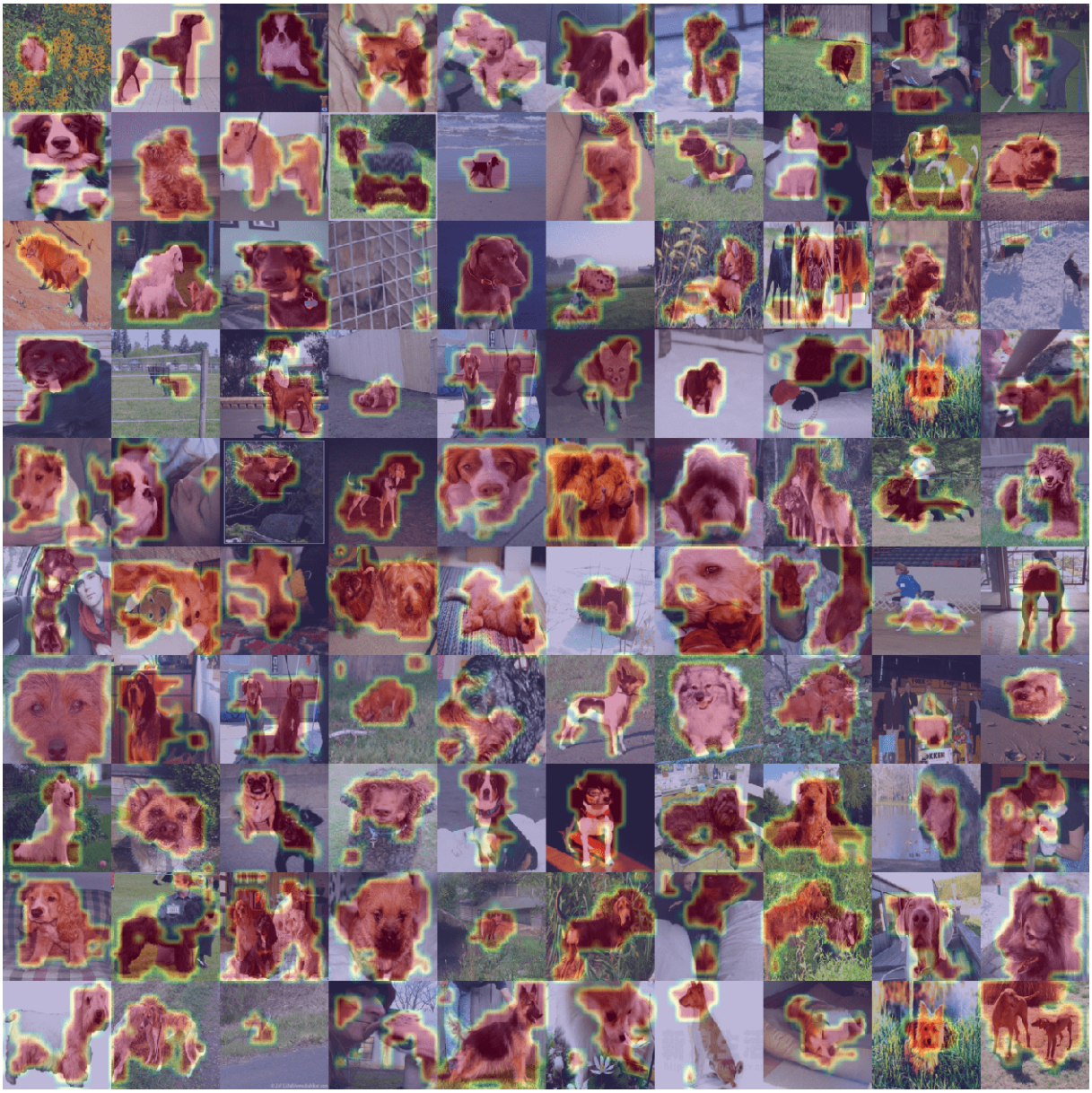

Figure S.29. Localization results of applying *mammal* classifier on the images containing the concept of *canine* from ImageNet.

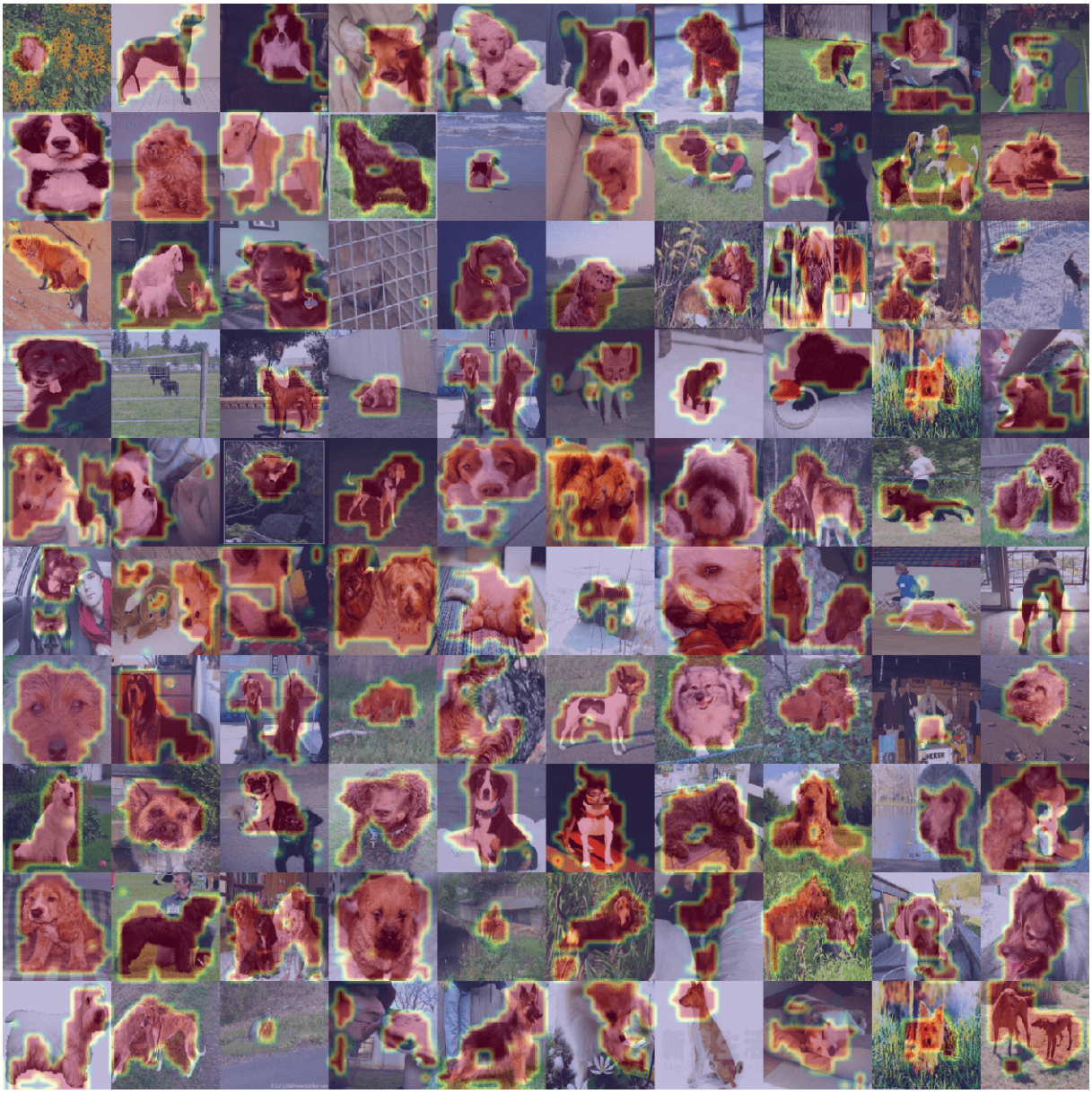

Figure S.30. Localization results of applying *carnivore* classifier on the images containing the concept of *canine* from ImageNet.

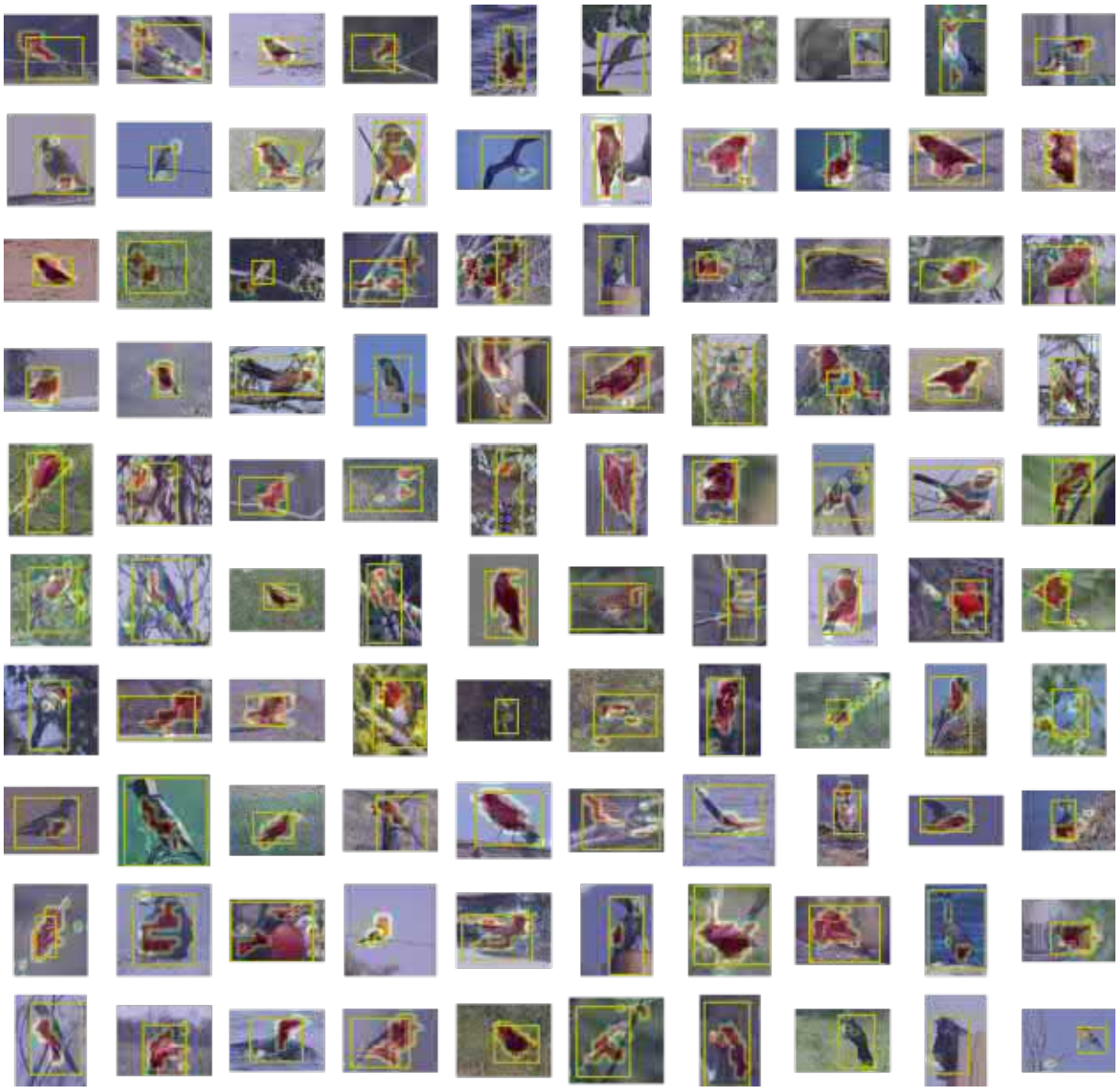

Figure S.31. Localization results of applying *animal* classifier trained on layer Mixed_6b of Inception v3 on the images from CUB-200-2011. The yellow bounding boxes are the groundtruth.

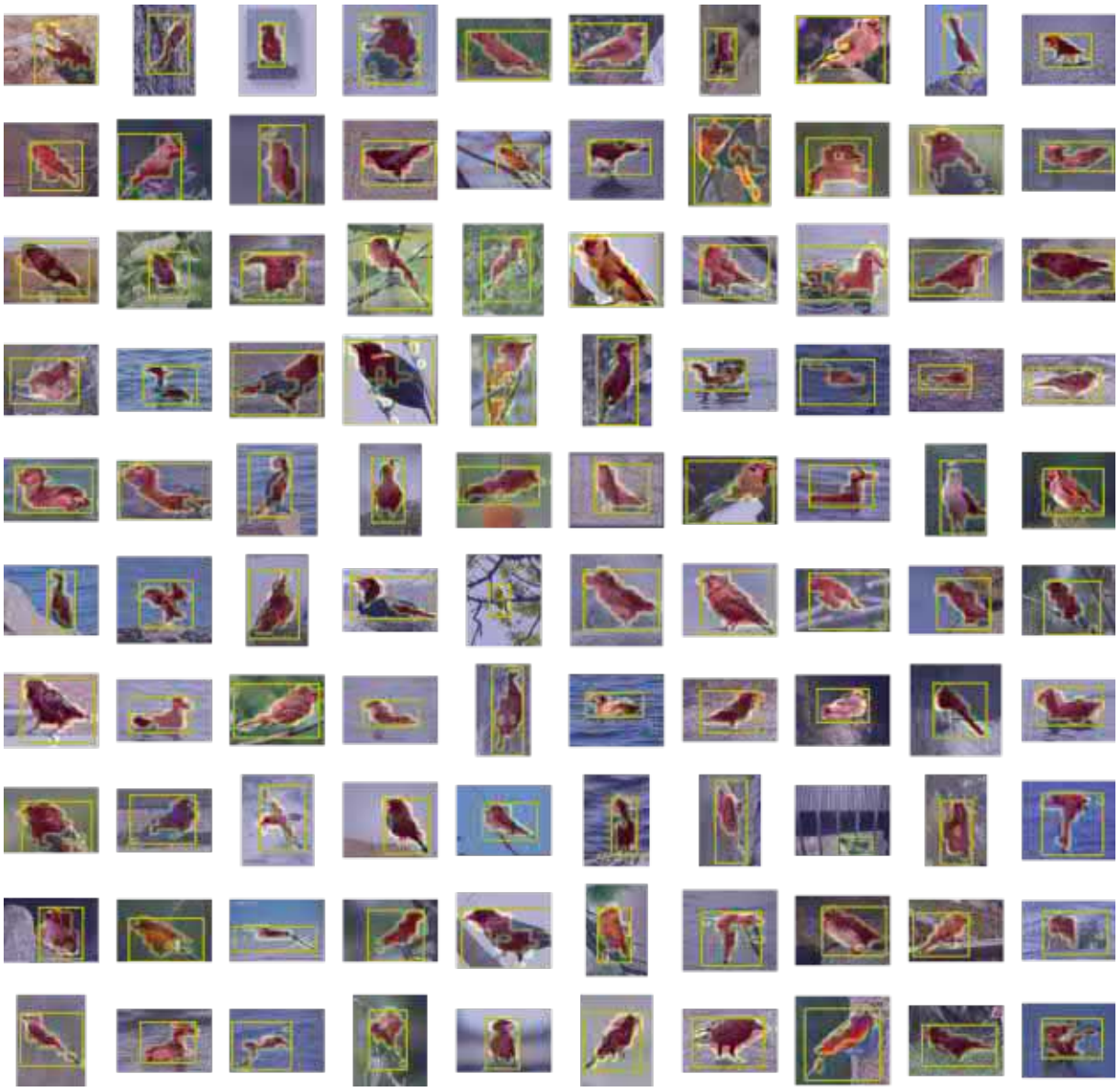

Figure S.32. Localization results of applying *animal* classifier trained on layer layer3.5 of ResNet50 on the images from CUB-200-2011. The yellow bounding boxes are the groundtruth.

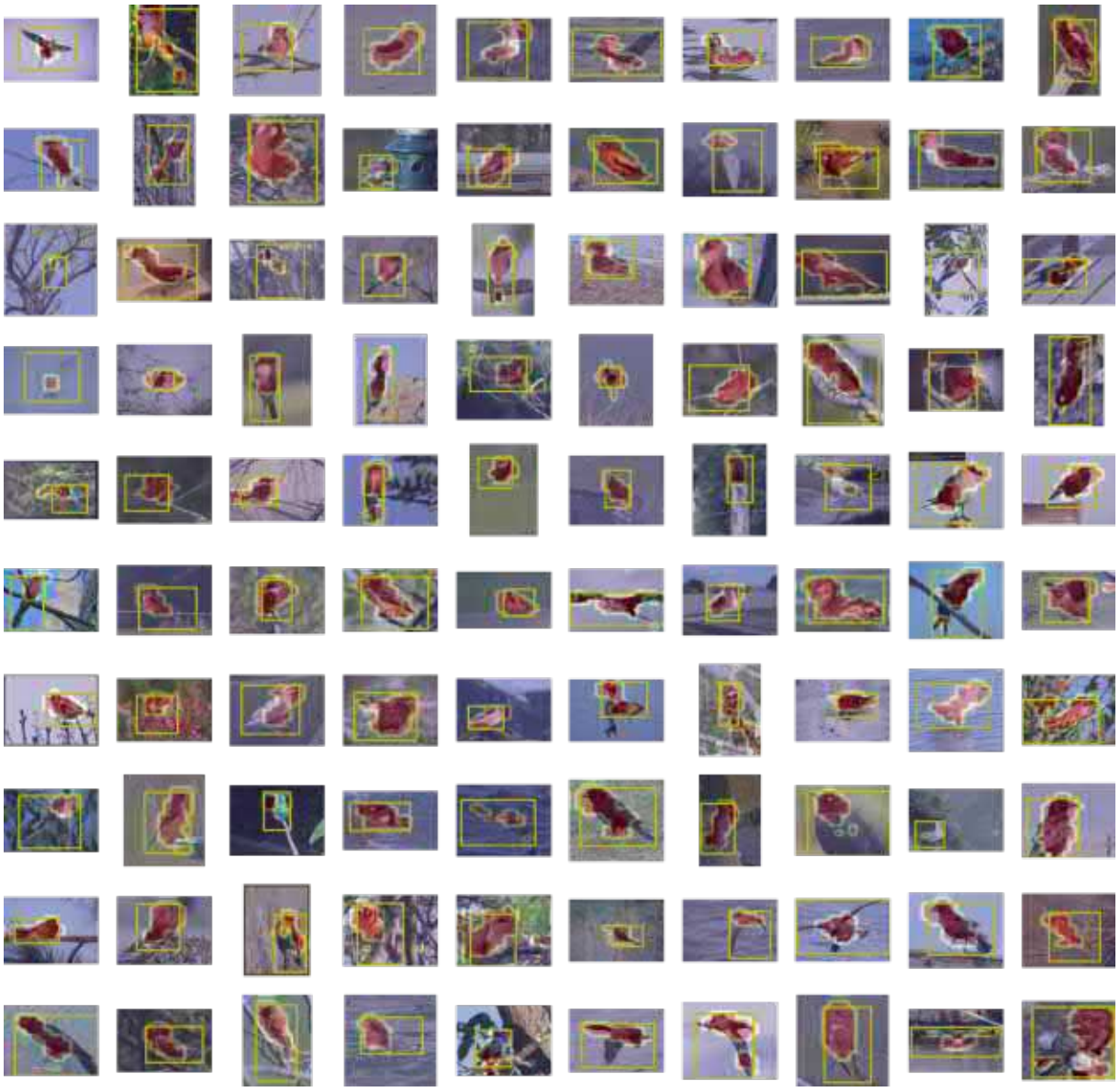

Figure S.33. Localization results of applying *animal* classifier trained on layer features.26 of VGG16 on the images from CUB-200-2011. The yellow bounding boxes are the groundtruth.

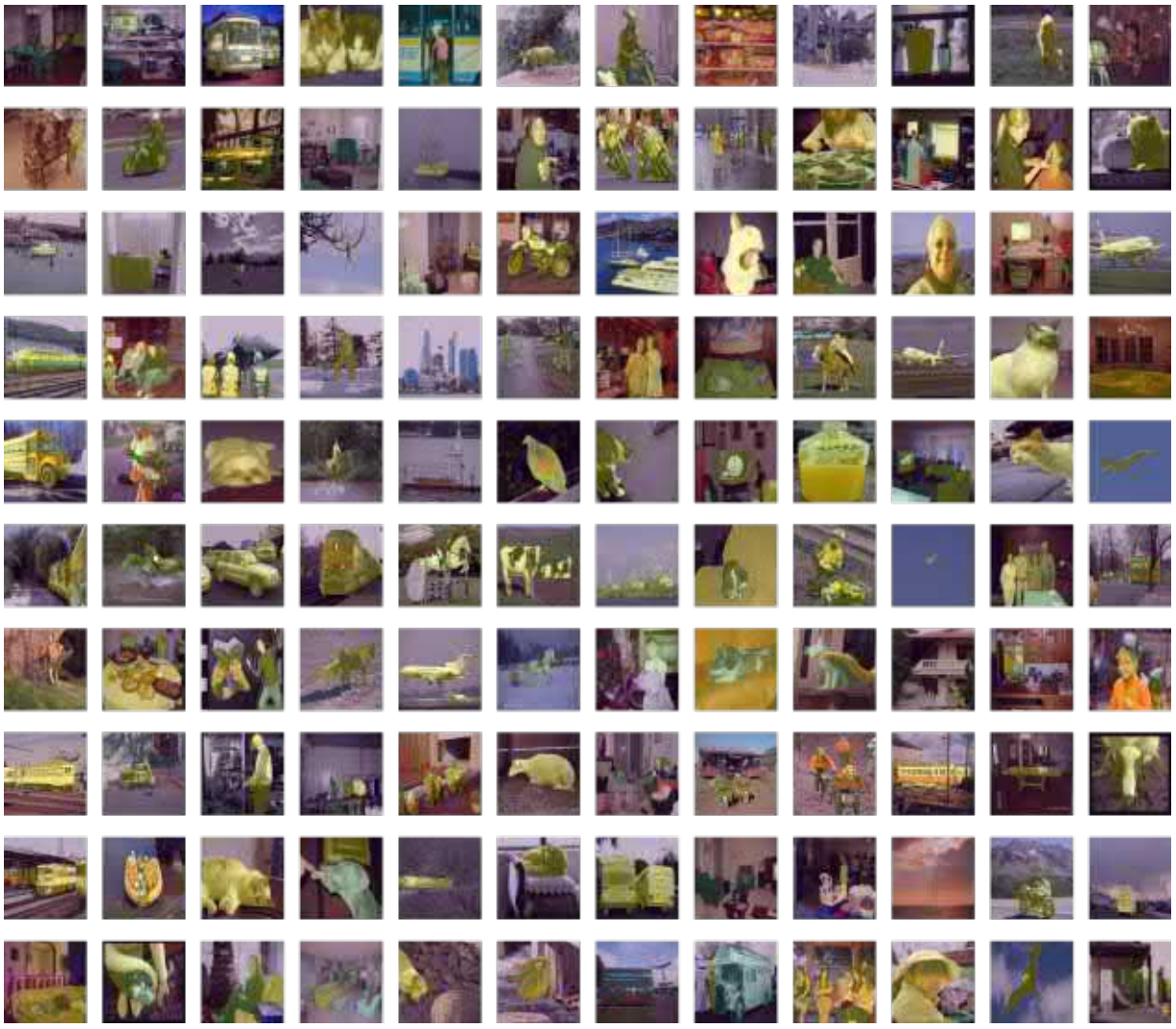

Figure S.34. Sample images from PASCAL VOC with masks indicating the target objects.

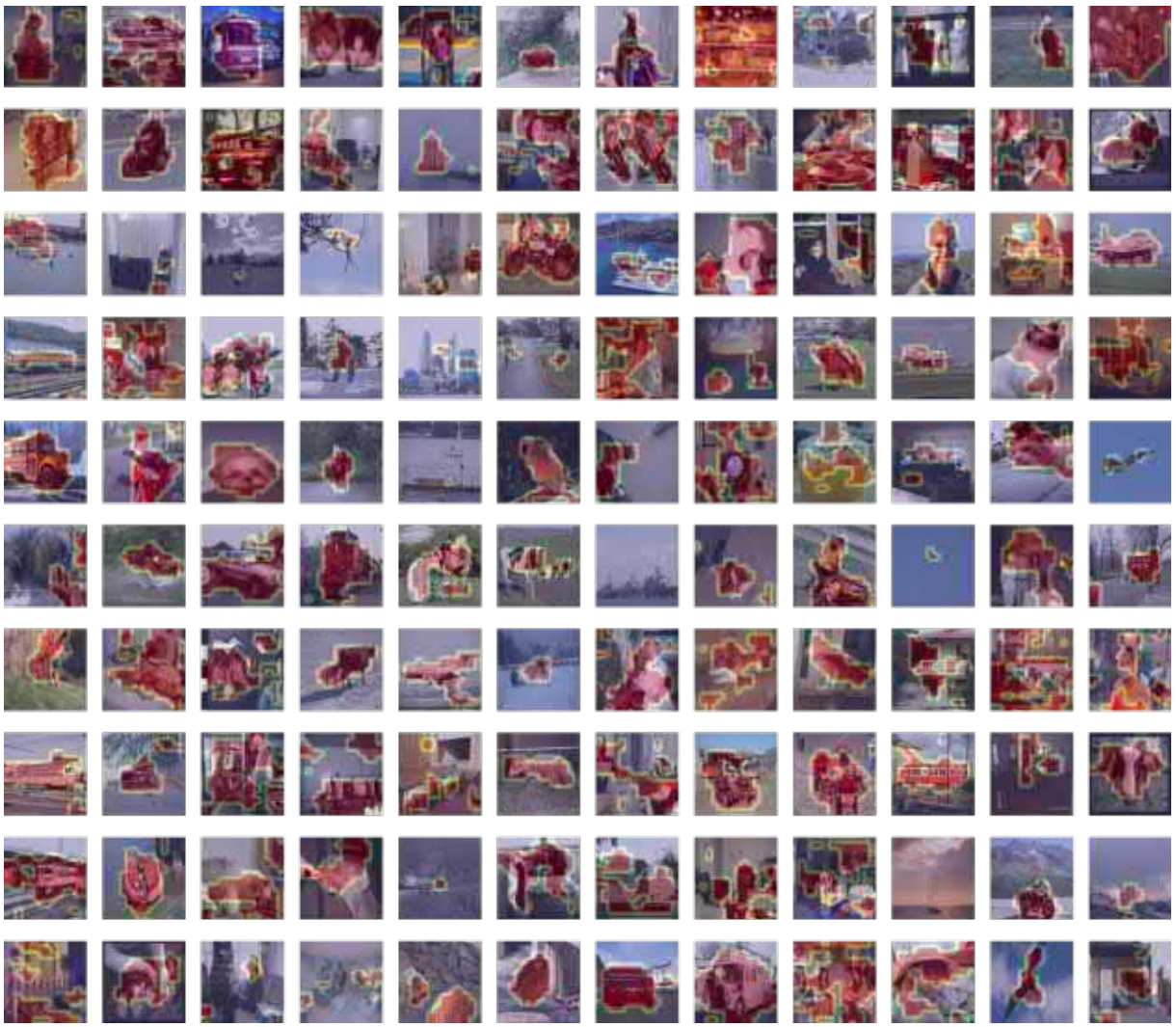

Figure S.35. Localization results of applying *whole* classifier on the sample images from PASCAL VOC. The classifier is trained on layer features.30 of VGG19 with 20 neurons selected by Shapley values.

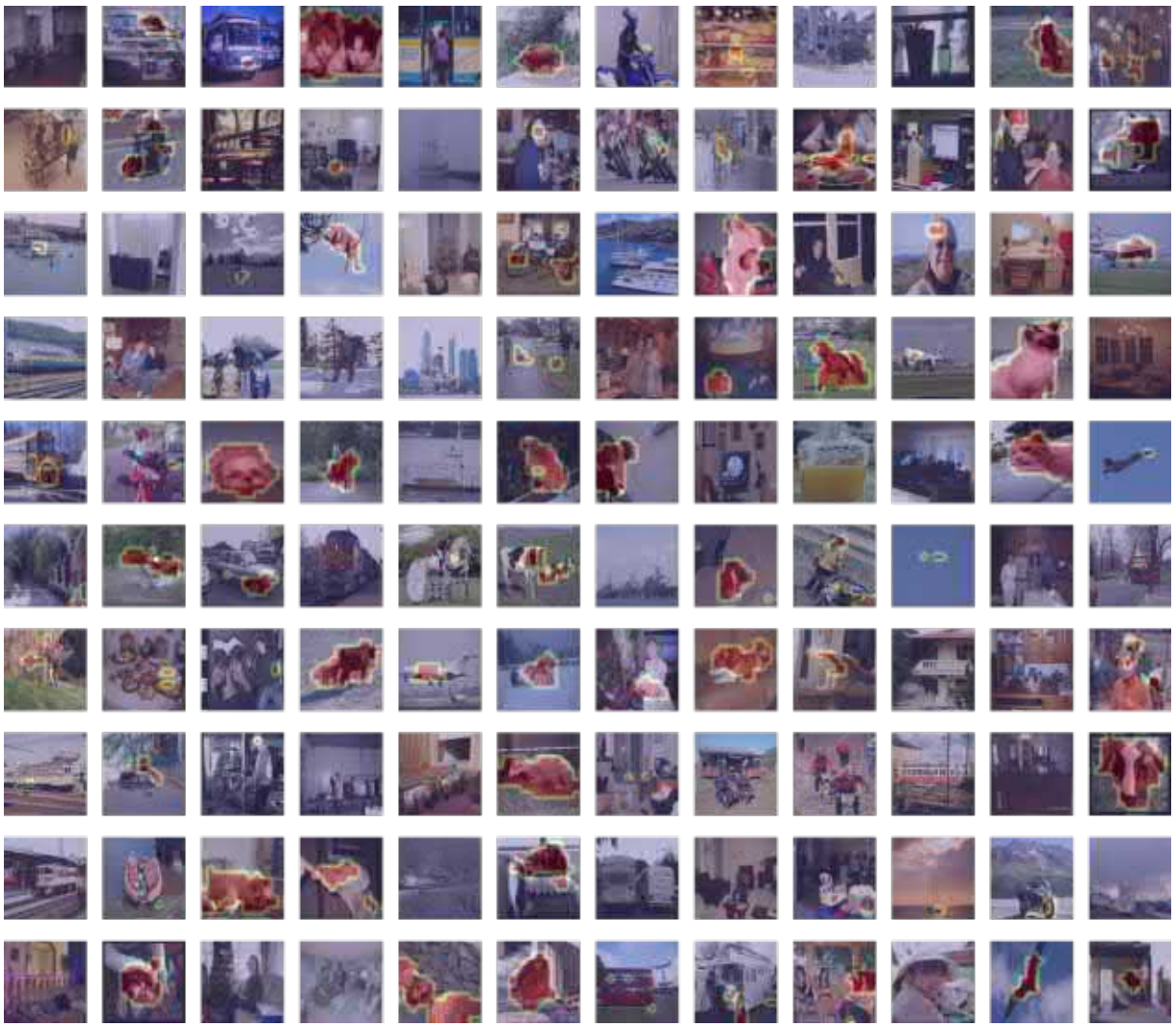

Figure S.36. Localization results of applying *animal* classifier on the sample images from PASCAL VOC. The classifier is trained on layer features.30 of VGG19 with 20 neurons selected by Shapley values.

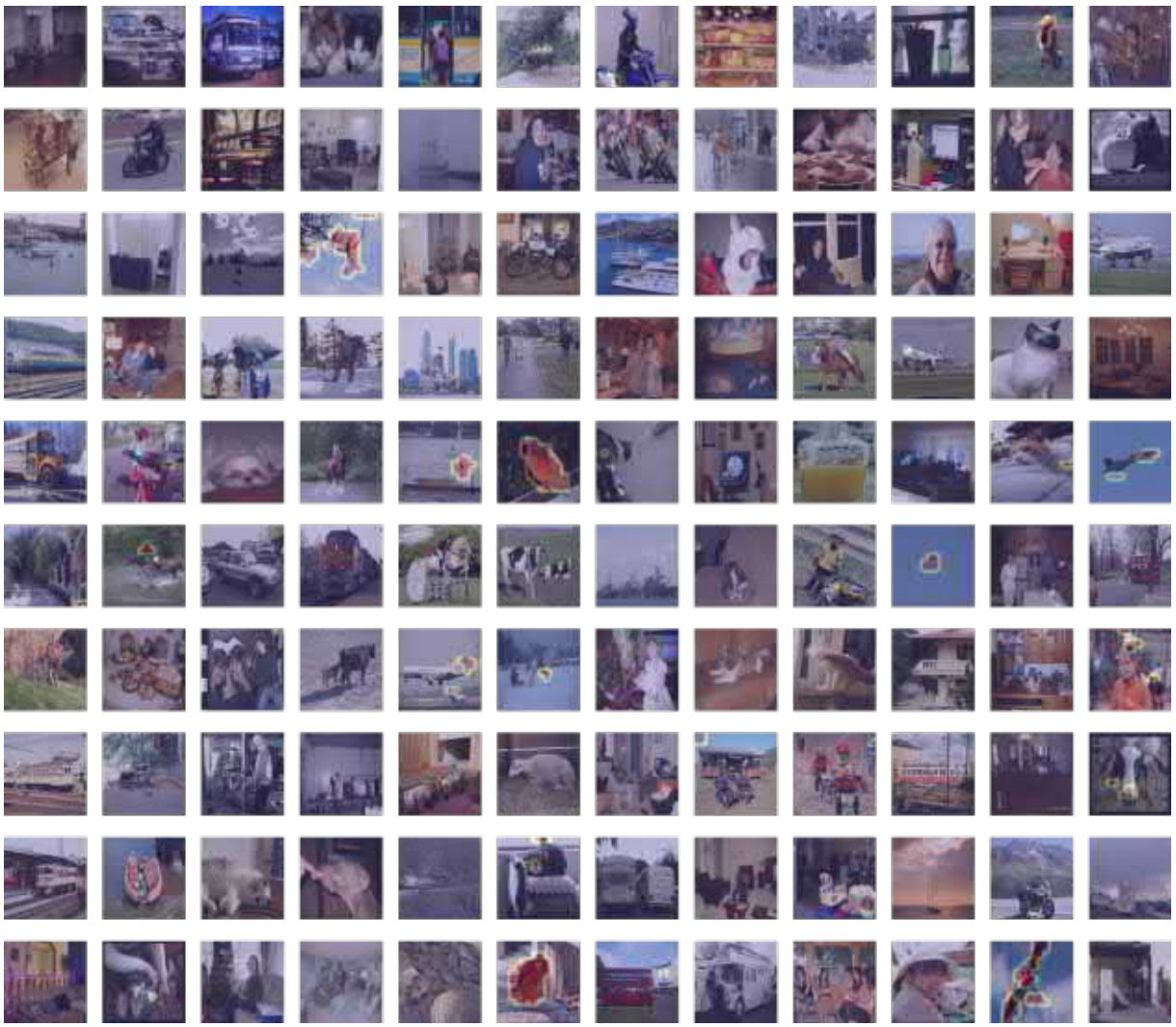

Figure S.37. Localization results of applying *bird* classifier on the sample images from PASCAL VOC. The classifier is trained on layer features.30 of VGG19 with 20 neurons selected by Shapley values.

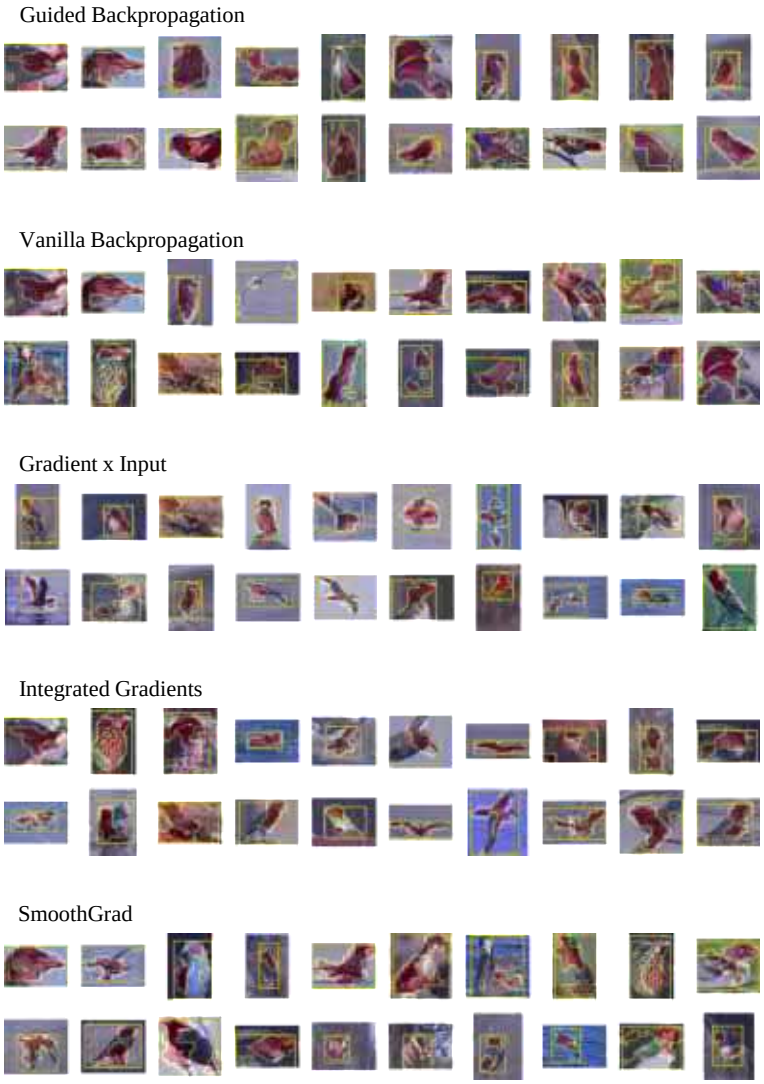

Figure S.38. Localization results of *animal* classifiers implemented with different modified saliency methods.

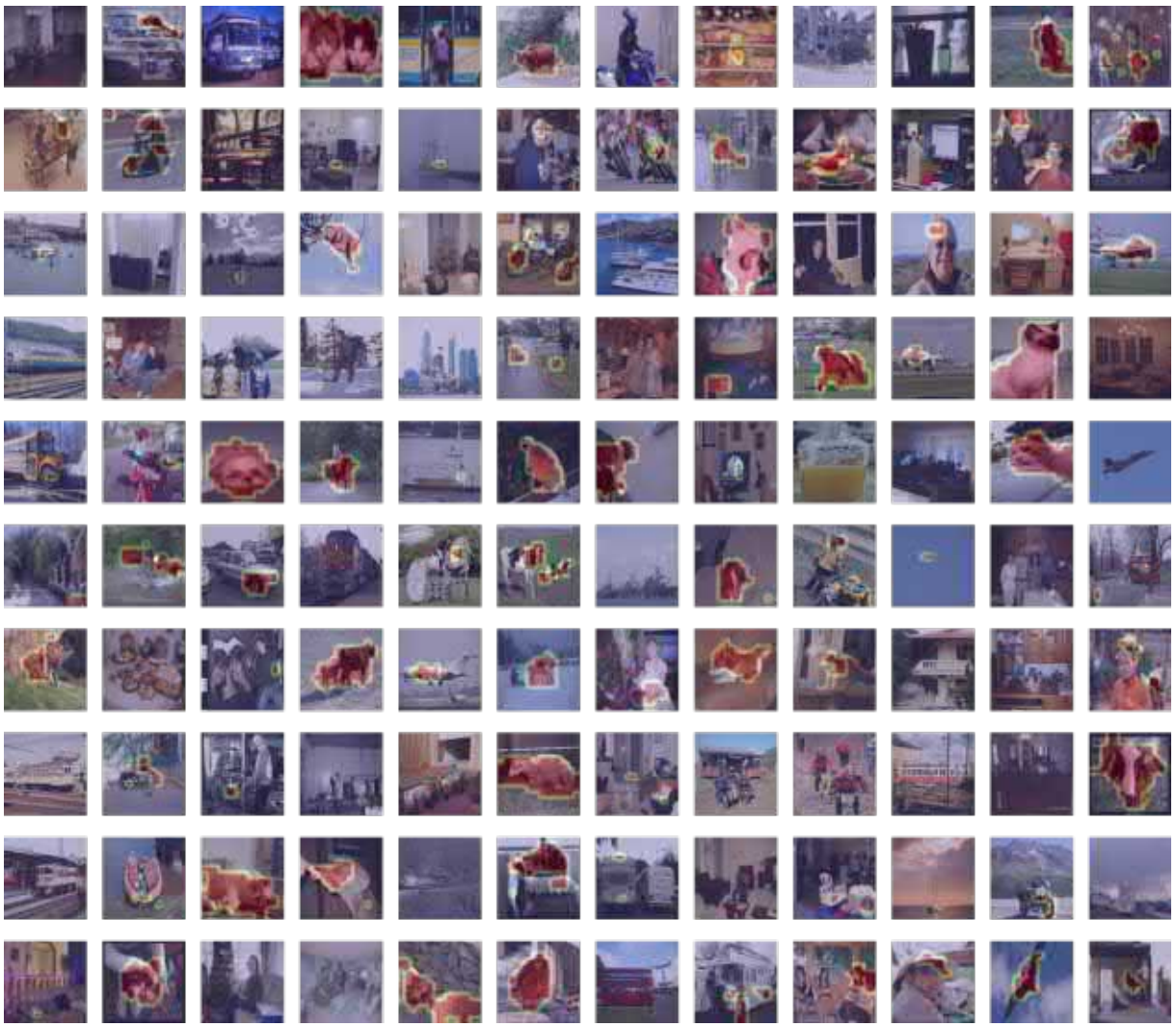

Figure S.39. Localization results of applying *whole* classifier on the sample images from PASCAL VOC. The classifier is trained on layer features.30 of VGG19 with 20 neurons selected by the coefficients of the linear classifier.

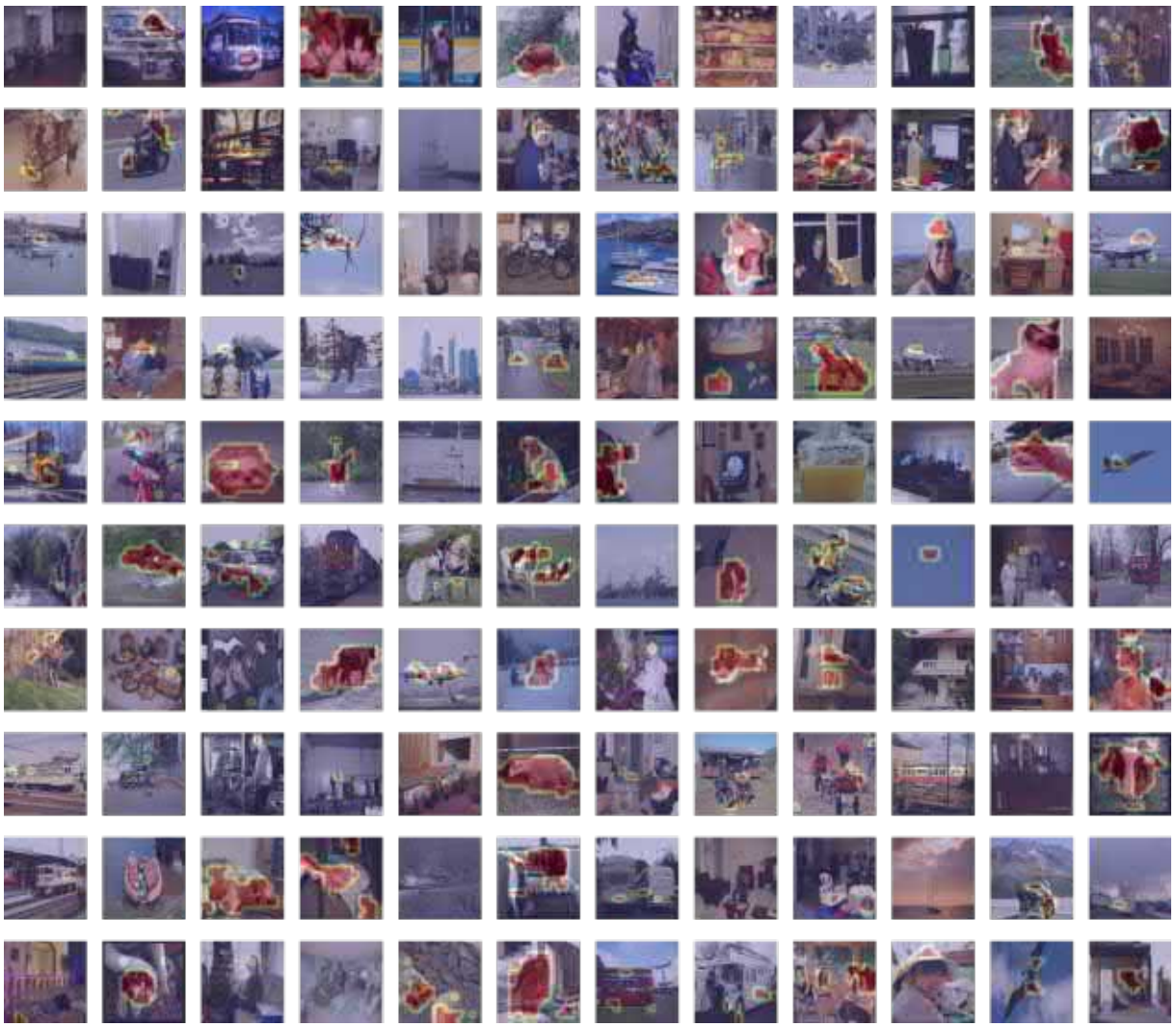

Figure S.40. Localization results of applying *whole* classifier on the sample images from PASCAL VOC. The classifier is trained on layer features.30 of VGG19 with 20 neurons randomly selected.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292*, 2018. 2

[2] Wonho Bae, Junhyug Noh, and Gunhee Kim. Rethinking class activation mapping for weakly supervised object localization. In *European Conference on Computer Vision*. Springer, 2020. 5

[3] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 5

[4] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 2021. 3

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 3

[6] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 3

[7] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016. 3

[8] Ashkan Khakzar, Sabrina Musatian, Jonas Buchberger, Icxel Valeriano Quiroz, Nikolaus Pinger, Soroosh Baselizadeh, Seong Tae Kim, and Nassir Navab. Towards semantic interpretation of thoracic disease and covid-19 diagnosis models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021. 3

[9] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019. 2

[10] Paras Lakhani, John Mongan, Chinmay Singhal, Quan Zhou, Katherine P Andriole, William F Auffermann, Prasanth Prasanna, Theresa Pham, Michael Peterson, Peter J Bergquist, et al. The 2021 siim-fisabio-rsna machine learning covid-19 challenge: Annotation and standard exam classification of covid-19 chest radiographs. *OSF Preprints*, 2021. 3

[11] Weizeng Lu, Xi Jia, Weicheng Xie, Linlin Shen, Yicong Zhou, and Jinming Duan. Geometry constrained weakly supervised object localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*. Springer, 2020. 5

[12] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 3

[13] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2

[14] Jinjie Mai, Meng Yang, and Wenfeng Luo. Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020. 5

[15] Meng Meng, Tianzhu Zhang, Qi Tian, Yongdong Zhang, and Feng Wu. Foreground activation maps for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 5

[16] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020. 1

[17] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016. 2

[18] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 1, 2

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[20] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 2

[21] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 2

[22] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*. PMLR, 2017. 2

[23] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 3

[24] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*. PMLR, 2019. 3, 4

[25] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 5

[26] Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 2009. 5

[27] Haolan Xue, Chang Liu, Fang Wan, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Danet: Divergent activation for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 5

[28] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[29] Chen-Lin Zhang, Yun-Hao Cao, and Jianxin Wu. Rethinking the route towards weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 5

[30] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 2018. 5

[31] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 5

[32] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 5

[33] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 5