Comparing Object Recognition in Humans and Deep Convolutional Neural Networks—An Eye Tracking Study

Leonard Elia van Dyck[1,2]*, Roland Kwitt[3], Sebastian Jochen Denzler[1] and Walter Roland Gruber[1,2]

[1] *Department of Psychology, University of Salzburg, Salzburg, Austria,* [2] *Center for Cognitive Neuroscience, University of Salzburg, Salzburg, Austria,* [3] *Department of Computer Science, University of Salzburg, Salzburg, Austria*

Deep convolutional neural networks (DCNNs) and the ventral visual pathway share vast architectural and functional similarities in visual challenges such as object recognition. Recent insights have demonstrated that both hierarchical cascades can be compared in terms of both exerted behavior and underlying activation. However, these approaches ignore key differences in spatial priorities of information processing. In this proof-of-concept study, we demonstrate a comparison of human observers (N = 45) and three feedforward DCNNs through eye tracking and saliency maps. The results reveal fundamentally different resolutions in both visualization methods that need to be considered for an insightful comparison. Moreover, we provide evidence that a DCNN with biologically plausible receptive field sizes called *vNet* reveals higher agreement with human viewing behavior as contrasted with a standard ResNet architecture. We find that image-specific factors such as category, animacy, arousal, and valence have a direct link to the agreement of spatial object recognition priorities in humans and DCNNs, while other measures such as difficulty and general image properties do not. With this approach, we try to open up new perspectives at the intersection of biological and computer vision research.

INTRODUCTION

In the last few years, advances in deep learning have turned rather simple convolutional neural networks, once developed to simulate the complex nature of biological vision, into sophisticated objects of investigation themselves. Especially the increasing synergy between neural and computer sciences has facilitated this interdisciplinary progress with the aim to enable machines to see and to further the understanding of visual perception in living organisms along the way. Computer vision has given rise to deep convolutional neural networks (DCNNs) that exceed human benchmark performance in key challenges of visual perception (Krizhevsky et al., 2012; He et al., 2015). Among them, the fundamental ability of core object recognition, which allows humans to identify an enormous number of objects despite their substantial variations in appearance

(DiCarlo et al., 2012) and thus to classify visual inputs into meaningful categories based on previously acquired knowledge (Cadieu et al., 2014).
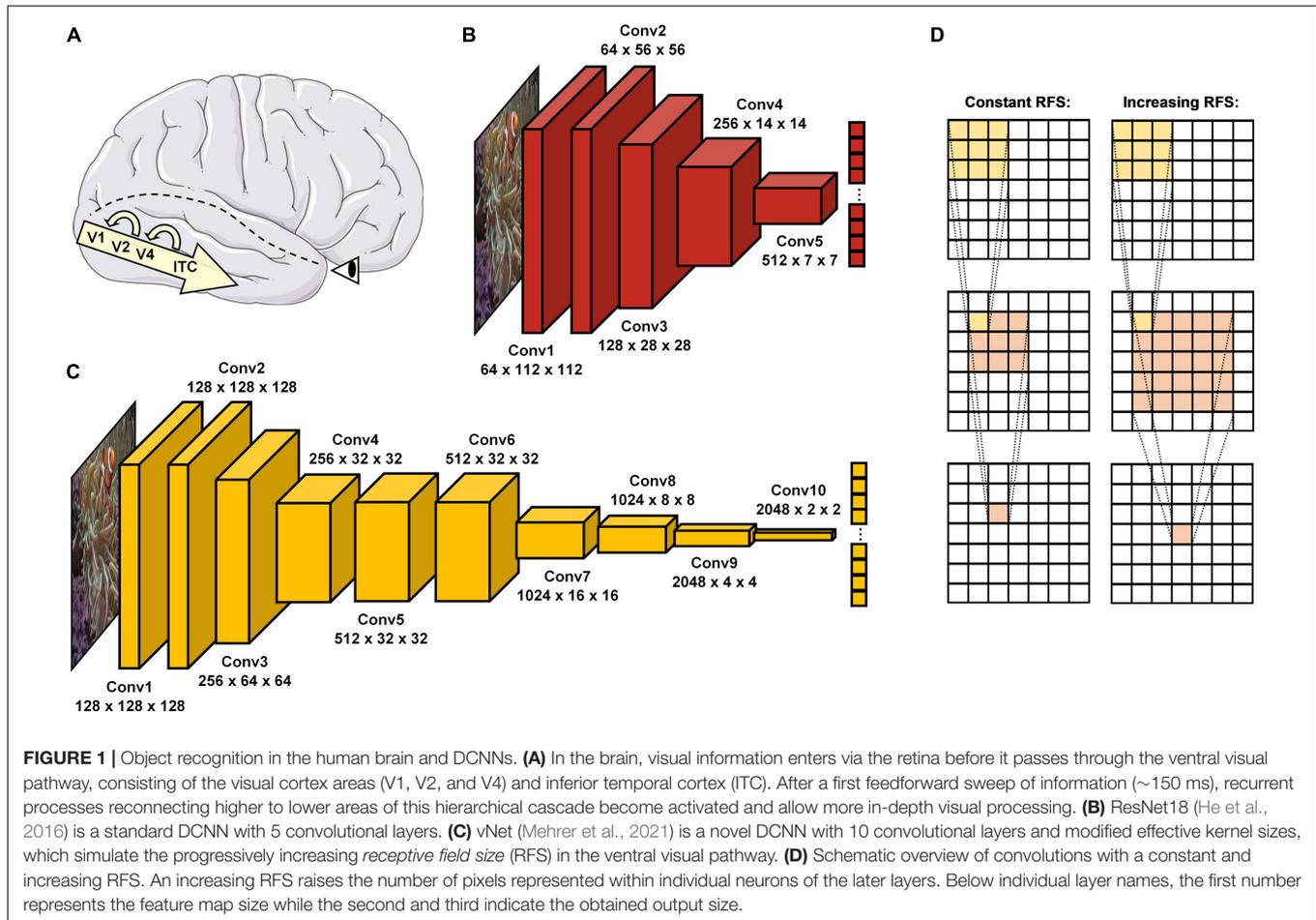
In the brain, this information processing task is solved particularly by the ventral visual pathway (Ishai et al., 1999), which passes information through a hierarchical cascade of retinal ganglion cells (RGC), lateral geniculate nucleus (LGN), visual cortex areas (V1, V2, and V4), and inferior temporal cortex (ITC) (Tanaka, 1996; Riesenhuber and Poggio, 1999; Rolls, 2000; DiCarlo et al., 2012). This organization shares vast similarities with the purely feedforward architectures of DCNNs in a way that visual information can pass through by means of a single end-to-end sweep (see **Figure 1**). While in most cases this processing mechanism seems to suffice for so called *early solved* natural images, a substantial body of literature proposes that especially *late-solved* challenge images benefit from recurrent processing through neural interconnections and loops (Lamme and Roelfsema, 2000; Kar et al., 2019; Kar and DiCarlo, 2020). Moreover, electrophysiological findings therefore suggest that recurrence may set in increasingly after around 150 ms to stimulus onset (DiCarlo and Cox, 2007; Cichy et al., 2014; Contini et al., 2017; Tang et al., 2018; Rajaei et al., 2019; Seijdel et al., 2020). Interestingly, DCNNs seem to face difficulties in recognizing exactly these late-solved (Kar et al., 2019) and manipulated challenge images (Dodge and Karam, 2017; Geirhos et al., 2017, 2018b; van Dyck and Gruber, 2020), which may require additional recurrent processing.

Naturally, images of objects, regardless of whether represented on a biological retina or in a computer matrix, are highly complex. Therefore, in the brain, object recognition is influenced by a number of bottom-up (Rutishauser et al., 2004) and top-down processes (Bar, 2003; Bar et al., 2006). In simplified terms, however, it is thought to solve this challenge by transferring given two-dimensional information into an invariant three-dimensional representation, encoded in an even higher-dimensional neuronal space (DiCarlo et al., 2012). As pointed out by Marr (1982), the two-dimensional image allows first and foremost the extraction of shape features such as edges and regions, the precursors of the so-called *primal sketch* (Marr, 1982, p. 37). Then, when textures and shades start to enrich the outline, a *2.5D sketch* (Marr, 1982, p. 37) emerges. Finally, in combination with previously acquired knowledge, the representation of an invariant *3D model* can be inferred (Marr, 1982, p. 37). These processing steps are reflected not only by the architecture of the ventral visual pathway but can also be found in its artificial replica. While in DCNNs, earlier layers are mostly sensitive to specific configurations of edges, blobs, and colors (e.g., the edge of an orange stripe), the following convolutions start to combine them into texture-like feature groups (e.g., an orange-white striped pattern), until later layers assemble whole object parts (e.g., the fin of an anemone fish) and eventually infer the object class label (e.g., an anemone fish). Interestingly, recent literature suggests that DCNNs trained on the ImageNet dataset (Deng et al., 2009; Russakovsky et al., 2015) are strongly biased toward texture and use it more frequently rather than shape information to classify images (Geirhos et al., 2018a). This preference seems to contradict findings in humans that clearly identify shape as the single most important cue for object recognition (Landau et al., 1988). In addition, several studies have examined the effect of context in human object recognition (Oliva and Torralba, 2007; Greene and Oliva, 2009). Contextual cues are representational regularities such as spatial layout (e.g., houses are usually located on the ground), inter-object dependencies (e.g., houses are usually attached to a street), point of view (e.g., houses are usually looked at from a specific perspective), and other summary statistics (i.e., entropy and power spectral density). Likewise, these same cues are certainly not meaningless to DCNNs. In fact, mainly work around removing or manipulating one of these hints such as the object's regular pose (Alcorn et al., 2019) or background (Beery et al., 2018) have demonstrated that indeed contextual learning, which is common practice in learning systems, can be found here as well (Geirhos et al., 2020).

Nevertheless, a key difference between humans and DCNNs lies within the modulating effects of appraisal. While millions of years of evolution have tuned *in vivo* object recognizers such as humans to seek and avoid different kinds of stimuli (e.g., find nourishment and avoid dangers quickly), this subjective experience of for example arousal and valence is lacking completely in *in silico* models. Furthermore, the neuroscientific literature contains many examples suggesting that particularly the automatic detection of fear-relevant and threatful objects is solved by an even faster subcortical route, which skips parts of the ventral visual pathway through shortcuts to the amygdala (Öhman, 2005; Pessoa and Adolphs, 2010). This for example might enable threat-superiority effects in terms of reaction times and reduced position effects in a visual search paradigm (Blanchette, 2006). While the plausibility of this so-called *low road* is highly discussed (Cauchoix and Crouzet, 2013), the differences in performance are well-documented.

As an interim summary, it can be noted that the ventral visual pathway and DCNNs suggest conceptual overlaps but also substantial differences in many regards. Hence, more recent evidence from Mehrer et al. (2021) further highlights the importance of biological plausibility for the fit between brain and DCNN activity, as their novel architecture called *vNet* simulates the progressively increasing foveal *receptive field size* (hereafter abbreviated as RFS) along the ventral visual pathway (Wandell and Winawer, 2015; Grill-Spector et al., 2017). However, as their analyses point out, this modification does not lead to higher congruence with for example fMRI activity of human observers when compared to a standard architecture such as AlexNet (Krizhevsky et al., 2012). This raises many questions about the definite impact of this RFS modification. Therefore, as vNet's hierarchical organization is designed to resemble that of the ventral visual pathway more accurately as compared to a standard DCNN, here we hypothesize that the major advantage of vNet may not be visible within more brain-like activations, as also not found by Mehrer et al. (2021), but rather more similar spatial priorities of information processing compared through eye tracking and GradCAM. Consequently, following an important distinction in human-machine comparisons by Firestone (2020), we believe that this resemblance in underlying *competence* should lead to higher similarity in object recognition behavior and further observable *performance*. As in this specific

FIGURE 1 | Object recognition in the human brain and DCNNs. **(A)** In the brain, visual information enters via the retina before it passes through the ventral visual pathway, consisting of the visual cortex areas (V1, V2, and V4) and inferior temporal cortex (ITC). After a first feedforward sweep of information (∼150 ms), recurrent processes reconnecting higher to lower areas of this hierarchical cascade become activated and allow more in-depth visual processing. **(B)** ResNet18 (He et al., 2016) is a standard DCNN with 5 convolutional layers. **(C)** vNet (Mehrer et al., 2021) is a novel DCNN with 10 convolutional layers and modified effective kernel sizes, which simulate the progressively increasing *receptive field size* (RFS) in the ventral visual pathway. **(D)** Schematic overview of convolutions with a constant and increasing RFS. An increasing RFS raises the number of pixels represented within individual neurons of the later layers. Below individual layer names, the first number represents the feature map size while the second and third indicate the obtained output size.

human-machine comparison rather divergent architectures are compared, we believe that a RFS modification can result in more similar spatial priorities of information processing and likewise object recognition behavior, without immediately suggesting a higher match between activity patterns of neural components and individual DCNN layers.
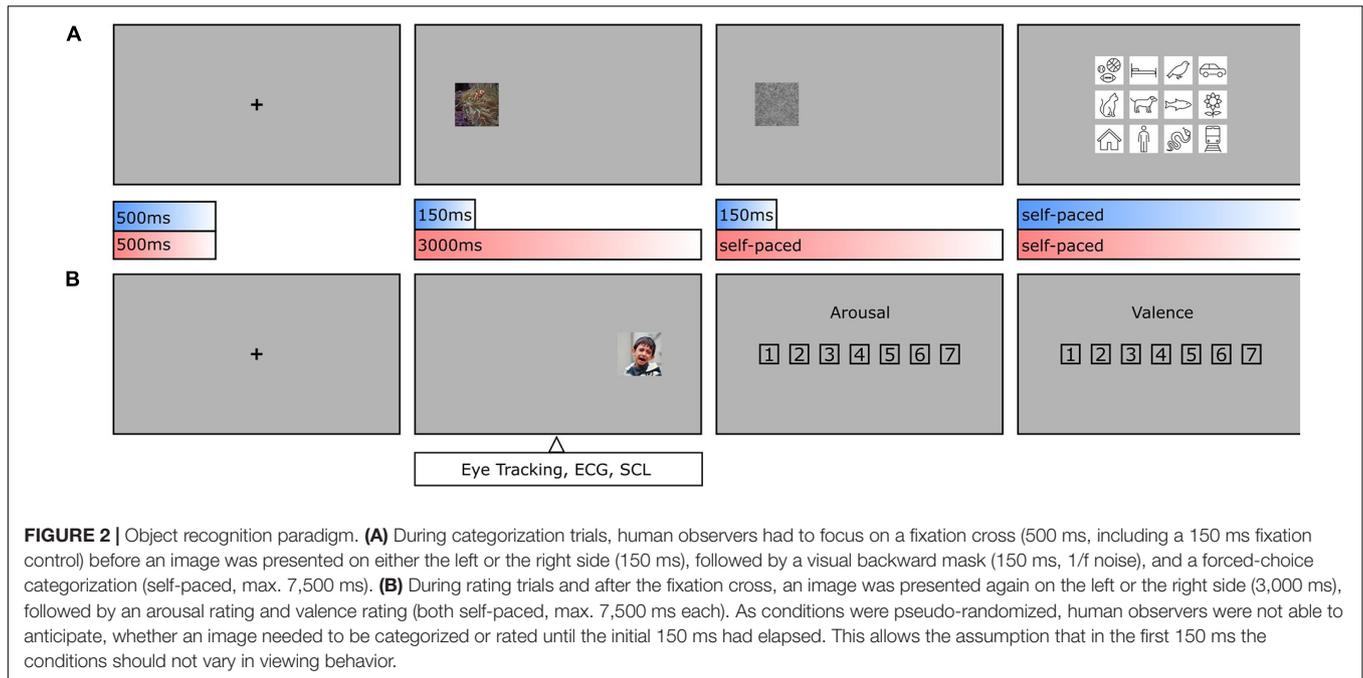
As DCNNs are getting more and more complex, several *attribution*-tools have been developed to understand (to some extent) their classifications. At first glance, these new visualization algorithms, also called *saliency maps*, resemble well-established methods in cognitive neuroscience. In eye tracking, the execution of a visual task is analyzed by mapping gaze behavior onto specific regions of interest, which receive special cognitive or computational priorities during information processing. While eye tracking and saliency maps share this concept, the methodological way this is achieved seems fundamentally different. In eye tracking measurements, a human observer is presented with a stimulus, which is only presented for a limited time, while viewing behavior is being recorded. Saliency maps such as *Gradient-weighted Class Activation Mapping*, also known as *GradCAM* (Selvaraju et al., 2017), extract class activations within a specific layer of the DCNN to explain obtained predictions. Therefore, the algorithm uses the gradient of the loss function to compute a weight for every feature map.

The weighted sum of these activations is class-discriminative and hence allows the localization and visualization of all relevant regions that contributed to the probability of a given class. However, DCNNs do not operate on a meaningful time scale (other than related to computational power) when trying to classify images. Despite the conceptual similarity between eye tracking and saliency maps, only a few attempts have been made to draw this comparison of black boxes (Ebrahimpour et al., 2019). Importantly, a major challenge in this field is to encourage and conduct fair human-machine comparisons, as it is only possible to infer similarities and differences if there are no fundamental constraints within the comparison itself (Firestone, 2020; Funke et al., 2020). In this study, we compare human eye tracking to DCNN saliency maps in an approximately species-fair object recognition task and examine a wide range of possible factors influencing similarity measures.

MATERIALS AND METHODS

General Procedure

In order to test our hypotheses, we designed a fair human-machine comparison that allowed us to investigate behavioral, eye tracking, and physiological data from human observers

FIGURE 2 | Object recognition paradigm. **(A)** During categorization trials, human observers had to focus on a fixation cross (500 ms, including a 150 ms fixation control) before an image was presented on either the left or the right side (150 ms), followed by a visual backward mask (150 ms, 1/f noise), and a forced-choice categorization (self-paced, max. 7,500 ms). **(B)** During rating trials and after the fixation cross, an image was presented again on the left or the right side (3,000 ms), followed by an arousal rating and valence rating (both self-paced, max. 7,500 ms each). As conditions were pseudo-randomized, human observers were not able to anticipate, whether an image needed to be categorized or rated until the initial 150 ms had elapsed. This allows the assumption that in the first 150 ms the conditions should not vary in viewing behavior.

performing a laboratory experiment, as well as predictions and activations of three DCNN architectures on identical visual stimuli and under roughly similar conditions. The main task in this experiment was to categorize briefly shown images based on a forced-choice format of 12 basic-level categories, namely *human*, *dog*, *cat*, *bird*, *fish*, *snake*, *car*, *train*, *house*, *bed*, *flower*, *ball* (see section Methods). Basic-level categories (e.g., *dog*) were chosen over detailed category concepts which are more often used in the field of computer vision (e.g., *border collie*), as they are more naturally utilized in human object classification (Rosch, 1999).

Human Observers—Eye Tracking Experiment

A total of 45 valid participants (28 female, 17 male) with an age between 18 and 31 years (M = 22.64, SD = 2.57) were tested in the eye tracking experiment. Participants were required to have normal or corrected-to-normal vision without problems of color perception and other eye diseases. Two participants with contact lenses had to be excluded from further analyses due to insufficient eye tracking precision. The experimental procedure was admitted by the University of Salzburg ethics committee, in line with the declaration of Helsinki and agreed to by participants via written consent before the experiment. Psychology students received accredited participation hours for taking part in the experiment.

The experiment consisted of two main tasks (see **Figure 2**). Participants had to concentrate on a fixation cross for 500 ms until an image appeared at center 12.93 degrees of visual angle away on the left or right side. If, as in one condition, the image was presented for a short duration of 150 ms, a visual backward mask (1/f noise) followed for the same duration, and the participant had to classify the presented object based on a forced-choice format of 12 basic-level categories by clicking on

the respective class symbol. If, as in the other condition, the image was presented for a long duration of 3000 ms, participants had to rate it afterward in its arousal (1 = *Very low*, 4 = *Neither low nor high*, 7 = *Very high*) and valence (1 = *Very negative*, 4 = *Neutral*, 7 = *Very positive*) on a scale from 1 to 7. As both classification and rating tasks were balanced out in occurrence and previously pseudo-randomized, it was impossible for the observers to differentiate between the two conditions before the short presentation time was exceeded. Based on this central assumption, both conditions should not vary in viewing behavior. Participants were familiarized with the experimental procedure during training trials which were excluded from further analyses. The whole experiment consisted of 420 test trials (210 per condition), took about 1 h to complete, and was divided into three blocks with resting breaks in between. In this way, a single participant classified one half of the entire dataset and rated the other half. To obtain categorization and rating results for all images, there were two versions of the experiment with interchanged conditions for both halves.

The participants performed the experiment in a laboratory room, where they were seated in front of a screen (1,920 × 1,080 pixels, 50 Hz) and had to place their head into a chin rest located at a distance of 60 cm. The right eye was tracked and recorded with an EyeLink 1000 (SR Research Ltd., Mississauga, ON, Canada) desktop mount, at a sampling rate of 1,000 Hz. For presentational purposes, original images were scaled by factor two onscreen (448 × 448) but stayed unchanged in image resolution (224 × 224). This way, the presented images had 11.52 degrees of visual angle in size. Recorded eye tracking data were preprocessed using DataViewer (Version 4.2, SR Research Ltd., Mississauga, ON, Canada) and analyzed after the participants gaze crossed an invisible boundary framing the entire image. In this way, participants were able to process

appearing images already peripherally for the first couple of milliseconds to allow a meaningfully programmed first fixation (see **Figure 3**). Fixations were compiled for the first 150 ms within the image. Here, x- and y-coordinates were downscaled again from expanded presentation size (448×448) to original image size (224×224). Average heatmaps were computed from individual sampling points of either the first 150 ms (= feedforward) or the entire presentation time after 150 ms (= recurrent) within the image using in-house built MATLAB scripts. The obtained heatmaps for all participants were averaged per image, Gaussian filtered with a standard deviation of 15 pixels, and normalized.

Additionally, before the eye tracking experiment started, participants were wired with ECG and SCL electrodes. Physiological measurements were collected using a Varioport biosignal recorder (Becker Meditec, Karlsruhe, Germany) and analyzed with ANSLAB (Blechert et al., 2016) (Version 2.51, Salzburg, Austria). Here, the relative change in mean activity from a baseline of 1000 ms before the image presentation to the time window of 3000ms during the image presentation was used.

Deep Convolutional Neural Network Training and GradCAM Saliency Maps

Our implementation (in PyTorch) of vNet follows the original GroupNorm variant of vNet from Mehrer et al. (2021), the ResNet18 architecture follows the original proposal from He et al. (2016). For training both network types, we use the (publicly available) Ecoset *training* set, constrained to the 12 categories described in section Methods below. As the number of images differs significantly across categories, we artificially balance the corpus by drawing (uniformly at random) N images per category, where N corresponds to the number of images in the smallest category (i.e., *fish*). During training, all ($\sim$15 k) images are first resized to a spatial resolution of 256×256, then cropped to the center square of size 224×224 and eventually normalized (by subtracting the channel-wise mean and dividing by the channel-wise standard deviation, computed from the training corpus). Note that no data augmentation is applied across all experiments. We minimize the cross-entropy loss using stochastic gradient descent (SGD) with momentum (0.9) and weight decay (1e–4) under a cosine learning rate schedule, starting at an initial learning rate of 0.01. We train for 80 epochs using a batch size of 128. Results for the fine-tuned ResNet18 are obtained by replacing the final linear classifier of an ImageNet-trained ResNet18, freezing all earlier layers, and fine-tuning for 50 epochs (in the same setup as described before). When referring to *early*, *middle*, and *late* layers in the manuscript, we refer to GradCAM outputs generated from activations after the 2nd, 6th, and 9th layer for vNet, and activations after the 1st, 3rd, and 4th ResNet18 block (as all ResNet architectures total four blocks). Eventually, if the aim is to investigate the similarity in spatial priorities of information processing that underlie object recognition behavior with the demonstrated approach, the output layer of a DCNN would suffice for the comparison. In this proof-of-concept study, however, we also include earlier layers for sanity checks and further model comparisons.
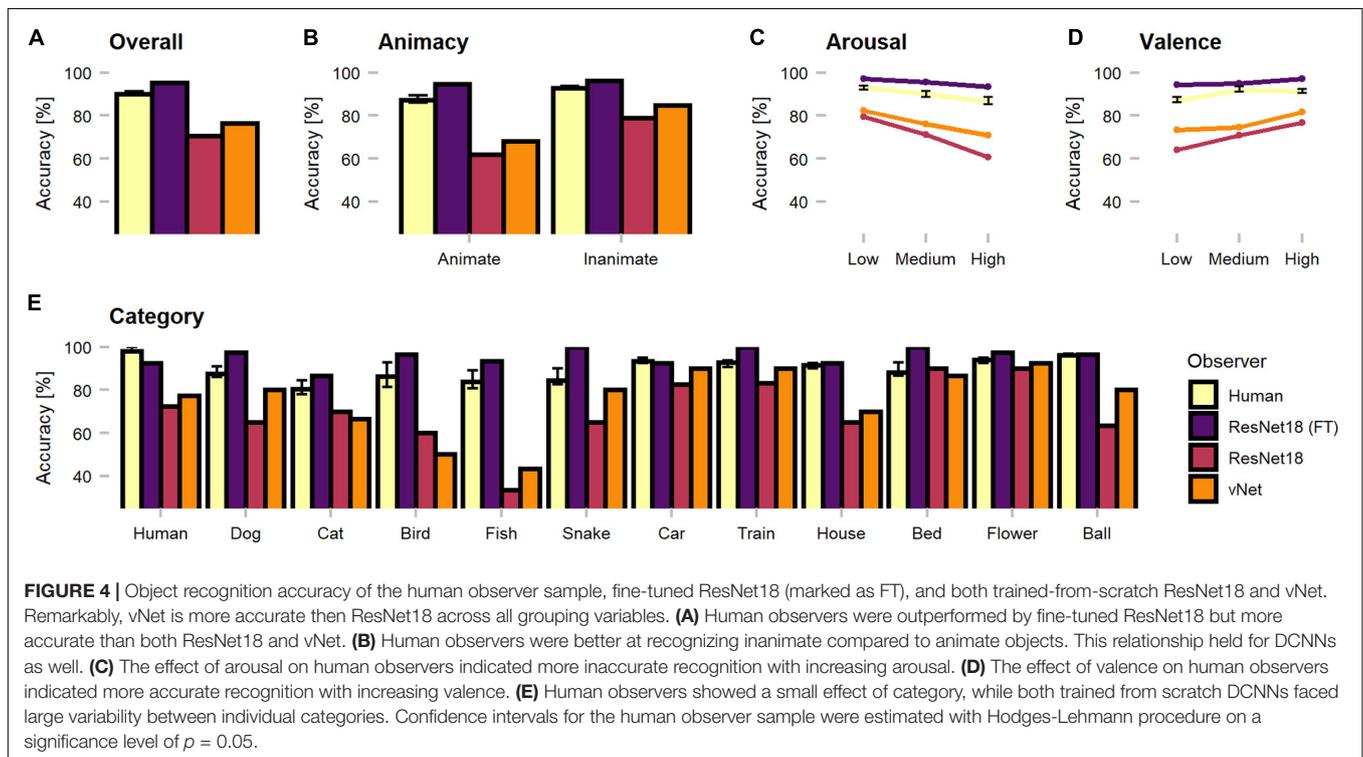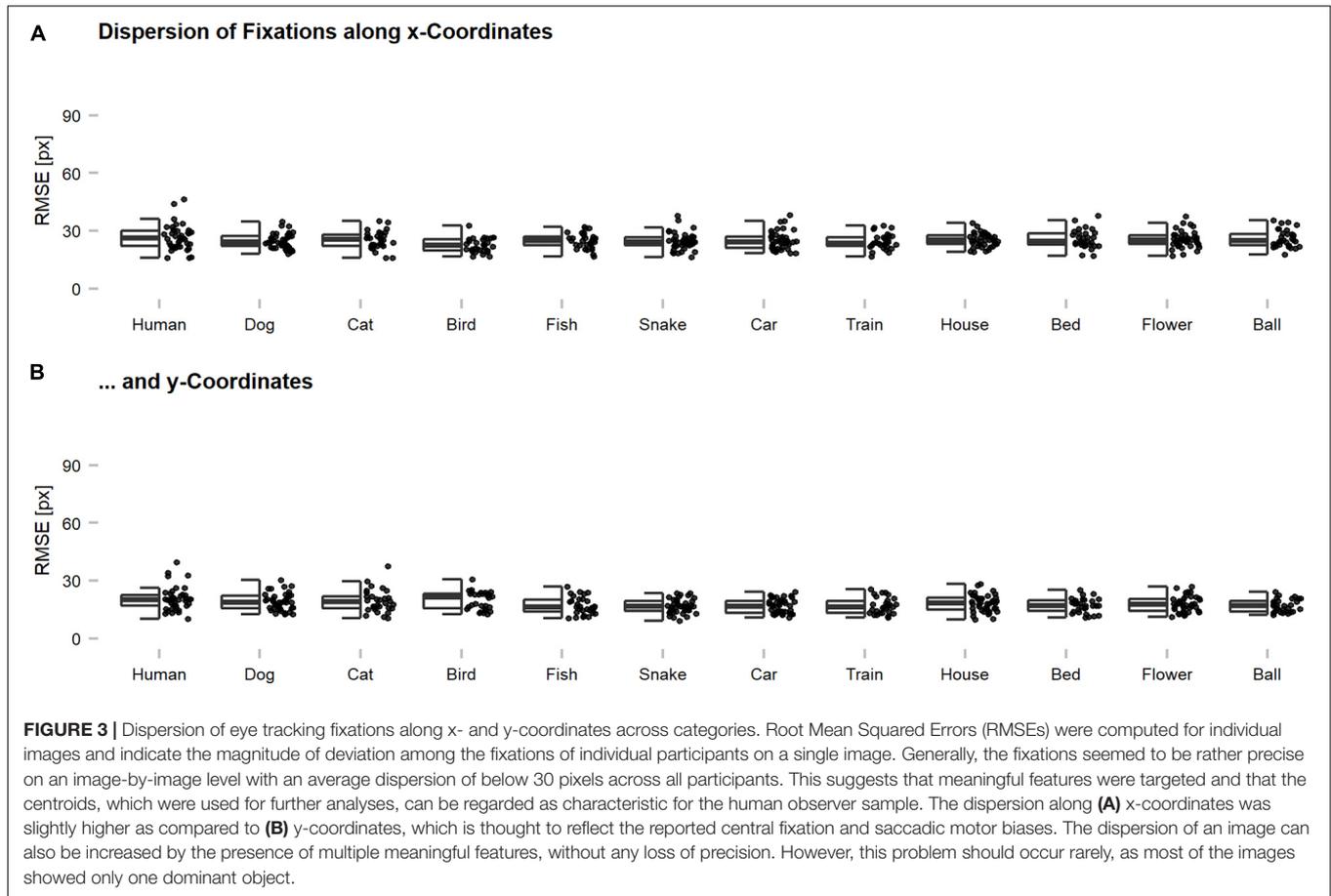
Dataset

Images were part of 12 basic-level categories from the ecologically motivated Ecoset dataset, which was created by Mehrer et al. (2021) in order to better capture the organization of human-relevant categories. The dataset consisted of 6 animate (namely *human*, *dog*, *cat*, *bird*, *fish*, and *snake*) and 6 inanimate categories (namely *car*, *train*, *house*, *bed*, *flower*, and *ball*) with 30 images per category. During preprocessing, images were randomly drawn from the test set, cropped toward the biggest possible central square, and resized to 224×224 pixels. All images were visually checked and excluded if multiple categories (e.g., *human* and *dog*), overlayed text, or image effects (e.g., grayscale images) were visible or the object was fully removed during preprocessing steps. In categories where less than 30 images from the test set remained, Ecoset images were complemented with ImageNet examples ($n = 28$, across 4 categories). Additionally, in 3 animate (namely *human*, *dog*, and *snake*) and 3 inanimate categories (namely *car*, *house*, *flower*), 10 images per category of the respective objects were added from the Open Affective Standardized Image Set (OASIS) (Kurdi et al., 2017). Here, arousal and valence ratings from a large number of participants ($N = 822$) were already available and increased the variability while serving as a sanity check for own ratings. In total, the dataset consisted of 420 test images.

RESULTS

Performance

The first set of analyses investigated object recognition performance in human observers and DCNNs. Therefore, human predictions obtained during categorization trials (see section Methods) were compared against model predictions. Generally, as the categorization data were not normally distributed, non-parametric tests were applied to compare the human observer sample against fixed-accuracy values of individual DCNNs. On average, human observers reached a recognition accuracy of 89.96%. One-sample Wilcoxon tests indicated that human observers were significantly outperformed by fine-tuned ResNet18 with 95.48% [$V = 0$, CI = (89.52, 91.43), $p < 0.001$, $r = 0.85$] but significantly more correct than both *trained-from-scratch* ResNet18 [$V = 1,035$, CI = (89.52, 91.43), $p < 0.001$, $r = 0.85$] and vNet [$V = 1,034$, CI = (89.52, 91.43), $p < 0.001$, $r = 0.85$] with accuracies of 70.48 and 76.43%, respectively. The results endorse both sides of the literature by demonstrating that especially DCNNs trained on large datasets can exceed human benchmark performance (Krizhevsky et al., 2012; Szegedy et al., 2015; He et al., 2016; Huang et al., 2017), but also simultaneously reminds of possible limits due to the amount of provided training data. Nevertheless, following analyses focus predominantly on the two equally trained DCNNs, as they are more suitable for a fair comparison of architectures.

Remarkably, vNet outperformed ResNet18 throughout all grouping variables (animacy, arousal, valence, and category) and generally seemed to be closer to the human benchmark level of accuracy (see **Figure 4**). However, across categories, applied Kruskal Wallis tests revealed that both ResNet18 [$X^2(29) = 60.18$,

FIGURE 3 | Dispersion of eye tracking fixations along x- and y-coordinates across categories. Root Mean Squared Errors (RMSEs) were computed for individual images and indicate the magnitude of deviation among the fixations of individual participants on a single image. Generally, the fixations seemed to be rather precise on an image-by-image level with an average dispersion of below 30 pixels across all participants. This suggests that meaningful features were targeted and that the centroids, which were used for further analyses, can be regarded as characteristic for the human observer sample. The dispersion along **(A)** x-coordinates was slightly higher as compared to **(B)** y-coordinates, which is thought to reflect the reported central fixation and saccadic motor biases. The dispersion of an image can also be increased by the presence of multiple meaningful features, without any loss of precision. However, this problem should occur rarely, as most of the images showed only one dominant object.

FIGURE 4 | Object recognition accuracy of the human observer sample, fine-tuned ResNet18 (marked as FT), and both trained-from-scratch ResNet18 and vNet. Remarkably, vNet is more accurate then ResNet18 across all grouping variables. **(A)** Human observers were outperformed by fine-tuned ResNet18 but more accurate than both ResNet18 and vNet. **(B)** Human observers were better at recognizing inanimate compared to animate objects. This relationship held for DCNNs as well. **(C)** The effect of arousal on human observers indicated more inaccurate recognition with increasing arousal. **(D)** The effect of valence on human observers indicated more accurate recognition with increasing valence. **(E)** Human observers showed a small effect of category, while both trained from scratch DCNNs faced large variability between individual categories. Confidence intervals for the human observer sample were estimated with Hodges-Lehmann procedure on a significance level of $p = 0.05$.

$p < 0.001$, $r = 0.14$] and vNet [$X^2(33) = 58.65$, $p < 0.001$, $r = 0.11$] performed significantly dissimilar to human observers and each other [$X^2(33) = 62.76$, $p = 0.001$, $r = 0.61$]. Moreover, based on previous findings, we hypothesized that human observers should be significantly better at recognizing animate compared to inanimate objects (New et al., 2007). However, Wilcoxon rank sum tests indicated the existence of this effect but in the opposite direction, as human observers were significantly more accurate in recognizing inanimate (Median = 93.27) compared to animate objects (Median = 88.68; W = 345.5, $p < 0.001$, $r = 0.55$). Additionally, ResNet18 and vNet mirrored this behavior with a substantial increase in accuracy from animate (ResNet18: 61.90%/vNet: 68.10%) to inanimate objects (ResNet18: 79.05%/vNet: 84.76%).

Furthermore, low, medium, and high arousal and valence groups, obtained by tertile splits of human observer ratings (33rd and 66th percentile, see section Methods), revealed a significant, negative Spearman rank correlation between human accuracy and arousal (rho = −0.22, $p < 0.001$). Interestingly, this relationship did not disappear when partial correlations controlling for animacy were computed (rho = −0.20, $p = 0.001$). Subsequent Kruskal Wallis tests hinted at an effect of arousal [$X^2(2) = 38.51$, $p < 0.001$, $r = 0.50$] and Bonferroni-corrected, pairwise Wilcoxon tests revealed a significant decrease of accuracy between all three increasing levels of arousal. A significant positive correlation (rho = 0.34, $p < 0.001$), partial correlation controlling for animacy (rho = 0.37, $p < 0.001$), and an effect [$X^2(2) = 36.90$, $p < 0.001$, $r = 0.48$] with a significant increase of accuracy between low and medium levels in Bonferroni-corrected *post-hoc* tests were found for valence. Similarly, the accuracies of ResNet18 and vNet seemed to follow both effects. Images that were assessed as more calm and positive by human observers during rating trials lead to better recognition performance in human observers and DCNNs.

It is fundamental to note that only the performance of vNet exhibited a significant, positive Spearman rank correlation with human observer performance (see **Table 1**; rho = 0.14, $p = 0.009$), which may indicate a better fit to human categorization behavior. Contrary to our expectations, image properties (namely entropy, shape, texture, and power spectral peak-to-mean ratio) seemed to be rather unrelated to performance. Yet, as expected, the parameters were highly correlated among each other and demonstrated the statistical regularities of complex natural images.

In order to shine more light on the classification errors made by human observers and DCNNs, which underlie the reported performances, categorization patterns were investigated (see **Figure 5**). Interestingly, as human observers seemed to have difficulties with relatively common classes (such as dog and cat or fish and snake), their classification behavior suggests that these conceptually similar classes could have lead to confusions. It should also be taken into account that other top-down and bottom-up influencing factors such as contextual cues may be especially similar between these classes. Moreover, both trained-from-scratch ResNet18 and vNet were found to misclassify images in similar ways.

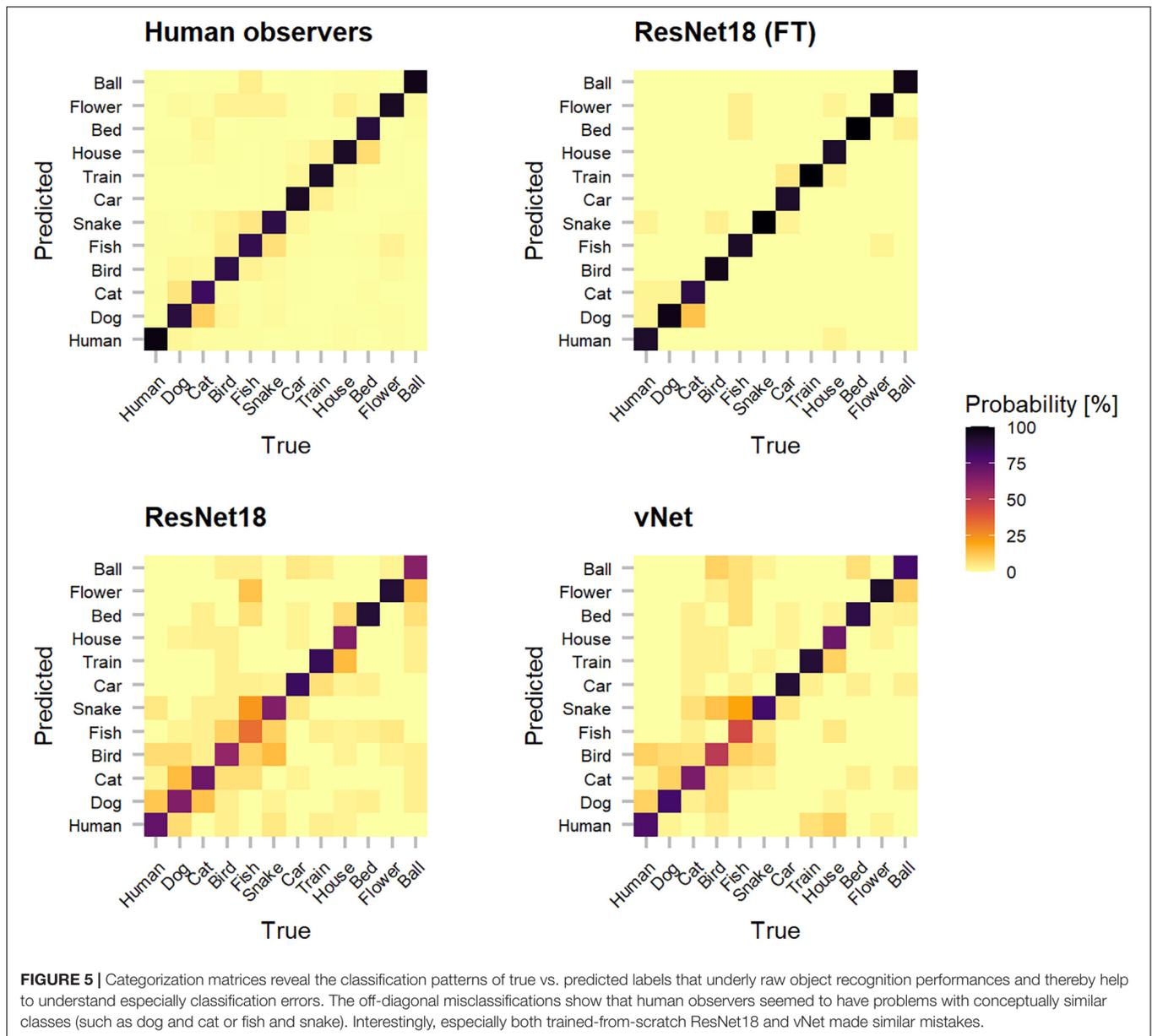## Fixations—Feedforward vs. Feedforward Processing

In an attempt to identify priorities during feedforward information processing in human observers and DCNNs, we inspected human fixations and global maxima of saliency maps. Fixations were compiled for the first 150 ms, the theoretical time of a feedforward pass, and assigned to 1 out of 16 equally sized target blocks. Similarly, for GradCAM, the centroid of the single highest scoring patch was defined as the global maximum and used as an equivalent with the respective target block. As displayed in **Figure 6**, we found that human fixations were subject to a *central fixation bias* (Tatler, 2007; Rothkegel et al., 2017), as most individual and almost all average fixations were located within the center blocks. Furthermore, a *saccadic motor bias* was visible, as average fixations of images presented on the left side were predominantly located on the right side of the image and vice versa. This pattern is thought to reflect the preference of the saccadic system for smaller amplitude eye movements over larger ones (Tatler et al., 2006). In most cases, however, meaningful fixations on object features could be clearly identified on the individual level.

Although we hypothesized GradCAM maxima to differ substantially from human fixations, the results of ResNet18 and vNet across early, middle, and late layers proposed fundamentally diverging distributions deeper down the architectures. While human fixations were found to be normally distributed on a

TABLE 1 | Spearman rank correlation between human observer accuracy, DCNN accuracy, and image properties.

	Human Acc.	ResNet18 (FT) Acc.	ResNet18 Acc.	vNet Acc.	Entropy	Shape	Texture
Human Acc.							
ResNet18 (FT) Acc.	0.071						
ResNet18 Acc.	0.030	0.236***					
vNet Acc.	0.127**	0.203***	0.551***				
Entropy	0.041	−0.071	−0.053	0.027			
Shape	0.016	−0.096*	−0.025	−0.019	0.299***		
Texture	0.036	−0.056	−0.044	0.000	0.432***	0.763***	
Peak-to-Mean	0.015	0.058	0.072	0.048	−0.271***	−0.591***	−0.874***

Acc. stands for Accuracy. Peak-to-Mean stands for Power Spectral Peak-to-Mean Ratio. Computed correlation used spearman-method with listwise-deletion. *p < 0.05, **p < 0.01, ***p < 0.001.
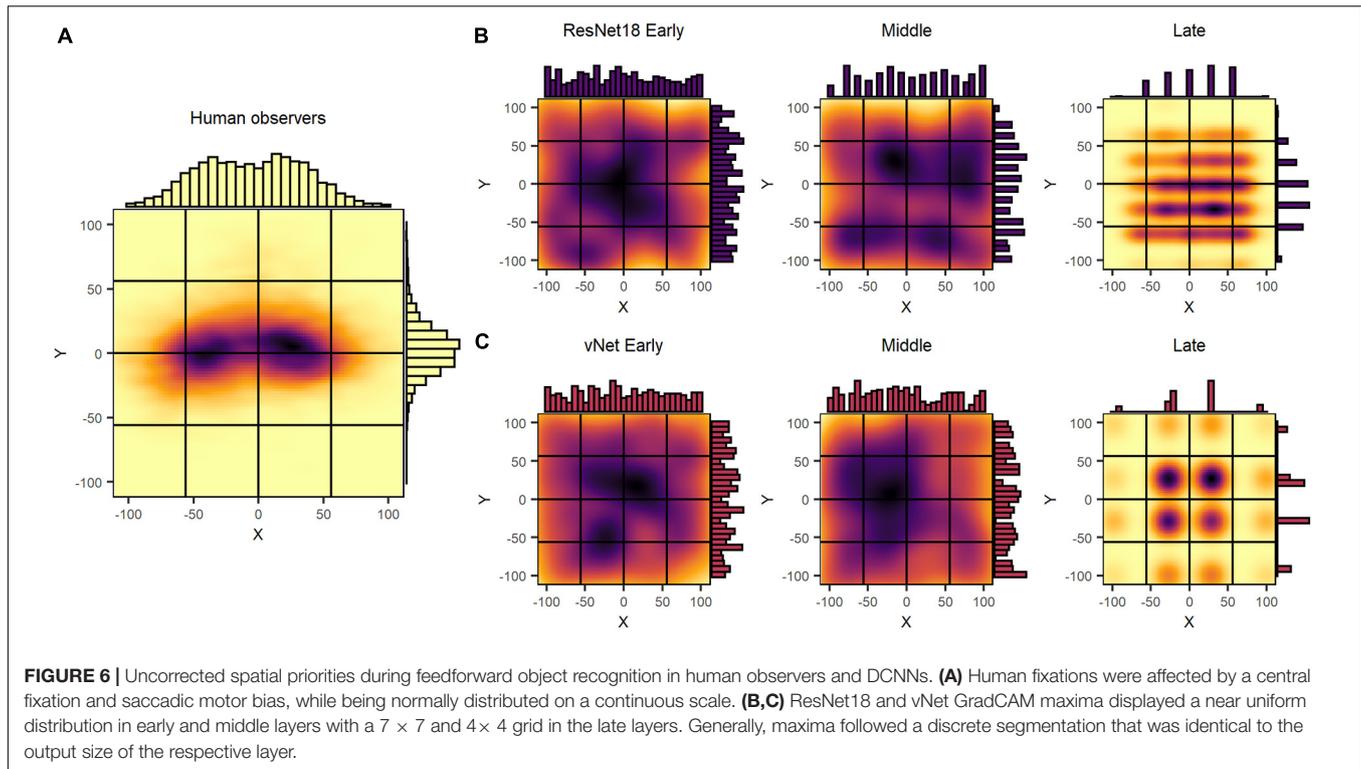
FIGURE 5 | Categorization matrices reveal the classification patterns of true vs. predicted labels that underlie raw object recognition performances and thereby help to understand especially classification errors. The off-diagonal misclassifications show that human observers seemed to have problems with conceptually similar classes (such as dog and cat or fish and snake). Interestingly, especially both trained-from-scratch ResNet18 and vNet made similar mistakes.

continuous scale of coordinates, GradCAM maxima, especially in middle and late layers, were located on discrete grids of different sizes. As these grids of possible maxima (ResNet18 Late = 7 × 7 and vNet Late = 4 × 4) were identical with the output sizes of the respective layers (see **Figures 1B,C**), we attributed this behavior to both the agglomerative nature of convolutions and the resulting technical aspects of how the GradCAM algorithm extracts class activations from layers.

To our knowledge, this characteristic of DCNN attribution methods has not been explicitly considered during previous human-machine comparisons so far. Since these findings restrict further comparisons of Euclidean distance measures between human fixations and GradCAM maxima, we proceeded by investigating only the specific target blocks, in which the respective fixations or maxima fell. As displayed in **Figure 7**,

this analysis promoted more similar object recognition priorities between human observers and vNet.

In principle, the agreement between human observers and DCNNs in target blocks (see **Figure 7**) was rather low. The results indicated persisting differences after correcting for the discovered mismatches in resolution, as vNet coincided with human target blocks in 17.14% compared to ResNet18 with 11.67%. Generally, the agreement seemed to be higher for animate objects compared to inanimate objects and vNet prioritized especially more human target blocks on images of specific animate objects (especially *human* and *dog*). These findings further strengthened our confidence that vNet compared to ResNet18 indeed utilizes spatial priorities that are more similar to those of human observers during feedforward object recognition. These insights offer compelling evidence for a more human-like vNet and

FIGURE 6 | Uncorrected spatial priorities during feedforward object recognition in human observers and DCNNs. **(A)** Human fixations were affected by a central fixation and saccadic motor bias, while being normally distributed on a continuous scale. **(B,C)** ResNet18 and vNet GradCAM maxima displayed a near uniform distribution in early and middle layers with a 7 × 7 and 4× 4 grid in the late layers. Generally, maxima followed a discrete segmentation that was identical to the output size of the respective layer.

generally the impact of a RFS modification in terms of fit to human eye tracking data.

Heatmaps—Recurrent vs. Feedforward Processing

Further analyses were conducted to compare human observers and DCNNs based on eye tracking and saliency heatmaps. Here, the focus was shifted away from feedforward mechanisms, as additionally recurrent processes with human eye movements after 150 ms were investigated. We hypothesized that differences, which had already existed during early processing (i.e., effects of animacy, arousal, and valence), should be amplified in the brain due to mostly top-down processes setting in during this time window. It is important to note that this analysis is rather unfair in its nature, as it compares feedforward processing in DCNNs with additional recurrent processing in humans. However, in the light of this knowledge, it is entirely possible to test further hypotheses. Generally speaking, the obtained heatmaps illustrated the expected idiosyncrasies. With a few exceptions, human observers fixated the specific object within the first milliseconds and later shifted their attention to more relevant features (such as faces or arousing image parts). In contrast, DCNNs displayed their hierarchical organization with activation of especially specific shape and texture features in early, feature groups in middle, and whole objects in late layers (see **Figure 8**).
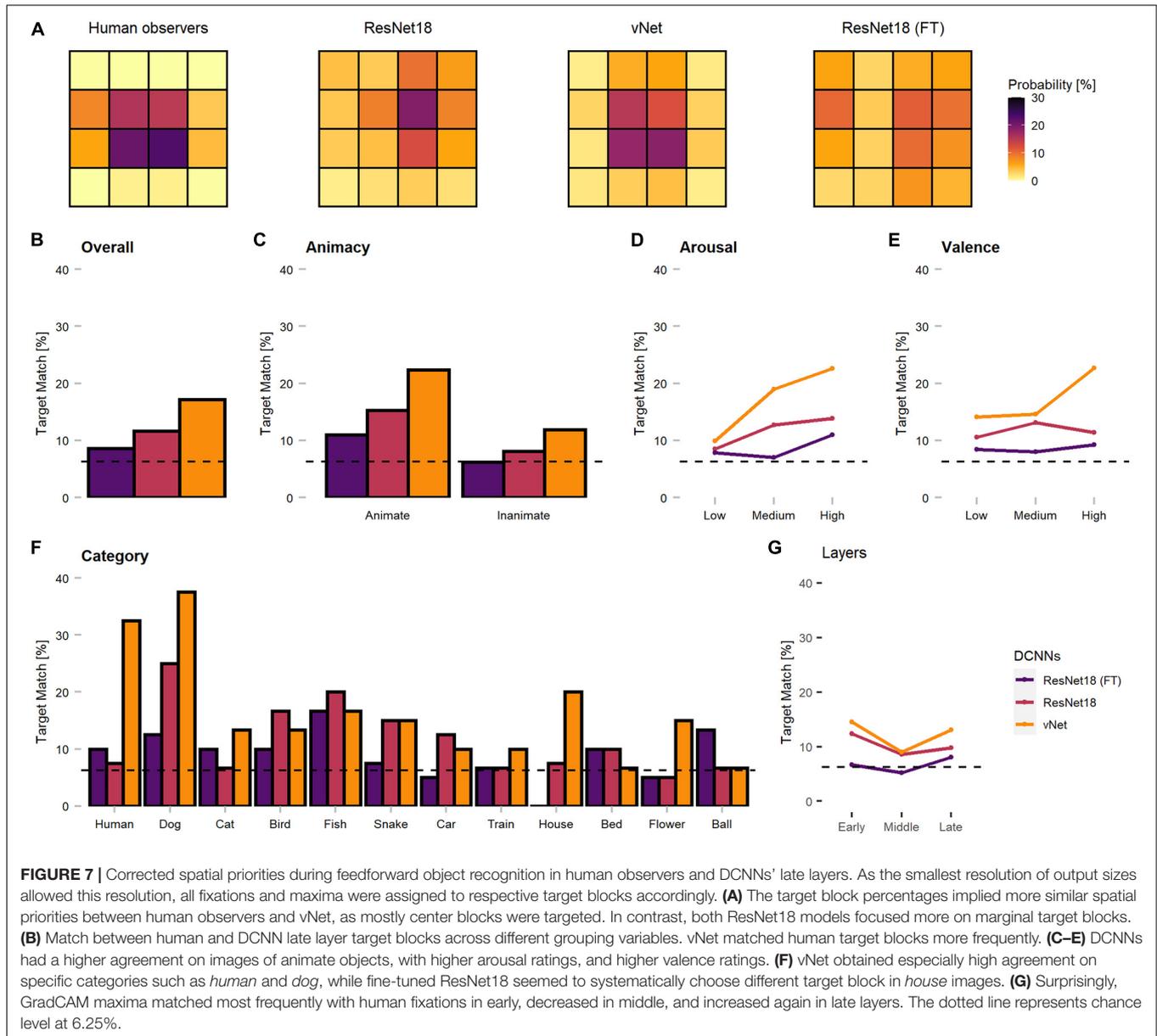
Mean absolute error (MAE), defined as the individual deviation of a GradCAM from its human heatmap equivalent, was computed for all individual images. On average, MAEs of 0.29 for ResNet18, and 0.22 for vNet were found. These results

fit well with previous findings by Ebrahimpour et al. (2019), who reported values of around 0.40 for a scene viewing task, and our previous outcomes promoting vNet as a better model for human eye tracking heatmaps. On top of that, MAE did not seem to be associated with general performance, as the fine-tuned ResNet18, which significantly outperformed human observers, reached only 0.32.

$$MAE = \frac{1}{Wx\,H} \sum_{x=1}^{W} \sum_{y=1}^{H} |E(x, y) - S(x, y)| \qquad (1)$$

Equation 1 mean absolute error (MAE) where W and H are the width and height of the original image in pixels (here 224 × 224), while E and S are the eye tracking heatmap and saliency map of the same image.

In order to link these results to the aforementioned control vs. challenge distinction by Kar et al. (2019), we treated images which were categorized correctly by human observers (avg. accuracy = 100%) and DCNNs as control images ($n = 145$) and images which were categorized correctly by human observers (avg. accuracy = 100%) but categorized incorrectly by ResNet18 and vNet as challenge images ($n = 60$). Analyses across individual layers showed that the reported difference in MAE between both DCNNs seemed to emerge in late layers (see **Figure 9**). These findings suggest that control and challenge images do not seem to be treated differently by the visual system in terms of their spatial priorities of information processing. Taken together with the findings of Kar et al. (2019) this means that the match between eye tracking and GradCAM data was not influenced by this distinction based on their difficulty.
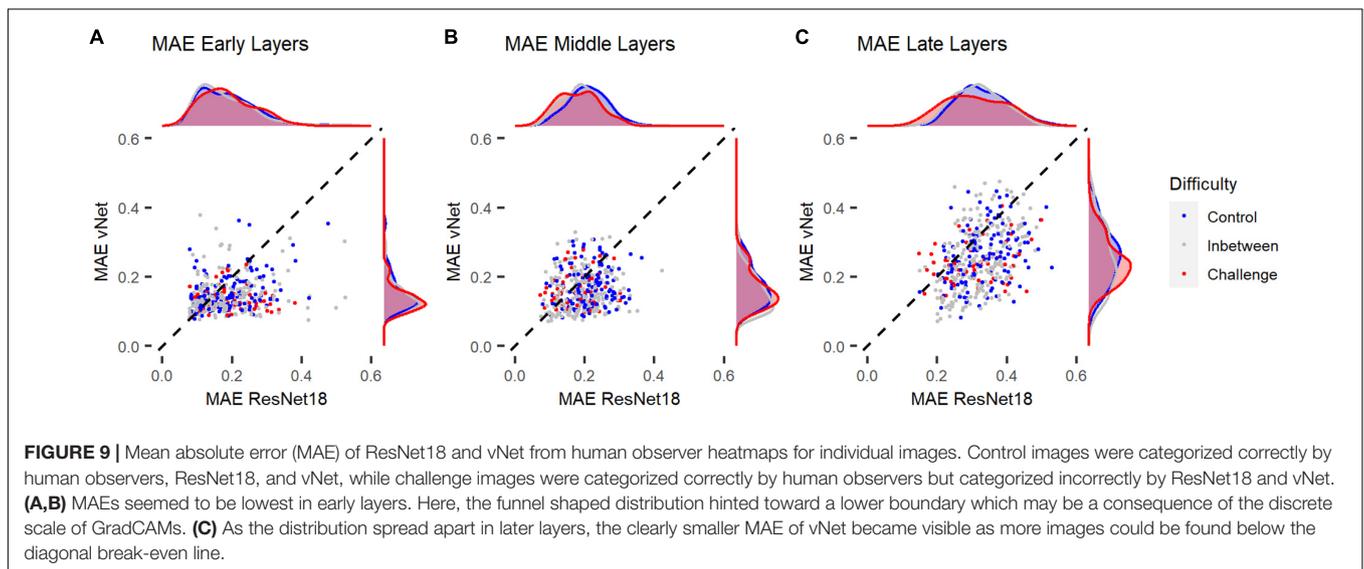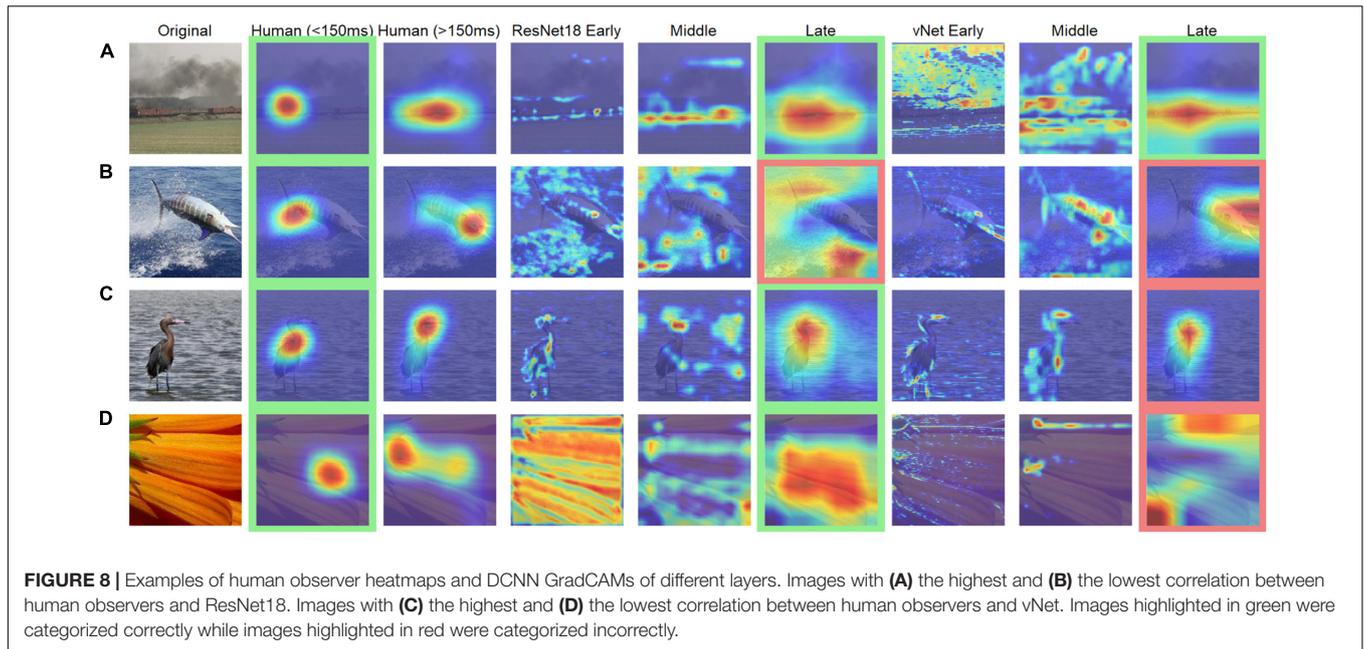
FIGURE 7 | Corrected spatial priorities during feedforward object recognition in human observers and DCNNs' late layers. As the smallest resolution of output sizes allowed this resolution, all fixations and maxima were assigned to respective target blocks accordingly. **(A)** The target block percentages implied more similar spatial priorities between human observers and vNet, as mostly center blocks were targeted. In contrast, both ResNet18 models focused more on marginal target blocks. **(B)** Match between human and DCNN late layer target blocks across different grouping variables. vNet matched human target blocks more frequently. **(C–E)** DCNNs had a higher agreement on images of animate objects, with higher arousal ratings, and higher valence ratings. **(F)** vNet obtained especially high agreement on specific categories such as *human* and *dog*, while fine-tuned ResNet18 seemed to systematically choose different target block in *house* images. **(G)** Surprisingly, GradCAM maxima matched most frequently with human fixations in early, decreased in middle, and increased again in late layers. The dotted line represents chance level at 6.25%.

Arousal and Valence

On average, human observers rated images with median scores of 4.04 in arousal and 4.27 in valence (see **Figure 10**). In terms of arousal, Kruskal Wallis tests suggested that animate objects received significantly higher scores (Median = 4.41) compared to inanimate objects (Median = 3.59; W = 37971, $p < 0.001$, $r = 0.28$), whereas for valence, no significant difference was discovered (W = 22555, $p = 0.685$, $r = 0.02$). Here, a substantial disagreement is evident, as mean heart rate [arousal: $X^2(2) = 0.62$, $p = 0.732$, $r = 0.02$/valence: $X^2(2) = 5.40$, $p = 0.067$, $r = 0.05$] and skin conductance response [arousal: $M = X$; $X^2(2) = 4.26$, $p = 0.119$, $r = 0.04$/valence: $M = X$; $X^2(2) = 0.53$, $p = 0.768$, $r = 0.03$] did not differ substantially between images of low, medium, and high arousal and valence ratings. However, available ratings from the OASIS dataset

(arousal: Median = 4.06/valence: Median = 3.92) were more or less consistent with our ratings.
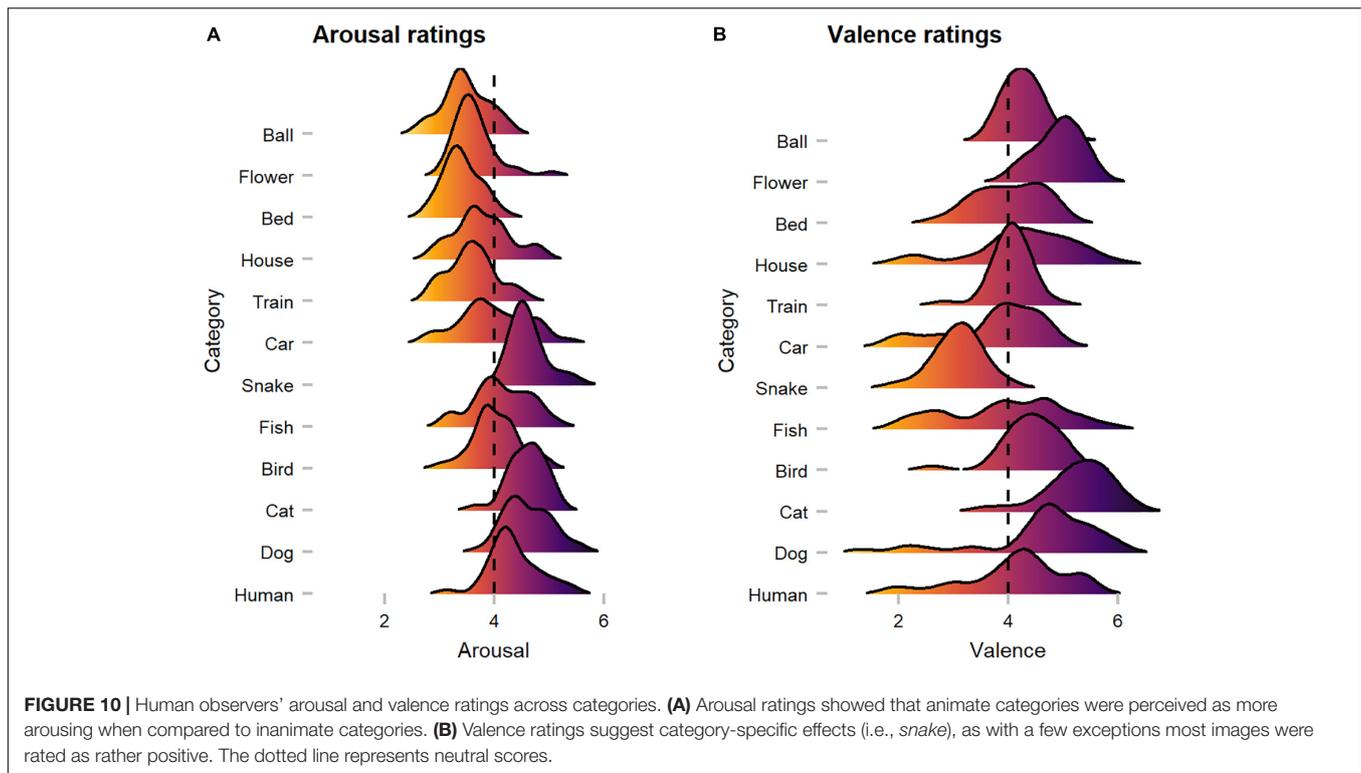
DISCUSSION

In this proof-of-concept study, we investigated the similarity of information processing in human observers and feedforward DCNN models during object recognition. For this purpose, human eye tracking heatmaps were compared to saliency maps of GradCAM, a customary attribution technique. Most importantly, during this endeavor, we found that GradCAM outputs, unlike eye tracking heatmaps, are produced on a discrete scale. While this is clear given the construction of DCNNs, to the best of our knowledge, this phenomenon has neither been regarded

FIGURE 8 | Examples of human observer heatmaps and DCNN GradCAMs of different layers. Images with **(A)** the highest and **(B)** the lowest correlation between human observers and ResNet18. Images with **(C)** the highest and **(D)** the lowest correlation between human observers and vNet. Images highlighted in green were categorized correctly while images highlighted in red were categorized incorrectly.

FIGURE 9 | Mean absolute error (MAE) of ResNet18 and vNet from human observer heatmaps for individual images. Control images were categorized correctly by human observers, ResNet18, and vNet, while challenge images were categorized correctly by human observers but categorized incorrectly by ResNet18 and vNet. **(A,B)** MAEs seemed to be lowest in early layers. Here, the funnel shaped distribution hinted toward a lower boundary which may be a consequence of the discrete scale of GradCAMs. **(C)** As the distribution spread apart in later layers, the clearly smaller MAE of vNet became visible as more images could be found below the diagonal break-even line.

in previous studies nor stated explicitly in the literature of the field. As a natural consequence, this finding constrains several established results, as for example by Ebrahimpour et al. (2019), and also our own heatmap comparisons, as different resolutions underlying these visualizations pose an evident challenge. Therefore, it seems necessary to draw attention to this fact, as it might endanger the fairness of human-machine comparisons (Firestone, 2020; Funke et al., 2020). We even believe to find more evidence for this problem as displayed in **Figure 9**, where especially in earlier layers MAE values seemed to hit a lower boundary which was possibly due to fundamental mismatches in resolution and therefore impossible to undercut. Nevertheless, final comparisons of fixations and maxima were not influenced by this effect, as we controlled for spatial inaccuracy

by proceeding with analyses on the target block level of the lowest maxima grid.

Generally, our results corroborate the assumption that the novel vNet architecture by Mehrer et al. (2021) captures human object recognition behavior more accurate compared to standard DCNNs commonly used throughout computer vision problems. Interestingly, this seemed to be the case on a performance level, as only vNet's performance was significantly correlated to human performance, and on a functional level of both before 150 ms, as it matched target blocks of human fixations spatially more consistent, and after 150 ms, where it yielded a lower MAE in late layers. On top of that, we argue that this higher similarity was not a side effect of higher performance compared to the equally trained ResNet18, as the fine-tuned ResNet18 model even

FIGURE 10 | Human observers' arousal and valence ratings across categories. **(A)** Arousal ratings showed that animate categories were perceived as more arousing when compared to inanimate categories. **(B)** Valence ratings suggest category-specific effects (i.e., *snake*), as with a few exceptions most images were rated as rather positive. The dotted line represents neutral scores.

outperformed human observers significantly and yet agreed the least with human fixations during feedforward processing. As already stated in previous literature (i.e., Dodge and Karam, 2017), the covariation in performance of human observers and DCNNs was found to be rather small. Nevertheless, these findings should be examined in the light of their relative impact, as vNet's correlation with human accuracy was not only of statistically significant importance, but also substantially higher. However, the reported effects should be treated with caution, as human-machine comparisons are prone to a wide range of confounding factors, either related to human cognition, such as the reported top-down and bottom-up processes, or related to deep learning problems, such as training settings and learning algorithms. We are aware of the fact that the number of parameters of both ResNet18 models and vNet is of necessity substantially different. In our view, these results emphasize the validity of a RFS modification as a method for designing more human-like models in computer vision. Moreover, as demonstrated by Luo et al. (2016), effective RFS follows a Gaussian distribution and become heavily increased by deep learning techniques such as subsampling (in most cases average or max pooling) and dilated convolutions, which are all commonly used in current architectures. Hence, as in contrast to DCNN saliency maps, human viewing behavior is highly focal, it would be interesting to see if the match of spatial priorities can be further increased in models that lack these computations. At the same time, future studies should target this topic by comparing DCNNs that only differ in RFS along their hierarchical architecture.

Moreover, we were able to identify control and challenge images based on the notion of Kar et al. (2019). Our analyses

suggested no effect of image difficulty on the similarity between eye tracking heatmaps and saliency maps. While the authors' original findings on neural activity show that control and challenge images are processed differently especially during late time periods (>150 ms), their results also propose that the two image groups share a similar early response and may not be treated differently by the visual system via the retina at all. Our reported null result regarding the viewing behavior may complement this line of argument well, as the authors even mentioned that on visual inspection no specific image properties differed between the groups. Unfortunately, in this case, the allocation to early- and late-solved images through neural recordings was not possible in the experimental setup at hand.

Surprisingly, in contradiction with earlier findings in humans (New et al., 2007), our results indicated that both human observers and DCNNs were more accurate in recognizing inanimate compared animate objects. This advantage for inanimate objects has been also reported on the level of basic categories by other studies before (Praß et al., 2013). Meanwhile, we discovered substantially more agreement between human and DCNN target blocks for animate objects. In turn, this suggests that both human observers and DCNNs were prioritizing more similar features during object recognition of animate objects. Therefore, we believe that this effect may be due to high efficient face processing mechanisms (Crouzet et al., 2010), which could have led to more similarity in specific *face-heavy* categories (such as *human* and *dog*). Furthermore, as animacy could not fully explain arousal and valence effects on a behavioral and

functional level of DCNNs, we argue that these effects could be a consequence of naturally learned optimization effects in visual systems, which have also been reported for image memorability judgments that automatically develop in DCNNs trained on object recognition and even predict variation in neural spiking activity (Jaegle et al., 2019; Rust and Mehrpour, 2020). This interpretation, however, needs to be treated with caution, as arousal and valence scores were obtained by human ratings which again underlie a wide range of effects such as for example acquired knowledge, attention, and memorability.

To summarize, in this paper we outline a novel concept of comparing human and computer vision during object recognition. In theory, this approach seems suitable for evaluating similarities and differences in priorities of information processing and may help to further pinpoint the specific impact of model adjustments toward more biological plausibility. We demonstrate this by showing that a RFS modification, which agrees conceptually with the ventral visual pathway, increases the model fit to human viewing behavior. Practically, we believe that our method will be improved by including different attribution techniques such as *Occlusion Sensitivity* (Zeiler and Fergus, 2014), which estimates class activations through a combination of occlusions and classifications, and thereby allow more adequate and comparable resolutions in the future. Furthermore, we hope that our idea will open up new perspectives on comparative vision at the intersection between biological and computer vision research.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

REFERENCES

Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.-S., et al. (2019). "Strike (with) a pose: neural networks are easily fooled by strange poses of familiar objects," in *Paper presented at the Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA. doi: 10.1109/CVPR.2019.00498

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *J. Cogn. Neurosci.* 15, 600–609. doi: 10.1162/089892903321662976

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., et al. (2006). Top-down facilitation of visual recognition. *Proc. Natl Acad. Sci. U.S.A.* 103:449. doi: 10.1073/pnas.0507062103

Beery, S., Van Horn, G., and Perona, P. (2018). "Recognition in Terra Incognita," in *Paper presented at the Proceedings of the European Conference on Computer Vision (ECCV)*, Munich. doi: 10.1007/978-3-030-01270-0_28

Blanchette, I. (2006). Snakes, spiders, guns, and syringes: how specific are evolutionary constraints on the detection of threatening stimuli? *Q. J. Exp. Psychol.* 59, 1484–1504. doi: 10.1080/02724980543000204

Blechert, J., Peyk, P., Liedlgruber, M., and Wilhelm, F. H. (2016). ANSLAB: integrated multichannel peripheral biosignal processing in psychophysiological science. *Behav. Res. Methods* 48, 1528–1545. doi: 10.3758/s13428-015-0665-1

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., et al. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput. Biol.* 10:e1003963. doi: 10.1371/journal.pcbi.1003963

Cauchoix, M., and Crouzet, S. (2013). How plausible is a subcortical account of rapid visual recognition? *Front. Hum. Neurosci.* 7:39. doi: 10.3389/fnhum.2013.00039

Cichy, R. M., Pantazis, D., and Oliva, A. (2014). Resolving human object recognition in space and time. *Nat. Neurosci.* 17, 455–462. doi: 10.1038/nn.3635

Contini, E. W., Wardle, S. G., and Carlson, T. A. (2017). Decoding the time-course of object recognition in the human brain: from visual features to categorical decisions. *Neuropsychologia* 105, 165–176. doi: 10.1016/j.neuropsychologia.2017.02.013

Crouzet, S. M., Kirchner, H., and Thorpe, S. J. (2010). Fast saccades toward faces: face detection in just 100 ms. *J. Vis.* 10, 16–16. doi: 10.1167/10.4.16

Deng, J., Dong, W., Socher, R., Li, L., Kai, L., and Li, F.-F. (2009). "ImageNet: a large-scale hierarchical image database," in *Paper Presented at the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL. doi: 10.1109/CVPR.2009.5206848

DiCarlo, J. J., and Cox, D. D. (2007). Untangling invariant object recognition. *Trends Cogn. Sci.* 11, 333–341. doi: 10.1016/j.tics.2007.06.010

DiCarlo, J. J., Zoccolan, D., Rust, and Nicole, C. (2012). How does the brain solve visual object recognition? *Neuron* 73, 415–434. doi: 10.1016/j.neuron.2012.01.010

Dodge, S., and Karam, L. (2017). "A study and comparison of human and deep learning recognition performance under visual distortions," in *Paper Presented at the 26th International Conference on Computer Communication and Networks (ICCCN)*, Vancouver, BC.

Ebrahimpour, M. K., Falandays, J. B., Spevack, S., and Noelle, D. C. (2019). "Do humans look where deep convolutional neural networks "attend"?," in *Paper*

Presented at the Advances in Visual Computing, Cham. doi: 10.1007/978-3-030-33723-0_5

Firestone, C. (2020). Performance vs. competence in human–machine comparisons. *Proc. Natl. Acad. Sci. U.S.A.* 117, 26562. doi: 10.1073/pnas.1905334117

Funke, C. M., Borowski, J., Stosio, K., Brendel, W., Wallis, T. S., and Bethge, M. (2020). The notorious difficulty of comparing human and machine perception. *arXiv* [Preprint] arXiv: 2004.09406, doi: 10.32470/CCN.2019.1295-0

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., et al. (2020). Shortcut learning in deep neural networks. *arXiv* [Preprint] arXiv:1312.6199, doi: 10.1038/s42256-020-00257-z

Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., and Wichmann, F. A. (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv* [Preprint] arXiv:1706.06969,

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., and Wichmann, F. A. (2018b). "Generalisation in humans and deep neural networks," in *Paper Presented at the Proceedings of the 32nd Conference on Neural Information Processing Systems (NIPS)*, Montréal, QC.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2018a). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv* [Preprint] arXiv:1811.12231,

Greene, M. R., and Oliva, A. (2009). Recognition of natural scenes from global properties: seeing the forest without representing the trees. *Cogn. Psychol.* 58, 137–176. doi: 10.1016/j.cogpsych.2008.06.001

Grill-Spector, K., Weiner, K. S., Kay, K., and Gomez, J. (2017). The functional neuroanatomy of human face perception. *Annu. Rev. Vis. Sci.* 3, 167–196. doi: 10.1146/annurev-vision-102016-061214

He, K., Zhang, X., Ren, S., and Sun, J. (2015). "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Paper Presented at the Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago. doi: 10.1109/ICCV.2015.123

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Paper Presented at the Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV. doi: 10.1109/CVPR.2016.90

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks," in *Paper Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI. doi: 10.1109/CVPR.2017.243

Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L., and Haxby, J. V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proc. Natl. Acad. Sci. U.S.A.* 96:9379. doi: 10.1073/pnas.96.16.9379

Jaegle, A., Mehrpour, V., Mohsenzadeh, Y., Meyer, T., Oliva, A., and Rust, N. (2019). Population response magnitude variation in inferotemporal cortex predicts image memorability. *eLife* 8:e47596. doi: 10.7554/eLife.47596

Kar, K., and DiCarlo, J. J. (2020). Fast recurrent processing via ventral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *bioRxiv* [Preprint] doi: 10.1101/2020.05.10.086959

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., and DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.* 22, 974–983. doi: 10.1038/s41593-019-0392-5

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Paper Presented at the Advances in Neural Information Processing Systems*, Lake Tahoe, NV.

Kurdi, B., Lozano, S., and Banaji, M. R. (2017). Introducing the open affective standardized image set (OASIS). *Behav. Res. Methods* 49, 457–470. doi: 10.3758/s13428-016-0715-3

Lamme, V. A., and Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* 23, 571–579. doi: 10.1016/s0166-2236(00)01657-x

Landau, B., Smith, L. B., and Jones, S. S. (1988). The importance of shape in early lexical learning. *Cogn. Dev.* 3, 299–321. doi: 10.1016/0885-2014(88)90014-7

Luo, W., Li, Y., Urtasun, R., and Zemel, R. (2016). "Understanding the effective receptive field in deep convolutional neural networks," in *Paper Presented at the Proceedings of the 30th International Conference on Neural Information Processing Systems*, Barcelona.

Marr, D. (1982). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. San Francisco, CA: W. H. Freeman.

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., and Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proc. Natl. Acad. Sci. U.S.A.* 118:e2011417118. doi: 10.1073/pnas.2011417118

New, J., Cosmides, L., and Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proc. Natl. Acad. Sci. U.S.A.* 104:16598. doi: 10.1073/pnas.0703913104

Öhman, A. (2005). The role of the amygdala in human fear: automatic detection of threat. *Psychoneuroendocrinology* 30, 953–958. doi: 10.1016/j.psyneuen.2005.03.019

Oliva, A., and Torralba, A. (2007). The role of context in object recognition. *Trends Cogn. Sci.* 11, 520–527. doi: 10.1016/j.tics.2007.09.009

Pessoa, L., and Adolphs, R. (2010). Emotion processing and the amygdala: from a 'low road' to 'many roads' of evaluating biological significance. *Nat. Rev. Neurosci.* 11, 773–782. doi: 10.1038/nrn2920

Praß, M., Grimsen, C., König, M., and Fahle, M. (2013). Ultra rapid object categorization: effects of level, animacy and context. *PLoS One* 8:e68051. doi: 10.1371/journal.pone.0068051

Rajaei, K., Mohsenzadeh, Y., Ebrahimpour, R., and Khaligh-Razavi, S.-M. (2019). Beyond core object recognition: recurrent processes account for object recognition under occlusion. *PLoS Comput. Biol.* 15:e1007001. doi: 10.1371/journal.pcbi.1007001

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025. doi: 10.1038/14819

Rolls, E. T. (2000). Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron* 27, 205–218. doi: 10.1016/s0896-6273(00)00030-1

Rosch, E. (1999). Principles of categorization. *Concepts* 189, 312–322. doi: 10.1016/B978-1-4832-1446-7.50028-5

Rothkegel, L. O. M., Trukenbrod, H. A., Schütt, H. H., Wichmann, F. A., and Engbert, R. (2017). Temporal evolution of the central fixation bias in scene viewing. *J. Vis.* 17:3. doi: 10.1167/17.13.3

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Rust, N. C., and Mehrpour, V. (2020). Understanding image memorability. *Trends Cogn. Sci.* 24, 557–568. doi: 10.1016/j.tics.2020.04.001

Rutishauser, U., Walther, D., Koch, C., and Perona, P. (2004). "Is bottom-up attention useful for object recognition?," in *Paper Presented at the Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*, Washington, DC.

Seijdel, N., Loke, J., van de Klundert, R., van der Meer, M., Quispel, E., van Gaal, S., et al. (2020). On the necessity of recurrent processing during object recognition: it depends on the need for scene segmentation. *bioRxiv* [Preprint] doi: 10.1101/2020.11.11.377655 bioRxiv: 2020.2011.2011.377655,

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Paper Presented at the Proceedings of the IEEE International Conference on Computer Vision*, Venice. doi: 10.1109/ICCV.2017.74

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *Paper Presented at the Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA. doi: 10.1109/CVPR.2015.7298594

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* 19, 109–139. doi: 10.1146/annurev.ne.19.030196.000545

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Ortega Caro, J., et al. (2018). Recurrent computations for visual pattern completion. *Proc. Natl. Acad. Sci. U.S.A.* 115, 8835–8840. doi: 10.1073/pnas.1719397115

Tatler, B. W. (2007). The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. *J. Vis.* 7, 4.1–17. doi: 10.1167/7.14.4

Tatler, B. W., Baddeley, R. J., and Vincent, B. T. (2006). The long and the short of it: spatial statistics at fixation vary with saccade amplitude and task. *Vis. Res.* 46, 1857–1862. doi: 10.1016/j.visres.2005.12.005

van Dyck, L. E., and Gruber, W. R. (2020). Seeing eye-to-eye? A comparison of object recognition performance in humans and deep convolutional neural networks under image manipulation. *arXiv* [Preprint] arXiv: 2007.06 294.,

Wandell, B. A., and Winawer, J. (2015). Computational neuroimaging and population receptive fields. *Trends Cogn. Sci.* 19, 349–357. doi: 10.1016/j.tics. 2015.03.009

Zeiler, M. D., and Fergus, R. (2014). "Visualizing and understanding convolutional networks," in *Paper presented at the European Conference on Computer Vision*, (Cham: Springer). doi: 10.1007/978-3-319-10590-1_53

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.