

Contributions of Shape, Texture, and Color in Visual Recognition

Yunhao Ge*, Yao Xiao*, Zhi Xu, Xingrui Wang, and Laurent Itti

University of Southern California

<https://github.com/gyhandy/Humanoid-Vision-Engine>

Abstract. We investigate the contributions of three important features of the human visual system (HVS) — shape, texture, and color — to object classification. We build a humanoid vision engine (HVE) that explicitly and separately computes shape, texture, and color features from images. The resulting feature vectors are then concatenated to support the final classification. We show that HVE can summarize and rank-order the contributions of the three features to object recognition. We use human experiments to confirm that both HVE and humans predominantly use some specific features to support the classification of specific classes (e.g., texture is the dominant feature to distinguish a zebra from other quadrupeds, both for humans and HVE). With the help of HVE, given any environment (dataset), we can summarize the most important features for the whole task (task-specific; e.g., color is the most important feature overall for classification with the CUB dataset), and for each class (class-specific; e.g., shape is the most important feature to recognize boats in the iLab-20M dataset). To demonstrate more usefulness of HVE, we use it to simulate the open-world zero-shot learning ability of humans with no attribute labeling. Finally, we show that HVE can also simulate human imagination ability with the combination of different features. We will open-source the HVE engine and corresponding datasets.

1 Introduction

The human vision system (HVS) is the gold standard for many current computer vision algorithms, on various challenging tasks: zero/few-shot learning [37,32,52,42,50], meta-learning [2,30], continual learning [45,54,60], novel view imagination [62,17], etc. Understanding the mechanism, function, and decision pipeline of HVS becomes more and more important. The vision systems of humans and other primates are highly differentiated. Although HVS provides us a unified image of the world around us, this picture has multiple facets or features, like shape, depth, motion, color, texture, etc. [16,23]. To understand the contributions of the most important three features — shape, texture, and color — in

* Yunhao Ge and Yao Xiao contributed equally

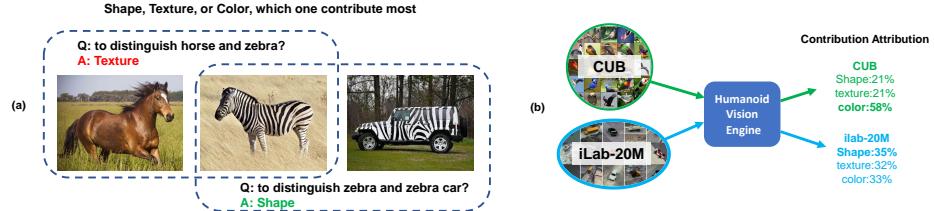


Fig. 1. Left: Contributions of Shape, Texture, and Color may be different among different scenarios/tasks. Here, texture is most important to distinguish zebra from horse, but shape is most important for zebra vs. zebra car. Right: Humanoid Vision Engine takes dataset as input and summarizes how shape, texture, and color contribute to the given recognition task in a pure learning manner (E.g., In ImageNet classification, shape is the most discriminative feature and contributes most to visual recognition).

visual recognition, some research compares the HVS with an artificial convolutional Neural Network (CNN). A widely accepted intuition about the success of CNNs on perceptual tasks is that CNNs are the most predictive models for the human ventral stream object recognition [7,61]. To understand which feature is more important for CNN-based recognition, recent paper shows promising results: ImageNet-trained CNNs are biased towards texture while increasing shape bias improves accuracy and robustness [33].

Due to the superb success of HVS on various complex tasks [37,2,45,62,19], human bias may also represent the most efficient way to solve vision tasks. And it is likely task-dependent (Fig. 1). Here, inspired by HVS, we wish to find a general way to understand how shape, texture, and color contribute to a recognition task by pure data-driven learning. The summarized feature contribution is important both for the deep learning community (guide the design of accuracy-driven models [33,22,15,6]) and for the neuroscience community (understanding the contributions or biases in human visual recognition) [35,59].

It has been shown by neuroscientists that there are separate neural pathways to process these different visual features in primates [1,11]. Among the many kinds of features crucial to visual recognition in humans, the shape property is the one that we primarily rely on in static object recognition [16]. Meanwhile, some previous studies show that surface-based cues also play a key role in our vision system. For example, [21] shows that scene recognition is faster for color images compared with grayscale ones and [40,38] found a special region in our brain to analyze textures. In summary, [9,8] propose that shape, color and texture are three separate components to identify an object.

To better understand the task-dependent contributions of these features, we build a Humanoid Vision Engine (HVE) to simulate HVS by explicitly and separately computing shape, texture, and color features to support image classification in an objective learning pipeline. HVE has the following key contributions: (1) Inspired by the specialist separation of the human brain on different features

[1,11], for each feature among shape, texture, and color, we design a specific feature extraction pipeline and representation learning model. (2) To summarize the contribution of features by end-to-end learning, we design an interpretable humanoid Neural Network (HNN) that aggregates the learned representation of three features and achieves object recognition, while also showing the contribution of each feature during decision. (3) We use HVE to analyze the contribution of shape, texture, and color on three different tasks subsampled from ImageNet. We conduct human experiments on the same tasks and show that both HVE and humans predominantly use some specific features to support object recognition of specific classes. (4) We use HVE to explore the contribution, relationship, and interaction of shape, texture, and color in visual recognition. Given any environment (dataset), HVE can summarize the most important features (among shape, texture, and color) for the whole task (task-specific) and for each class (class-specific). To the best of our knowledge, we provide the first fully objective, data-driven, and indeed first-order, quantitative measure of the respective contributions. (5) HVE can help guide accuracy-driven model design and performs as an evaluation metric for model bias. For more applications, we use HVE to simulate the open-world zero-shot learning ability of humans which needs no attribute labels. HVE can also simulate human imagination ability across features. We open-source the HVE engine and corresponding dataset.

2 Related Works

In recent years, more and more researchers focus on the interpretability and generalization of computer vision models like CNN [48,24] and vision transformer [12]. For CNN, many researchers try to explore what kind of information is most important for models to recognize objects. Some paper show that CNNs trained on the ImageNet are more sensitive to texture information [22,15,6]. But these works fail to quantitatively explain the contribution of shape, texture, color as different features, comprehensively in various datasets and situations. While most recent studies focus on the bias of Neural Networks, exploring the bias of humans or a humanoid learning manner is still under-explored and inspiring.

Besides, many researchers contribute to the generalization of computer vision models and focus on zero/few-shot learning [37,32,57,50,18,10], novel view imagination [62,17,20], open-world recognition [3,28,27], etc. Some of them tackled these problems by feature learning — representing an object by different features, and made significant progress in this area [55,39,62]. But, there still lacks a clear definition of what these properties look like or a uniform design of a system that can do humanoid tasks like generalized recognition and imagination.

3 Humanoid Vision Engine

The goal of the humanoid vision engine (HVE) is to summarize the contribution of shape, texture, and color in a given task (dataset) by separately computing the three features to support image classification, similar to the way that humans

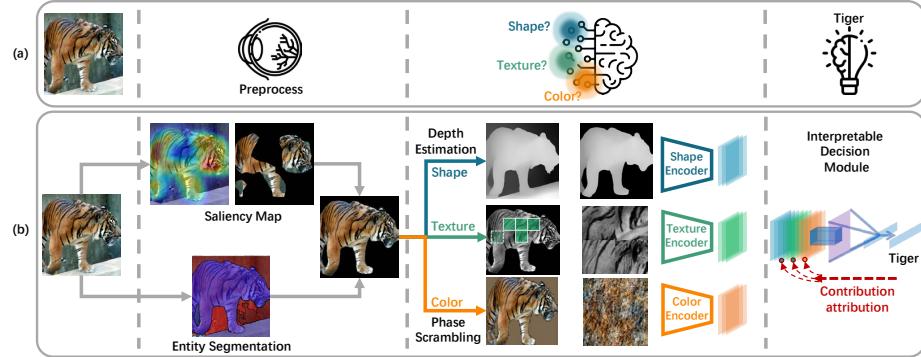


Fig. 2. Pipeline for humanoid vision engine (HVE). (a) shows how will humans' vision system deal with an image. After humans' eyes perceive the object, the different parts of the brain will be activated. The human brain will organize and summarize that information to get a conclusion. (b) shows how we design HVE to correspond to each part of the human's vision system.

recognize objects. During the pipeline and model design, we borrow the findings of neuroscience researchers on the structure, mechanism, and function of HVS [1,11,16,21,40,38]. We use end-to-end learning with backpropagation to simulate the learning process of humans and to summarize the contribution of shape, texture, and color. The advantage of end-to-end training is that we can avoid introducing human bias, which may influence the objective of contribution attribution (e.g., we do not introduce handcrafted elementary shapes as done in Recognition by Components [4]). We only use data-driven learning, a straightforward way to understand the contribution of each feature from effectiveness perspective, and we can easily generalize HVE to different tasks (datasets). As shown in Fig. 2, HVE consists of (1) a **humanoid image preprocessing pipeline**, (2) **feature representation** for shape, texture, and color, and (3) a **humanoid neural network** that aggregates the representation of each feature and achieves interpretable object recognition.

3.1 Humanoid Image Preprocessing and Feature Extraction

As shown in Fig. 2 (a), humans (or primates) can localize an object intuitively in a complex scene before we recognize what it is [29]. Also, there are different types of cells or receptors in our primary visual cortex extracting specific information (like color, shape, texture, shading, motion, etc) information from the image [16]. In our HVE (Fig. 2 (b)), for an input raw image $I \in \mathbb{R}^{H \times W \times C}$, we first parse the object from the scene as preprocessing and then extract our defined shape, texture, and color features I_s, I_t, I_c , for the following humanoid neural network.

Image Parsing and Foreground Identification. As shown in the preprocessing part of Fig.2 (b), we use the entity segmentation method [41] to simulate the process of parsing objects from a scene in our brain. Entity segmentation is an open-world model and can segment the object from the image without labels. This method aligns with human behavior, which can (at least in some cases; e.g., autostereograms [29]) segment an object without deciding what it is. After we get the segmentation of the image, we use a pre-trained CNN and GradCam [47] to find the foreground object among all masks. (More details in Supp.)

We design three different feature extractors after identifying the foreground object segment: shape extractor, texture extractor, and color extractor, similar to the separate neural pathways in the human brain which focus on specific property [1,11]. The three extractors focus only on the corresponding features, and the extracted features, shape I_s , texture I_t , and color I_c , are disentangled from each other.

Shape Feature Extractor For the shape extractor, we want to keep both 2D and 3D shape information while eliminating the information of texture and color. We first use a 3D depth prediction model [44,43] to obtain the 3D depth information of the whole image. After element-wise multiplying the 3D depth estimation and 2D mask of the object, we obtain our shape feature $I_s \in \mathbb{R}^{H \times W}$. We can notice that this feature only contains 2D shape and 3D structural information (the 3D depth) and without color or texture information (Fig. 2(b)).

Texture Feature Extractor In texture extractor, we want to keep both local and global texture information while eliminating shape and color information. Fig. 3 visualizes the extraction process. First, to remove the color information, we convert the RGB object segmentation to a grayscale image. Next, we cut this image into several square patches with an adaptive strategy (the patch size and location are adaptive with object sizes to cover more texture information). If the overlap ratio between the patch and the original 2D object segment is larger than a threshold τ , we add that patch to a patch pool (we set τ to be 0.99 in our experiments, which means the over 99% of the area of the patch belongs to the object). Since we want to extract both local (one patch) and global (whole image) texture information, we randomly select 4 patches from the patch pool and concatenate them into a new texture image (I_t). (More details in Supp.)

Color Feature Extractor We use two methods to represent the color feature for a given image I . The first method is phase scrambling, which is popular in psychophysics and in signal processing [36,53]. Phase scrambling transforms the image into the frequency domain using the fast Fourier transform (FFT). In the frequency domain, the phase of the signal is then randomly scrambled, which destroys shape information while preserving color statistics. Then we use IFFT to transfer back to image space and get $I_c \in \mathbb{R}^{H \times W \times C}$. I_c and I have the same distribution of pixel color values (Fig. 2(b)).

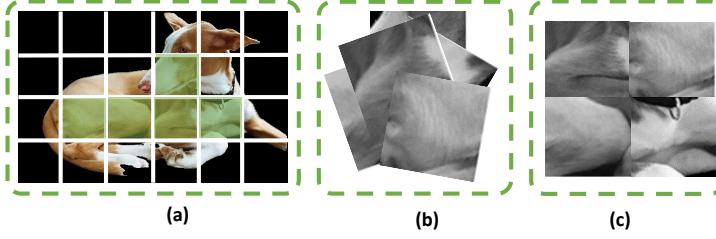


Fig. 3. Pipeline for extracting texture feature: (a) Crop images and compute the overlap ratio between 2D mask and patches. Patches with overlap > 0.99 are shown in a green shade. (b) add the valid patches to the patch pool. (c) randomly choose 4 patches from the pool and concatenate them to obtain a texture image I_t .

We also used simple color histograms (see suppl.) as an alternative, but the results were not as good, hence we focus here on the phase scrambling approach for color representation.

3.2 Humanoid Neural Network

After preprocessing, we have three features, i.e. shape I_s , texture I_t , color I_c of an input image I . To simulate the separate neural pathways in humans' brains for different feature information [1,11], we design three feature representation encoders for shape, texture, and color, respectively. Shape feature encoder E_s takes a 3D shape feature I_s as input and outputs the shape representation ($V_s = E_s(I_s)$). Similarly, texture encoder E_t and color encoder E_c take the texture patch image I_t or color phase scrambled image I_c as input, after embedded by E_t (or E_c), we get the texture feature V_t and color feature V_c .

We use ResNet-18 [24] as the backbone for all feature encoders to project the three types of features to the corresponding well-separated embedding spaces. It is hard to define the ground-truth label of the distance between features. Given that the objects from the same class are relatively consistent in shape, texture, and color, the encoders can be trained in the classification problem independently instead, with the supervision of class labels. After training our encoders as classifiers, the feature map of the last convolutional layer will serve as the final feature representation. (More details in the appendix.)

To aggregate separated feature representations and conduct object recognition, we freeze the three encoders and train a contribution interpretable aggregation module Aggr_θ , which is composed of two fully-connected layers (Fig. 2 (b) right). (More exploration in Appendix B.3). We concatenate V_s, V_t, V_c and send it to Aggr_θ . The output is denoted as $p \in \mathbb{R}^n$, where n is the number of classes.

$$p = \text{Aggr}_\theta(\text{concat}(V_s, V_t, V_c)). \quad (1)$$

We also propose a gradient-based *contribution attribution* method to interpret the contributions of shape, texture, and color to the classification decision,

respectively. Take the shape feature as an example, given a prediction p and the probability of class k , namely p^k , we compute the gradient of p^k with respect to the shape feature V_s . We define the gradient as shape importance weights α_s^k :

$$\alpha_s^k = \frac{\partial p^k}{\partial V_s} \quad \alpha_t^k = \frac{\partial p^k}{\partial V_t} \quad \alpha_c^k = \frac{\partial p^k}{\partial V_c}. \quad (2)$$

Then we can calculate element-wise product between V_s and α_s^k to get the final shape contribution S_s^k . In other words, S_s^k represents the “contribution” of shape feature to classifying this image as class k .

$$S_s^k = \text{ReLU} \left(\sum \alpha_s^k V_s \right). \quad (3)$$

We can do the same thing to get texture contribution S_t^k and color contribution S_c^k . After getting the feature contributions for each image, we can calculate the average value of all images in this class to assign feature contributions to each class (class-specific bias) and the average value of all classes to assign feature contributions to the whole dataset (task-specific bias).

4 Experiments

In this section, we first show the effectiveness of feature encoders on representation learning (Sec. 4.1); then we show the contribution interpretation performance of Humanoid NN on different feature-biased datasets in ImageNet (Sec. 4.2); We use human experiments to confirm that both HVE and humans predominantly use some specific features to support the classification of specific classes (Sec. 4.3); Then we use HVE to summarize the contribution of shape, texture, and color on different datasets (CUB[58] and iLab-20M[5]) (Sec. 4.4).

4.1 Effectiveness of Feature Encoders

To show that our three feature encoders focus on embedding their corresponding sensitive features, we handcrafted three subsets of ImageNet [31]: shape-biased dataset (D_{shape}), texture-biased dataset (D_{texture}), and color-biased dataset (D_{color}). **Shape-biased dataset** containing 12 classes, where the classes were chosen which intuitively are strongly determined by shape (e.g., vehicles are defined by shape more than color). **Texture-biased dataset** uses 14 classes which we believed are more strongly determined by texture. **Color-biased dataset** includes 17 classes. The intuition of class selection of all three datasets will be verified by our results in Table 1 with further illustration in Sec. 4.2. All these datasets use 80% / 20% of training and testing data (each class with around 800 training images and 200 testing images; images are randomly selected in each class). The details of classes contained in our biased datasets are shown in Fig. 4.

If our feature extractors actually learned their corresponding features with *feature-constructive* latent spaces, their T-SNE results will show clear clusters in

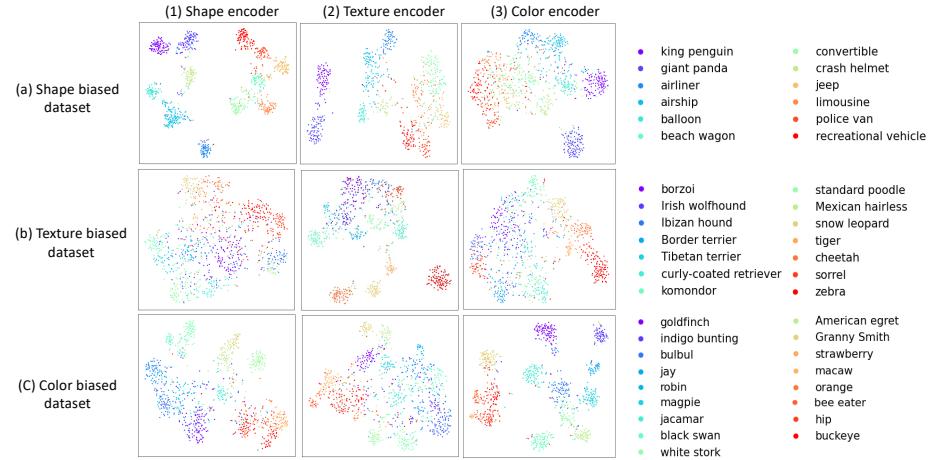


Fig. 4. T-SNE results of feature encoders on their corresponding biased datasets

Table 1. The “original” column means the accuracy of Resnet18 on the original images is our upper bound. “shape”, “texture” and “color” columns represent the accuracy of feature nets. The “all” column shows results of the Humanoid Neural Network that combines the 3 feature nets. It approaches the upper bound, suggesting that the split into 3 feature nets preserved most information needed for image classification.

accuracy	original	shape	texture	color	all
Shape biased dataset	97%	90%	84%	71%	95%
Texture biased dataset	96%	64%	81%	65%	91%
Color biased dataset	95%	70%	73%	82%	92%

the feature-biased datasets. “Bias” here means we can classify the objects based on the biased feature easily, but it is more difficult to make decisions based on the other two features. As shown in Fig. 1, distinguishing horse and zebra is a texture-biased task, while zebra vs. zebra car is a shape-biased task.

After pre-processing the original images and getting their feature images, we input the feature images into feature encoders and get the T-SNE results shown in Fig. 4. Each row represents one feature-biased dataset and each column is bounded with one feature encoder, each image shows the results of one combination. For instance, row (a) shows that the 12 classes in our shape-biased dataset are well separated in the embedding space of our shape encoder (column 1), but they are not separated in the embedding space of the texture (column 2) or color (column 3) encoders. Likewise color and texture encoders, their T-SNE results are separated perfectly on corresponding datasets (diagonal) but not as well on others’ datasets (off-diagonal), which shows that our feature encoders are predominantly sensitive to the corresponding features.

4.2 Effectiveness of Humanoid Neural Network

We can use feature encoders to serve as classifiers after adding fully-connected layers. As these classifiers classify images based on corresponding feature representation, we call them *feature nets*. We tested the accuracy of feature nets on these three biased datasets. As shown in Table 1, a ResNet-18 trained on the original segmented images (without explicit separated features, e.g. Fig. 2 (b) tiger without background) provided an upper bound for the task. We find that feature net consistently obtains the best performance on their own biased dataset (e.g., on the shape-biased dataset, shape net classification performance is better than that of the color net or texture net). If we combine these three feature nets with the interpretable aggregation module, the classification accuracy is very close to the upper bound, which means our vision system can classify images based on these three features almost as well as based on the full original color images. This demonstrates that we can obtain most information of original images by our feature nets, and our aggregation and interpretable decision module actually learned how to combine those three features by end-to-end learning.

Table 2 shows the quantitative contribution summary results of Humanoid NN (Sec. 3.2). For task-specific bias on each dataset, shape plays a dominant role in shape-biased tasks, and texture, color also contribute most to their related biased tasks.

Table 2. Contributions of features for different biased datasets.

contribution	shape ratio	texture ratio	color ratio
Shape biased dataset	47%	34%	19%
Texture biased dataset	5%	65%	30%
Color biased dataset	11%	19%	70%

4.3 Human Experiments

Intuitively, we expect that humans may rely on different features to classify different objects (Fig. 1). To show this, we designed human experiments that asked participants to classify reduced images with only shape, texture, or color features. If an object is mainly recognizable based on shape for humans, we could then check whether it is also the same for HVE, and likewise for color and texture.

Experiments Design As shown in Table. 1, the three datasets (Fig. 4) have a clear bias towards corresponding features. We asked the participants to classify objects in the corresponding dataset based on one single feature image computed by one of our feature extractors (Fig. 5). The reduced image contained only shape, texture, or color information. To make sure that the participants understood the class definitions well, we also showed them two example images



Fig. 5. Sample question for the human experiment. (a) A test image (left) is first converted into shape, color, and texture images using our feature extractors. (b) On a given trial, human participants are presented with one shape, color, or texture image, along with 2 reference images for each class in the corresponding dataset (not shown here, see suppl. for a screenshot of an experiment trial). Participants are asked to guess the correct object class from the feature image.

of each class at the bottom of the screen with the assigned class label written below (for instance, some participants may not have been familiar with what the "beach wagon" ImageNet class is). Participants were asked to choose the correct class label for the reduced image (from 12/14/17 classes in shape/texture/color datasets).

Human Performance Results The results here are based on 3270 trials, 109 participants. We publicly posted our questionnaires on the internet and sent the questionnaire link to the students and machine learning researchers in different universities. The accuracy for different feature questions on different biased datasets can be seen in Table 3. Human performance is quite similar to our feature nets' performance (compare Table 1 with Table 3). On the shape-biased dataset, both human and feature nets attain the highest accuracy with shape. The same for the color and texture biased datasets. Thus, both HVE and humans predominantly use some specific features to support object recognition of specific classes. Interestingly, humans can perform not badly on all three biased datasets with shape feature.

Table 3. Humans' recognition accuracy of different feature images on different biased datasets

	accuracy	shape	texture	color
Shape biased dataset	90.0%	49.0%	16.8%	
Texture biased dataset	33.1%	40.0%	11.1%	
Color biased dataset	32.3%	19.7%	46.5%	

4.4 Contributions Attribution in Different Tasks

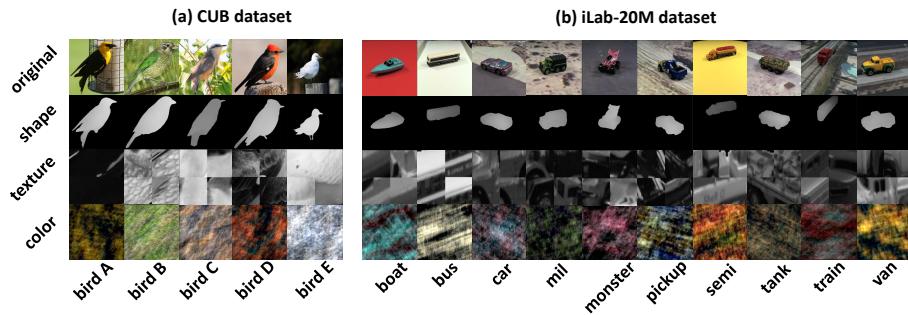
With our vision system, we can summarize the task-specific bias and class-specific bias for any dataset. This enables several applications: (1) Guide accuracy-driven model design; e.g., [33, 22, 15, 6] use dataset bias to guide the model design.

Table 4. class-specific bias for each class in iLab-20M

ratio	boat	bus	car	mil	monster	pickup	semi	tank	train	van
shape	40%	35%	44%	18%	36%	28%	40%	36%	31%	40%
texture	32%	31%	40%	30%	34%	20%	31%	32%	34%	27%
color	28%	34%	16%	52%	30%	53%	29%	32%	35%	33%

Our method provides objective summarization of dataset bias. (2) Evaluation metric for model bias. Our method can help correct an initially wrong model bias on some datasets (e.g., that most CNN trained on ImageNet are texture biased [22,33]). (3) Substitute human intuition to obtain more objective summarization with end-to-end learning. We implemented the biased summarization experiments on two datasets, CUB [58] and iLab-20M [5]. Fig. 1(b) shows the task-specific biased results. The results of CUB are reasonable since CUB is a dataset of birds, which means all the classes in CUB have a similar shape with feather textures, hence color may indeed be the most discriminative feature. We show sample images of CUB in Fig. 6 (a).

As for iLab (Fig. 6 (b)), its task-specific bias is so even that we cannot reach a clear conclusion about which feature is more important. So we implement the class-specific biased experiments on iLab and summarize the class biases in Table 4. It is interesting to find that the dominant feature is different for different classes. Take the boat as an example, it is strongly shape-biased, as indeed the shape of boats is so distinct in this dataset of vehicles. The results for military vehicles (mil) also deserve attention. The aggregation module shows that we can rely on color to distinguish mil. This conclusion is intuitive as the color of the mil is always green. We show more example images of iLab in the appendix.

**Fig. 6.** Overview of CUB and iLab-20M dataset

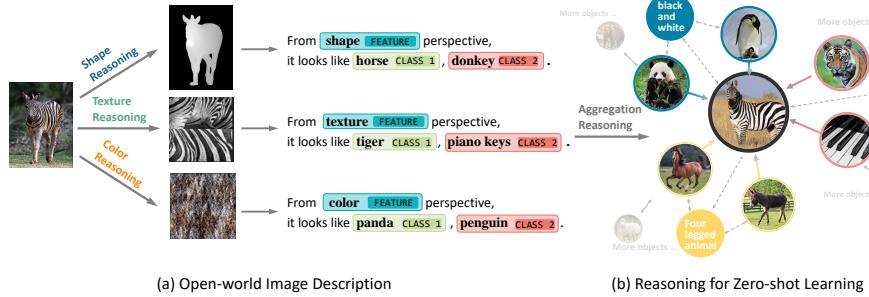


Fig. 7. The zero-shot learning method with HVE. We first describe the novel image in the perspective of shape, texture, and color. Then we use ConceptNet as common knowledge to reason and predict the label.

5 More Humanoid Applications with HVE

To further explore more applications with HVE, we use HVE to simulate the visual reasoning process of humans and propose a new solution for conducting open-world zero-shot learning without predefined attribute labels (Sec. 5.1). We also use HVE to simulate human imagination ability through cross-feature retrieval and imagination (Sec. 5.2).

5.1 Open-world Zero-shot Learning with HVE

Zero-shot learning is a challenging task, where, at test time, a learner needs to classify samples from classes never seen during training. Most current methods [37,32,13] need humans to provide detailed attribute labels for each image, which is costly in time and energy. However, given an image from an unseen class, humans can still *describe* it with their learned knowledge. For example, we may use horse-like shape, panda-like color, and tiger-like texture to describe an unseen class zebra, if we know horse, panda, and tiger. We use HVE to simulate this feature-wise open-world image description by feature retrieval and ranking (Sec. 5.1). Based on the image description, humans may conduct further *reasoning or consulting* with the help of an outside knowledge pool (e.g., WordNet [34]) to predict the class of the given unseen image. Similarly, we propose a feature-wise open-world zero-shot learning (reasoning) pipeline with the help of ConceptNet [51] (Sec. 5.1). The whole process can be seen in Fig. 7.

Step 1: Description In this section, we introduce how to use HVE to provide feature-wise description for any unseen class image without predefined attribute labels. First, to represent learnt knowledge, we use feature extractors E_s, E_t, E_c (described in Sec. 3.2) to get the shape, texture, and color representation $V_s^{(k,i)}, V_t^{(k,i)}, V_c^{(k,i)}$ of the i_{th} image of seen class k . Then, given an unseen

class image I_{un} , we use the same feature extractors to get its feature-wise representation V'_s, V'_t, V'_c . To retrieve learnt classes as description, we calculate the average distance d_m^k between I_{un} and images of other class k in the latent space on feature m . Here $m \in \{s, t, c\}$ is “shape”, “texture” or “color”.

$$d_m^k = \frac{1}{n_k} \sum_{i \in \mathcal{T}_k} d_m^{(k,i)} = \frac{1}{n_k} \sum_{i \in \mathcal{T}_k} \|V'_m - V_m^{(k,i)}\|_2, \quad (4)$$

where \mathcal{T}_k represents images of seen class k in the training set and n_k is the number of images in class k .

In this way, we can find the top K closest classes of I_{un} from shape perspective, and we call these K classes shape roots R_s . We do the same operation to texture and color representation and get R_t, R_c . Now, we can describe I_{un} using our roots. For example, as shown in Fig. 7(a), for the unseen class zebra, $R_s = \{\text{horse, donkey}\}$, $R_t = \{\text{tiger, piano keys}\}$, $R_c = \{\text{panda, penguin}\}$, and we can describe its shape, texture, and color based on these classes.

Step 2: Open-world classification To further predict the actual class of I_{un} based on the feature-wise description, we use ConceptNet as common knowledge to conduct reasoning. As shown in Fig. 7(b), for every feature root R_s, R_t, R_c , we retrieve their common attribute as $R_a^m = \{\bigcap_{r \in R_m} p_r\}$ in ConceptNet, where p_r is the neighbors of r in ConceptNet, (e.g., stripe is common attribute root of $R_t = \{\text{tiger, piano keys}\}$). We form a reasoning root pool R^* consisting of feature roots R_s, R_t, R_c obtained during image description, and shared attribute roots R_a^s, R_a^t, R_a^c . The reasoning roots will be our *evidence* for reasoning. For every root $r \in R^*$, we can search its neighbors in ConceptNet, which are treated as possible candidate classes for I_{un} . All candidates form a possible candidate pool P , which contains all hypothesis classes. Now we have two pools, root pool R^* and candidate pool P . For every candidate $p_i \in P$, we calculate the ranking score of p_i as:

$$\bar{S}(p_i) = \sum_{r_j \in R^*} \cos(\mathcal{E}(p_i), \mathcal{E}(r_j)), \quad (5)$$

where $\mathcal{E}(\cdot)$ is the word embedding in ConceptNet and $\cos(A, B)$ means cosine similarity between A and B .

We choose the candidate having the highest-ranking score as our predicted label. We build a prototype zero-shot learning dataset with 34 seen classes as the training set and 5 unseen classes as the test set, using 200 images per class. We calculate the accuracy of unseen classes (Table 5). Although not perfect, results are well above the chance level (1/5=20%). The performance of each feature extractor and encoder in HVE is the core of open-world classification applications. For baseline method, we conduct prototypical networks [49] on the same open-world zero-shot learning task with the same dataset and same backbone as our method. It is hard for the prototypical network to directly conduct zero-shot so we provide one-shot for each test class and use its one-shot setting. More details are in the appendix.

Table 5. Accuracy of unseen class for zero-shot learning

Method	fowl	zebra	wolf	sheep	apple
Prototype (one-shot)	19%	16%	17%	21%	74%
Ours (zero-shot)	78%	87%	63%	72%	98%

5.2 Cross Feature Imagination with HVE

We show HVE has the potential to simulate human imagination ability. We humans can intuitively imagine an object when seeing one aspect of a feature, especially when this feature is prototypical (contribute most to classification). For instance, we can imagine a zebra when seeing its stripe (texture). This process is similar but harder than the classical image generation task since the input features modality here *dynamic* which can be any feature among shape, texture, or color. To solve this problem, using HVE, we separate this procedure into two steps: (1) **cross feature retrieval** and (2) **cross feature imagination**. Given any feature (shape, texture, or color) as input, cross-feature retrieval focuses on finding the most possible two other features. Cross-feature imagination can then generate a whole object based on a group of shapes, textures, and color features.

Cross Feature Retrieval In order to reasonably retrieve the most possible other two corresponding features given only one feature (among shape, texture, or color), we learn a feature agnostic encoder that projects the three features into one same feature space and makes sure that the features belonging to the same class are in the nearby regions.

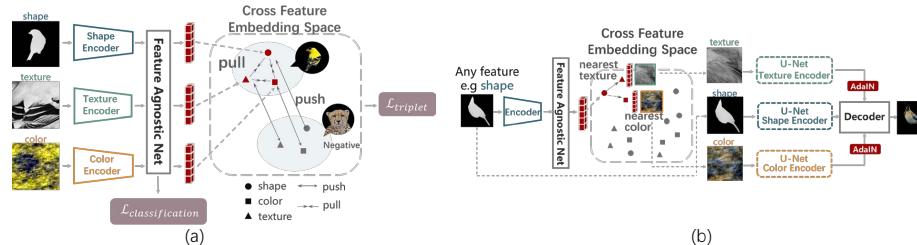


Fig. 8. (a) The structure and training process of the cross-feature retrieval model. E_s , E_t , E_c are the same encoders in Sec. 3.2. The feature agnostic net then projects them to shared feature space for retrieval. (b) The process of cross-feature imagination. After retrieval, we design a cross-feature pixel2pixel GAN model to generate the final image.

As shown in Fig. 8(a), during training, the shape I_s , texture I_t and color I_c are first sent into the corresponding frozen encoders E_s , E_t , E_c , which are the same

Table 6. Cross-features retrieval accuracy on biased datasets (DS).

input retrieval	shape			texture			color		
	shape	texture	color	shape	texture	color	shape	texture	color
shape biased DS	86%	81%	74%	76%	77%	66%	64%	61%	60%
texture biased DS	52%	51%	41%	67%	73%	63%	52%	57%	54%
color biased DS	59%	56%	54%	56%	61%	59%	67%	75%	75%

encoders in Sec. 3.2. Then all of the outputs are projected into a cross-feature embedding space by a feature agnostic net \mathcal{M} , which contains three convolution layers. We also add a fully connected layer to predict the class labels of the features. We use cross-entropy loss \mathcal{L}_{cls} to regularize the prediction label and a triplet loss $\mathcal{L}_{\text{triplet}}$ [46] to regularize the projection of \mathcal{M} . For any input feature x (e.g., a bird A shape), positive sample x_{pos} are either same class same modality (another bird A shape) or same class different feature modality (a bird A texture or color); negative sample x_{neg} are any features from different class. $\mathcal{L}_{\text{triplet}}$ pulls the embedding of x closer to that of the positive sample x_{pos} , and pushes it apart from the embedding of the negative sample x_{neg} . The triplet loss is defined as follow:

$$\mathcal{L}_{\text{triplet}} = \max(\|\mathcal{F}(x) - \mathcal{F}(x_{\text{pos}})\|_2 - \|\mathcal{F}(x) - \mathcal{F}(x_{\text{neg}})\|_2 + \alpha, 0), \quad (6)$$

where $\mathcal{F}(\cdot) := \mathcal{M}(E(\cdot))$, E is one of the feature encoders. α is the margin size in the feature space between classes, $\|\cdot\|_2$ represents ℓ_2 norm.

We test the retrieval model in all three biased datasets (Fig. 4) separately. In the retrieval process, given any feature of any object, we can map it into the cross feature embedding space by the corresponding encoder net and the feature agnostic net. Then we apply the ℓ_2 norm to find the other two features closest to the input one as output. The output is correct if they belong to the same class as the input.

For each dataset, we retrieve the three types of feature pair by pair, and the accuracy is shown in Table 6. The retrieval performs better when the input feature is the dominant of the dataset, which again verifies the feature bias in each dataset.

Cross Feature Imagination To stimulate imagination, we propose a cross-feature imagination model to generate a plausible final image with the input and retrieved features. The procedure of imagination is shown in Fig. 8(b). Inspired by the pixel2pixel GAN[26] and AdaIN[25] in the style transfer, we design a cross-feature pixel2pixel GAN model to generate the final image. The GAN model is trained and tested on the three biased datasets. In Fig. 9, we show more results of the generation model. The results show that our model satisfactorily generates the object from a single feature. From the comparison between (c) and (d), we can clearly find that they are alike from the perspective of the corresponding input feature, but the imagination results preserve the feature from the retrieval features. The imagination variance also shows the feature contributions from a generative view: if the given feature is the dominant feature

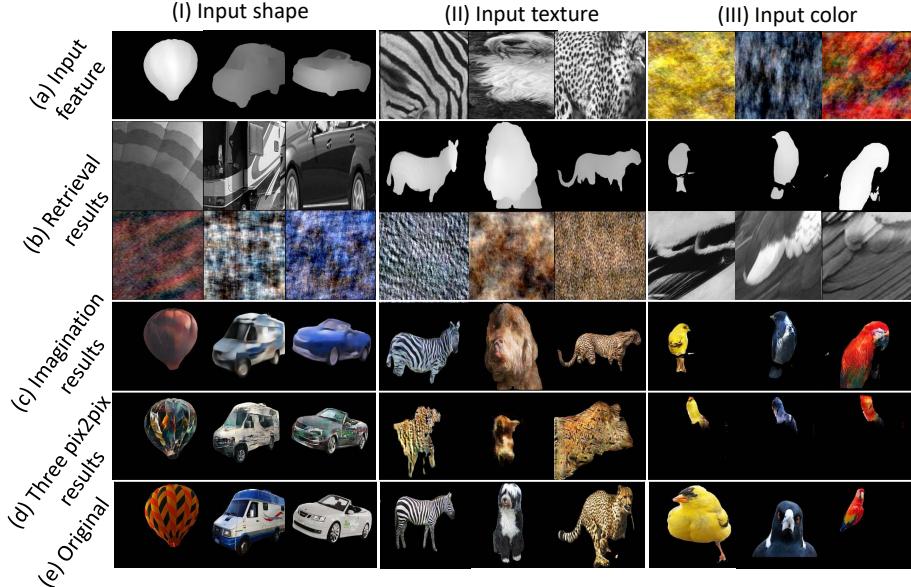


Fig. 9. Results of imagination with shape, texture, and color feature input (columns I, II, III). Line (a) is the input feature. Line (b) is the retrieved features given (a). Line (c) is the imagination results with HVE and our GAN model. Line (d) is results of baseline pix2pix GANs, which is only fed with input feature. Line (e) is the original images to which the input features belong. Our model can reasonably “imagine” the object given a single feature.

of a class (contribute most in classification. e.g., the stripe of zebra), then the retrieved features and imagined images have smaller variance (most are zebras); While non-dominant given feature (shape of zebra) lead to large imagination variance (can be any horse-like animals). We create a baseline by using three pix2pix GANs (same structure as ours), where each pix2pix GAN is responsible for one specific feature (e.g., take one modality of feature as input and imagine the raw image). FID comparison shows in Table. 7. More details of the GAN model are in the appendix.

Table 7. Cross-features imagination quality comparison.

	FID (\downarrow)	shape input	texture input	color input
Three pix2pix GANs	123.915	188.854	203.527	
Ours	96.871	105.921	52.846	

6 Conclusion

To explore the task-specific contribution of shape, texture, and color features in human visual recognition, we propose a humanoid vision engine (HVE) that explicitly and separately computes these features from images and then aggregates them to support image classification. With the proposed contribution attribution method, given any task (dataset), HVE can summarize and rank-order the task-specific contributions of the three features to object recognition. We use human experiments to show that HVE has a similar feature contribution to humans on specific tasks. We show that HVE can help simulate more complex and humanoid abilities (e.g., open-world zero-shot learning and cross-feature imagination) with promising performance. HVE is not a perfect simulation of HVS due to the limited learning ability and complexity of each component. These results are the first step towards better understanding the contributions of object features to classification, zero-shot learning, imagination, and beyond.

Acknowledgments This work was supported by C-BRIC (one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA), DARPA (HR00112190134) and the Army Research Office (W911NF2020053). The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.

References

1. Amir, Y., Harel, M., Malach, R.: Cortical hierarchy reflected in the organization of intrinsic connections in macaque monkey visual cortex. *Journal of Comparative Neurology* **334**(1), 19–46 (1993)
2. Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M.W., Pfau, D., Schaul, T., Shillingford, B., De Freitas, N.: Learning to learn by gradient descent by gradient descent. In: Advances in neural information processing systems. pp. 3981–3989 (2016)
3. Bendale, A., Boult, T.: Towards open world recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1893–1902 (2015)
4. Biederman, I.: Recognition-by-components: a theory of human image understanding. *Psychological review* **94**(2), 115 (1987)
5. Borji, A., Izadi, S., Itti, L.: ilab-20m: A large-scale controlled object dataset to investigate deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2221–2230 (2016)
6. Brendel, W., Bethge, M.: Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net (2019), <https://openreview.net/forum?id=SkfMWhAqYQ>
7. Cadieu, C.F., Hong, H., Yamins, D.L., Pinto, N., Ardia, D., Solomon, E.A., Majaj, N.J., DiCarlo, J.J.: Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology* **10**(12), e1003963 (2014)
8. Cant, J.S., Goodale, M.A.: Attention to form or surface properties modulates different regions of human occipitotemporal cortex. *Cerebral Cortex* **17**(3), 713–731 (2007)
9. Cant, J.S., Large, M.E., McCall, L., Goodale, M.A.: Independent processing of form, colour, and texture in object perception. *Perception* **37**(1), 57–78 (2008)
10. Cheng, H., Wang, Y., Li, H., Kot, A.C., Wen, B.: Disentangled feature representation for few-shot image classification. CoRR **abs/2109.12548** (2021), <https://arxiv.org/abs/2109.12548>
11. DeYoe, E.A., Carman, G.J., Bandettini, P., Glickman, S., Wieser, J., Cox, R., Miller, D., Neitz, J.: Mapping striate and extrastriate visual areas in human cerebral cortex. *Proceedings of the National Academy of Sciences* **93**(6), 2382–2386 (1996)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), <https://openreview.net/forum?id=YicbFdNTTy>
13. Fu, Y., Xiang, T., Jiang, Y.G., Xue, X., Sigal, L., Gong, S.: Recent advances in zero-shot recognition: Toward data-efficient understanding of visual content. *IEEE Signal Processing Magazine* **35**(1), 112–125 (2018)
14. Gao, M., Du, Y., Ai, H., Lao, S.: A hybrid approach to pedestrian clothing color attribute extraction. In: 2015 14th IAPR International Conference on Machine Vision Applications (MVA). pp. 81–84. IEEE (2015)
15. Gatys, L.A., Ecker, A.S., Bethge, M.: Texture and art with deep neural networks. *Current opinion in neurobiology* **46**, 178–186 (2017)

16. Gazzaniga, M.S., Ivry, R.B., Mangun, G.: Cognitive neuroscience. the biology of the mind,(2014) (2006)
17. Ge, Y., Abu-El-Haija, S., Xin, G., Itti, L.: Zero-shot synthesis with group-supervised learning. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), <https://openreview.net/forum?id=8wqCDnBmnrt>
18. Ge, Y., Xiao, Y., Xu, Z., Li, L., Wu, Z., Itti, L.: Towards generic interface for human-neural network knowledge exchange (2021)
19. Ge, Y., Xiao, Y., Xu, Z., Zheng, M., Karanam, S., Chen, T., Itti, L., Wu, Z.: A peek into the reasoning of neural networks: Interpreting with structural visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2195–2204 (2021)
20. Ge, Y., Xu, J., Zhao, B.N., Itti, L., Vineet, V.: Dall-e for detection: Language-driven context image synthesis for object detection. arXiv preprint arXiv:2206.09592 (2022)
21. Gegenfurtner, K.R., Rieger, J.: Sensory and cognitive contributions of color to the recognition of natural scenes. *Current Biology* **10**(13), 805–808 (2000)
22. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net (2019), <https://openreview.net/forum?id=Bygh9j09Kx>
23. Grill-Spector, K., Malach, R.: The human visual cortex. *Annu. Rev. Neurosci.* **27**, 649–677 (2004)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
25. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1501–1510 (2017)
26. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
27. Jain, L.P., Scheirer, W.J., Boult, T.E.: Multi-class open set recognition using probability of inclusion. In: European Conference on Computer Vision. pp. 393–409. Springer (2014)
28. Joseph, K., Khan, S., Khan, F.S., Balasubramanian, V.N.: Towards open world object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5830–5840 (2021)
29. Julesz, B.: Binocular depth perception without familiarity cues: Random-dot stereo images with controlled spatial and temporal properties clarify problems in stereopsis. *Science* **145**(3630), 356–362 (1964)
30. Khodadadeh, S., Böldöni, L., Shah, M.: Unsupervised meta-learning for few-shot image classification. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada.* pp. 10132–10142 (2019), <https://proceedings.neurips.cc/paper/2019/hash/fd0a5a5e367a0955d81278062ef37429-Abstract.html>

31. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25**, 1097–1105 (2012)
32. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 951–958. IEEE (2009)
33. Li, Y., Yu, Q., Tan, M., Mei, J., Tang, P., Shen, W., Yuille, A., Xie, C.: Shape-texture debiased neural network training. arXiv preprint arXiv:2010.05981 (2020)
34. Miller, G.A.: Wordnet: a lexical database for english. *Communications of the ACM* **38**(11), 39–41 (1995)
35. Oliva, A., Schyns, P.G.: Diagnostic colors mediate scene recognition. *Cognitive Psychology* **41**(2), 176–210 (2000). <https://doi.org/https://doi.org/10.1006/cogp.1999.0728>, <https://www.sciencedirect.com/science/article/pii/S001002859907284>
36. Oppenheim, A.V., Lim, J.S.: The importance of phase in signals. *Proceedings of the IEEE* **69**(5), 529–541 (1981)
37. Palatucci, M., Pomerleau, D., Hinton, G.E., Mitchell, T.M.: Zero-shot learning with semantic output codes. In: Bengio, Y., Schuurmans, D., Lafferty, J.D., Williams, C.K.I., Culotta, A. (eds.) *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada*. pp. 1410–1418. Curran Associates, Inc. (2009), <https://proceedings.neurips.cc/paper/2009/hash/1543843a4723ed2ab08e18053ae6dc5b-Abstract.html>
38. Peuskens, H., Claeys, K.G., Todd, J.T., Norman, J.F., Van Hecke, P., Orban, G.A.: Attention to 3-d shape, 3-d motion, and texture in 3-d structure from motion displays. *Journal of cognitive neuroscience* **16**(4), 665–682 (2004)
39. Prabhudesai, M., Lal, S., Patil, D., Tung, H., Harley, A.W., Fragkiadaki, K.: Disentangling 3d prototypical networks for few-shot concept learning. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), <https://openreview.net/forum?id=-Lr-u0b42he>
40. Puce, A., Allison, T., Asgari, M., Gore, J.C., McCarthy, G.: Differential sensitivity of human visual cortex to faces, letterstrings, and textures: a functional magnetic resonance imaging study. *Journal of neuroscience* **16**(16), 5205–5215 (1996)
41. Qi, L., Kuen, J., Wang, Y., Gu, J., Zhao, H., Lin, Z., Torr, P., Jia, J.: Open-world entity segmentation. arXiv preprint arXiv:2107.14228 (2021)
42. Rahman, S., Khan, S., Porikli, F.: A unified approach for conventional zero-shot, generalized zero-shot, and few-shot learning. *IEEE Transactions on Image Processing* **27**(11), 5652–5667 (2018)
43. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. ArXiv preprint (2021)
44. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2020)
45. Schlimmer, J.C., Fisher, D.: A case study of incremental concept induction. In: AAAI. vol. 86, pp. 496–501 (1986)
46. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)

47. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
48. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), <http://arxiv.org/abs/1409.1556>
49. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Advances in neural information processing systems **30** (2017)
50. Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 4077–4087 (2017), <https://proceedings.neurips.cc/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html>
51. Speer, R., Chin, J., Havasi, C.: Conceptnet 5.5: An open multilingual graph of general knowledge. In: Thirty-first AAAI conference on artificial intelligence (2017)
52. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1199–1208 (2018)
53. Thomson, M.G.: Visual coding and the phase structure of natural scenes. Network: Computation in Neural Systems **10**(2), 123 (1999)
54. Thrun, S., Mitchell, T.M.: Lifelong robot learning. Robotics and autonomous systems **15**(1-2), 25–46 (1995)
55. Tokmakov, P., Wang, Y.X., Hebert, M.: Learning compositional representations for few-shot recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6372–6381 (2019)
56. Van De Weijer, J., Schmid, C., Verbeek, J., Larlus, D.: Learning color names for real-world applications. IEEE Transactions on Image Processing **18**(7), 1512–1523 (2009)
57. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. Advances in neural information processing systems **29**, 3630–3638 (2016)
58. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
59. Walther, D.B., Chai, B., Caddigan, E., Beck, D.M., Fei-Fei, L.: Simple line drawings suffice for functional mri decoding of natural scene categories. Proceedings of the National Academy of Sciences **108**(23), 9661–9666 (2011). <https://doi.org/10.1073/pnas.1015666108>, <https://www.pnas.org/doi/abs/10.1073/pnas.1015666108>
60. Wen, S., Rios, A., Ge, Y., Itti, L.: Beneficial perturbation network for designing general adaptive artificial intelligence systems. IEEE Transactions on Neural Networks and Learning Systems (2021)
61. Yamins, D.L., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., DiCarlo, J.J.: Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the national academy of sciences **111**(23), 8619–8624 (2014)

62. Zhu, J.Y., Zhang, Z., Zhang, C., Wu, J., Torralba, A., Tenenbaum, J., Freeman, B.: Visual object networks: Image generation with disentangled 3d representations. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018), <https://proceedings.neurips.cc/paper/2018/file/92cc227532d17e56e07902b254dfad10-Paper.pdf>

Appendix

A Details of Human Vision Engine (HVE)

A.1 Image Parsing and Foreground Identification

As described in the main paper Sec. 3.1, we use entity segmentation and foreground object identification to simulate the preprocessing behavior of the human vision system. An illustration is shown in Fig. 10.

The entity segmentation can parse an input image and output a set of binary images masks. Each mask represents a single object or stuff in the image. For an input raw image $I_{\text{raw}} \in \mathbb{R}^{H \times W \times C}$, we denote the image mask sets as $\{I_{\text{mask},k}\}$, where $k = 1, 2, \dots$, denotes each different object. For each image mask $I_{\text{mask},k}$, pixel $I_{\text{mask},k}^{(i,j)}$ equals to 1 if and only if pixel $I_{\text{raw}}^{(i,j)}$ belongs to objects k , otherwise equals to 0.

For foreground identification, we borrow the learned knowledge from a pre-trained model which already learn the foreground class. As Grad-CAM [47] could generate class-specific saliency map M_{cam} given the model’s prediction, which represent how important each pixel contribute to the specific prediction. So the pixel with a higher activation value is more likely belong to foreground. We first use a pretrained model to generate a saliency map M_{cam} with Grad-Cam, and then we can get the binary attention map M_{att} based on M_{cam} . For mask M_{att} , pixel $M_{\text{att}}^{(i,j)}$ equals to 1 if and only if pixel $M_{\text{cam}}^{(i,j)} > \tau$ (we set τ to be the median value of M_{cam}), otherwise equals to 0.

After getting M_{att} , we can compute Intersection over union (IoU) score of each image mask $I_{\text{mask},k}$,

$$S_k = \text{SUM}(I_{\text{mask},k} \cap M_{\text{att}}) / \text{SUM}(I_{\text{mask},k})$$

Where $\text{SUM}()$ calculates the number of pixels. Then we choose the one with the highest score as the foreground mask.

A.2 Feature Extractor

In HVE, with the preprocessed image $I \in \mathbb{R}^{H \times W \times C}$, we use the three independent feature extractors to obtain the corresponding image feature: shape image (I_s), texture image (I_t), color image (I_c). Here, we will introduce more details about the texture and color extractor.

Texture Extractor Fig. 11 visualizes the process of extracting texture. First, to remove the color information, we convert the RGB object segmentation to a grayscale image. Since we want to get rid of the influence of background and extract the local texture feature as well as the global texture feature, we compute the maximum circumscribed rectangle of the object by its 2D mask and resize this rectangle part to get a 224×224 image. In this way, we can eliminate the

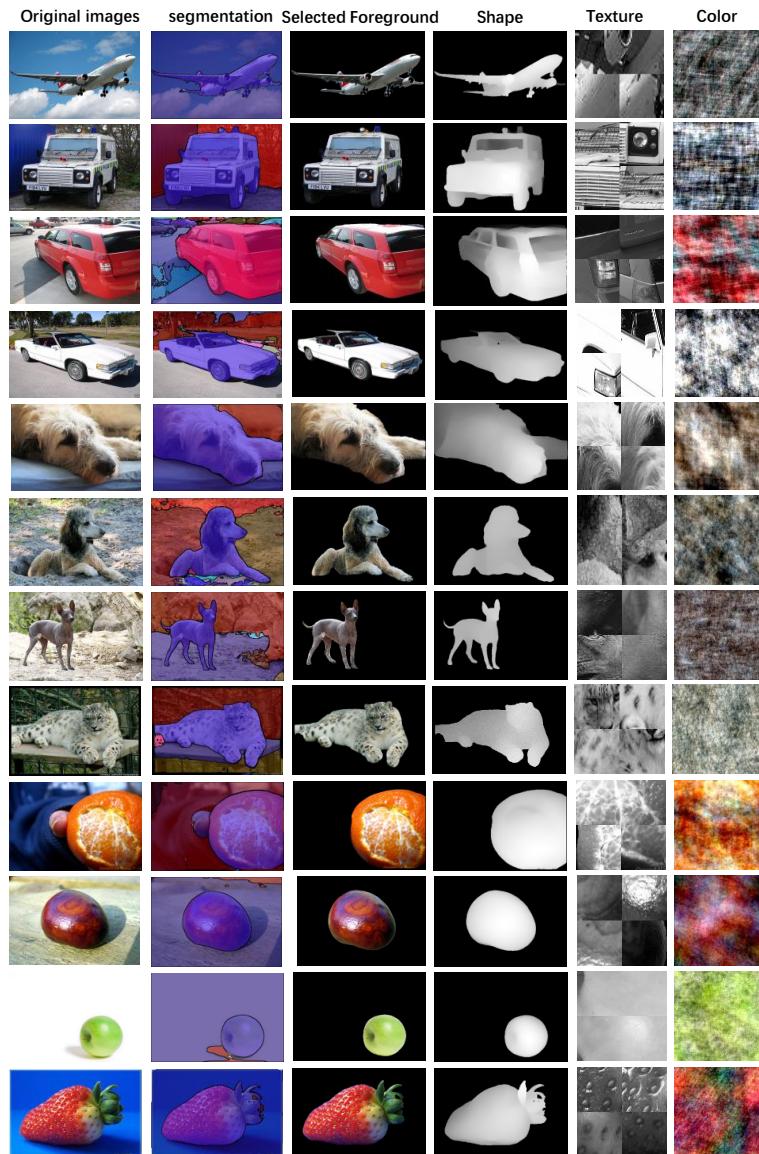
**Fig. 10.** Example of the preprocessing result and feature extraction.



Fig. 11. The process for extracting texture. We cut the image of the foreground object into several square patches as shown in (a) and select the patch pool as shown in (b). (c) is the final texture feature, the concatenation of k randomly selected patches from the patches pool.

most background and focus on the object. Next, we will cut this new image into several square patches. Specifically, We first cut this image into several square patches, as shown in Fig. 11 (a), if the overlap ratio between the patch and the original 2D object segment is larger than a threshold τ , we will add them to a patch pool (we set τ to be 0.99 in our experiments, which means the over 99% area of the patch belong to the object), as shown in Fig. 11 (b). Since we want to extract both local texture information (one patch) and global texture information (whole image), we randomly select k patches from the patch pool and concatenate them to a new texture image (I_t). We set $k = 4$ because we want to concatenate those patches to a square image, which means k should be a square number. If we set $k = 1$, we can get the maximum square patch of the object but we lose some small, local texture. If we set $k = 9$, the patch will be too small to contain useful information. To dynamically fit the object size and minimize the texture information loss. The number of patches we cut from the original image is dynamic. We will first cut the image into 9 square patches, if we can get $k = 4$ valid patches that can be added into the patch pool from these 9 patches, we will stop processing this image. However, if we cannot get 4 valid patches, we will cut the original images into 16, 25, 36... smaller patches until we can get 4 valid patches. After concatenation, we can get the texture image I_t , such as Fig. 11 (c). More results are shown in Fig. 10.

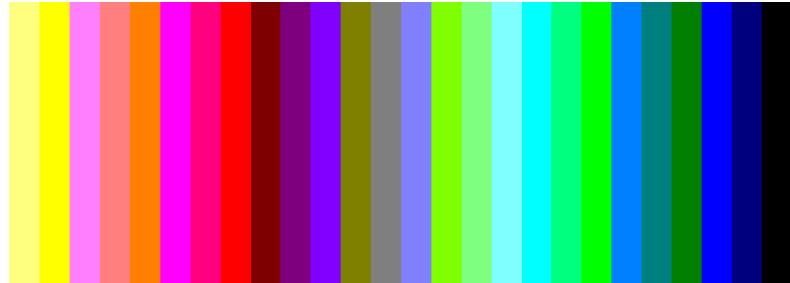


Fig. 12. 27 Color centers for the color block.

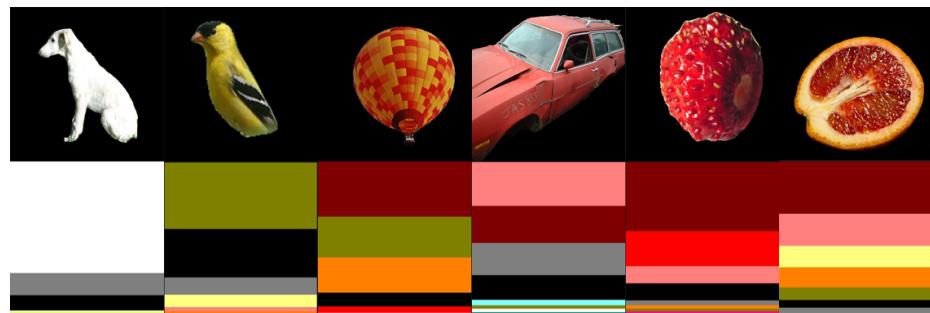


Fig. 13. Example of color block representation. The first row is original images, while the second row is the color blocks of those images.

Color Extractor We have two different ways to extract the color feature I_c , phase scrambling, and color blocks.

Phase scrambling For a given image $I \in \mathbb{R}^{H \times W \times C}$ and a random matrix $z \in \mathbb{R}^{H \times W}$, we use the 2D fast fourier transform (FFT) on channel j of the image and get the output $I_j^* \in \mathbb{C}^{H \times W}$, where $j = 1, 2, \dots, C$. We then caculate their modulus r_j and angle θ_j . Similarly, we apply the FTT to the random matrix z and get the transformed result $z^* \in \mathbb{C}^{H \times W}$ and its angle φ . With r_j, θ_j, φ , we can construct a new complex variable $T_j = r \cdot e^{i(\theta_j + p\varphi)}$, where $p \in [0, 1]$ is a scramble factor. After the mapping backing by 2D inverse fast fourier transform and rescale the result to range $[0, 255]$, the output will be the channel j of our color feature $I_c \in \mathbb{R}^{H \times W \times C}$. This process can be described formally as:

$$I_j^* = \text{FFT}(I_j) = r_j \cdot e^{i\theta_j} \quad (7)$$

$$z^* = \text{FFT}(z) = s \cdot e^{i\varphi}, \quad (8)$$

$$T_j = r \cdot e^{i(\theta_j + p\varphi)}, \quad (9)$$

$$I_{c,j} = \text{rescale}(\text{IFFT}(T_j)), \quad (10)$$

where $j = 1, 2, \dots, C$ and $I_{c,j}$ is the channel j of I_c . More results are shown in Fig. 10.

Color Blocks The second method uses statistical color histogram representation [56,14]. The original RGB color space is a three dimensional cubic $\{(r, g, b) | r, g, b \in [0, 255]\}$. In order to represent the distribution of color for each image, we first choose 27 center points which are uniformly distributed in the entire color space. The colors we choose are shown in Fig. 12. For an input image, we assign each pixel to its closest center point by calculating their manhattan distance. By counting how many pixels belong to each color center and calculating the percentage, we can summarize the result to a color block image of size 224×224 as our color feature I_c (the examples are shown in Fig. 13). The color block consists of various stripes in different widths and colors. The color of stripes comes from the color center and the width represents how many percent of pixels this center covered in the input image. For instance, if there are 10% pixels assign to white and 90% pixels of black, we will generate a image whose 10% pixels are RGB (255, 255, 255) and 90% pixels are RGB (0, 0, 0). This image (I_c) will not contain any shape or texture information. Fig. 13 shows some examples of the color blocks.

Compared with phase scrambling, this method is more intuitive to understand but may lose information when approximating RGB value in color space to color centers.

A.3 Details about the Humanoid Neural Network

The three encoders (E_s, E_t, E_c) use ResNet18 as backbone. E_s takes shape feature with size $224 \times 224 \times 1$ as input, while E_t takes texture feature with size

Table 8. Architecture of the humanoid neural network. N is the number of labels, depending on which dataset is being used.

Layer	Input → Output Shape	Layer Information
Shape Encoder	(224, 224, 1) → (7, 7, 512)	ResNet18
Texture Encoder	(224, 224, 1) → (7, 7, 512)	ResNet18
Color Encoder	(224, 224, 3) → (7, 7, 512)	ResNet18
Concatenation Layer	$3 \times (7, 7, 512) \rightarrow (7, 7, 1536)$	-
Pooling Layer	$(7, 7, 1536) \rightarrow (1, 1, 1536)$	Average Polling-($k7 \times 7, s1, p0$)
Flatten Layer	$(1, 1, 1536) \rightarrow 1536$	-
Hidden Layer	$1536 \rightarrow 512$	Linear, ReLU
Output Layer	$512 \rightarrow N$	Linear, Softmax

$224 \times 224 \times 1$ and E_c takes color feature with size $224 \times 224 \times 3$. They all produce the feature with size $7 \times 7 \times 512$. When training the encoders, these output features are then passed into a fully connected layer, which output the vectors with length N , the number of classes in the dataset.

During training the interpretable aggregation module, we freeze the three encoders and concatenate their output features along the channel dimension into a tensor of size $7 \times 7 \times 1536$. This tensor is the input of the interpretable aggregation module, which is a two-layer MLP. The final output is a vector with length N , the number of classes in the dataset. The summary of the structure of the three encoders and the aggregation module is shown in Table 8.

We use Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and set batch size to 64, learning rate is 0.001.

B Experiments Details

B.1 Influence of Random Selection in Texture Feature Extractor

We re-ran 10 times main paper Table.2 using different random seeds for texture (patch selections and shuffling the order sequence). The mean and std shown in Table. 9.

B.2 More quantitative contribution summary results of Humanoid NN

To further evaluate the contribution of shape texture and color, for each feature V (e.g., shape feature Vs), we compute the accuracy if we only use the rest features (e.g., combine texture Vt and color Vc) as input, and calculate the

Table 9. Influence of Random Selection in Texture Feature Extractor.

contribution	Shape	Texture	Color
Shape-bias DS	$47.1\% \pm 2.7\%$	$34.5\% \pm 2.2\%$	$18.4\% \pm 1.4\%$
Texture-bias DS	$5.2\% \pm 1.8\%$	$62.4\% \pm 2.3\%$	$32.3\% \pm 1.6\%$
Color-bias DS	$12.4\% \pm 2.1\%$	$19.1\% \pm 1.9\%$	$68.5\% \pm 3.6\%$

accuracy drop compared with using all three features as input (last column in Table 1). That accuracy drop represents the “unsubstitutability” or the “necessity” or “non-redundant contribution” of that feature for visual recognition. The results (Table. 10) are consistent with our previous results (main paper Table 1) on the contribution/importance of each feature. For example, in the shape biased dataset, the largest accuracy drop is when we remove the shape feature.

Table 10. More quantitative contribution summary results of Humanoid NN

accuracy	Shape + Texture	Shape + Color	Texture + Color	all
Shape biased DS	94%	92%	86%	95%
Texture biased DS	89%	80%	83%	90%
Color biased DS	83%	88%	87%	92%

B.3 Different interpretable aggregation module

We conducted experiments to substitute the non-linear MLP with simple average pooling followed by an output classification layer. Contributions are shown in table. 11. While the numerical results differ, the ordering and conclusions remain (e.g., shape texture is most important in the shape-biased dataset). The experiments results show that the contribution result is robust to the aggregation module.

Table 11. Average pooling interpretable aggregation module

Ave-Pool	Shape ratio	Texture ratio	Color ratio
Shape biased DS	48%	40%	12%
Texture biased DS	1%	79%	20%
Color biased DS	1%	4%	95%

B.4 Sample Question of Human Experiments

We designed human experiments that asked participants to classify reduced images with only shape, texture, or color features. Here, as shown in Fig. 14, we demonstrate a screenshot of an experiment trial in our experiments.

B.5 Contributions of Features in Different Tasks

Table 12. Local bias for CUB classes.

ratio		shape	texture	color
Bewick Wren	44.8%	21.3	33.8%	
Gadwall	43.6%	24.2%	32.2%	
Brown Pelican	41.6%	33.0%	25.4%	
American Pipit	6.8%	72.7%	20.4%	
Eared Grebe	5.6%	70.3%	24.1%	
Harris Sparrow	13.5%	64.1%	22.4%	
Kentucky Warbler	0.1%	0.7%	99.2%	
Cape May Warbler	0.3%	1.2%	98.5%	
Gray crowned Rosy Finch	1.1%	1.5%	97.4%	

Fig. 15 shows more example images of the iLab dataset [5]. In the figure, we can see that the military vehicles are always in the unique green, which matches with the result in our experiment that color is the most discriminative feature to classify military among other vehicles.

We also compute local bias for each class in CUB dataset [58]. For each feature among shape, texture, and color, we select the top 3 classes whose feature contributions are the highest, and we show the bias for these 9 classes in Table. 12. To show more details of these top classes for each feature, we show the example images in Fig. 16. For the first three-row classes, the shape is the most discriminative feature to classify them among other birds; for the middle three-row classes, the texture is the most discriminative feature, while color is the most discriminative feature to classify the last three-row classes.

C Details of Humanoid Application

C.1 Process of Open-world Zero-shot Learning

As describe in main paper Sec.5.1, we conduct open-world zero-shot learning with HVE. Here, we describe more details about the experiment. Given an unseen class image, we can use open world description (Step 1) to get its shape root, texture root, color root. For example, given a wolf image, we can get $R_s = \{\text{hyena, fox, tiger}\}$, $R_t = \{\text{fox, dog}\}$, $R_c = \{\text{dog}\}$. We provide more examples in table.13.



If we only show the shape information of one object, which class do you think this object belongs to? (we show two sample images for each class)

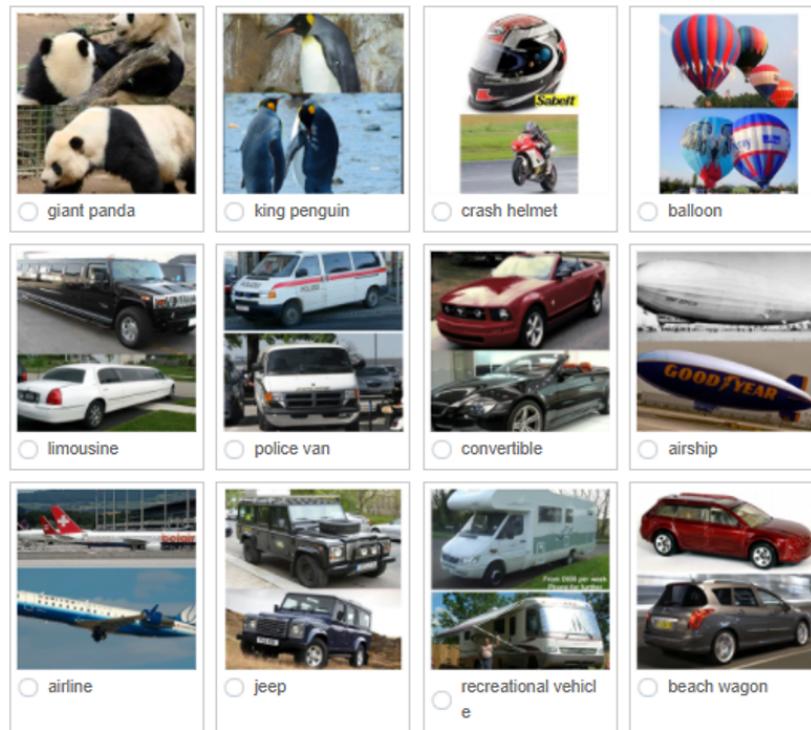


Fig. 14. A screenshot of an experiment trial in our human experiments.

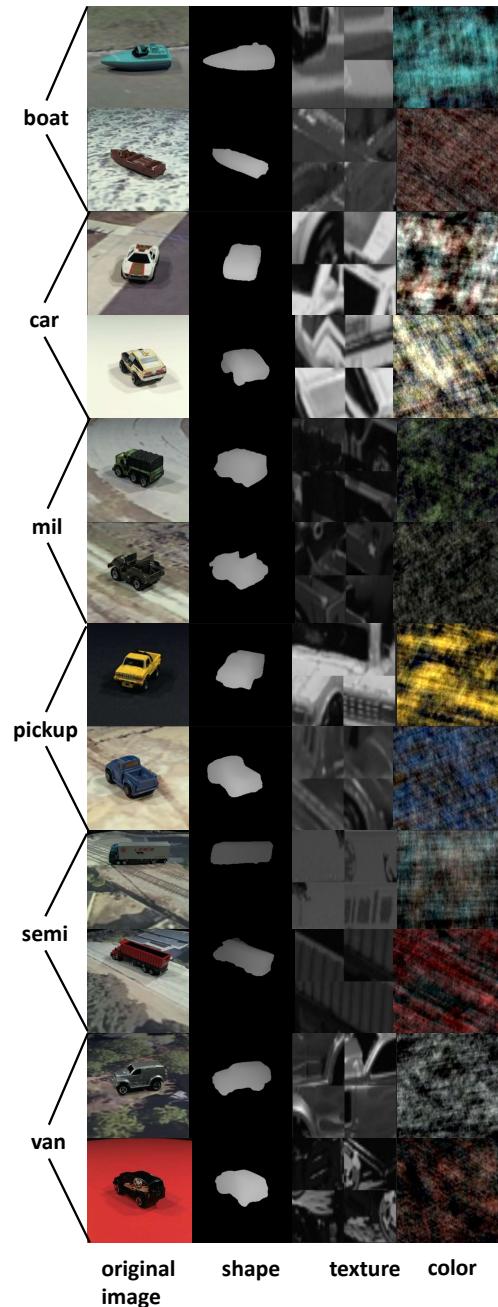


Fig. 15. Example images for ilab-20M classes.

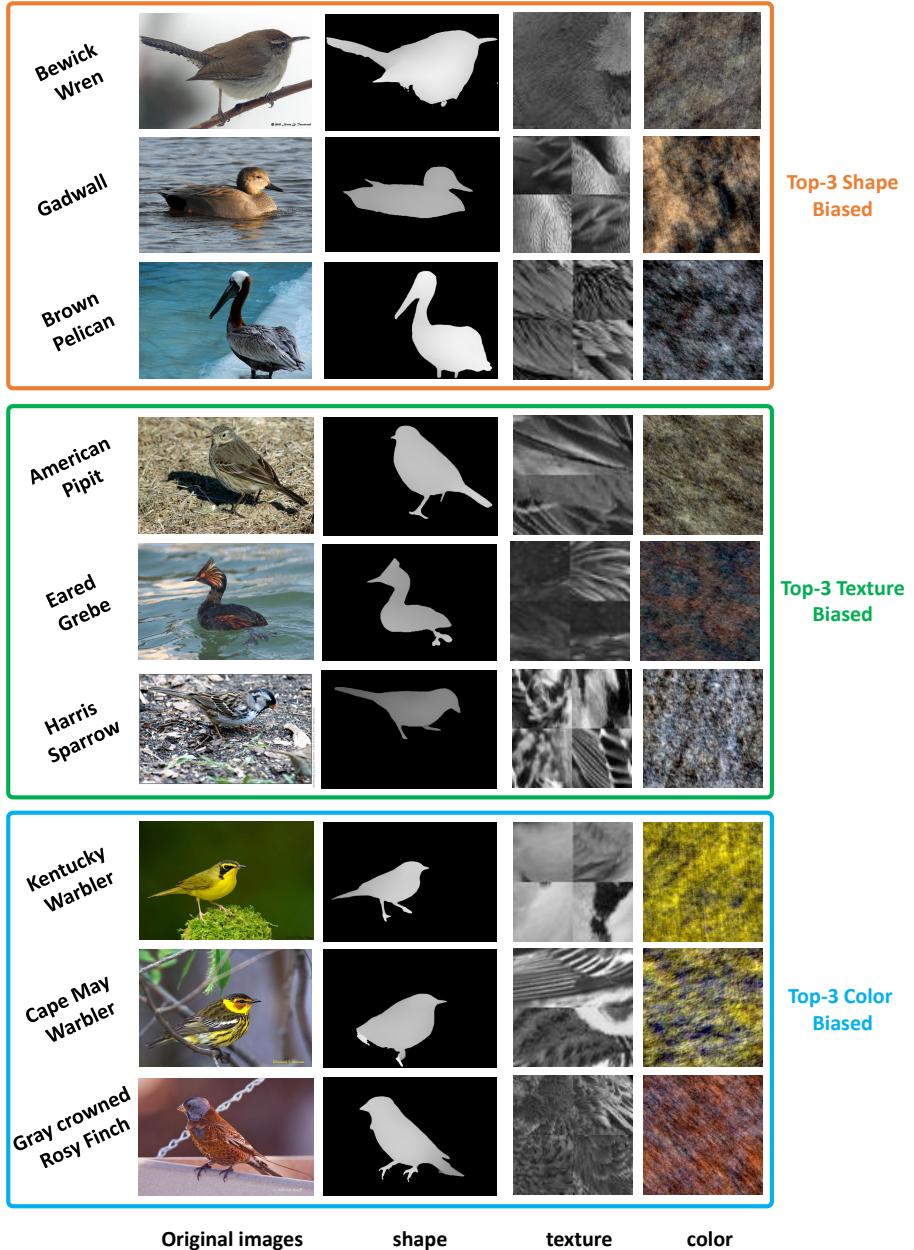


Fig. 16. Example images for CUB classes. The first 3 rows show the top-3 classes in CUB whose shape contribution is highest; The middle 3 rows show top-3 classes whose texture contribution is highest; The last 3 rows show top-3 classes whose color contribution is highest. Table. 12 shows the quantity biased results for these classes.

Class	zebra	fowl	wolf	sheep	apple
shape root	horse, donkey	turkey, goose cock	hyena, fox, tiger	goat, bull	tomato, orange, pear
texture root	tiger, piano keys	turkey, penguin	fox, dog	dog, goat	cherry, tomato
color root	panda, penguin	turkey	dog	elephant, goat	cherry, tomato

Table 13. More open-world zero-shot learning result.

After we get $R_s = \{\text{hyena, fox, tiger}\}$, $R_t = \{\text{fox, dog}\}$, $R_c = \{\text{dog}\}$, we use ConceptNet [51] and word embedding to predict its label (Step 2 in main paper Sec. 5.1). Specifically, if we directly use the candidate pool got from ConceptNet and calculate their ranking score, the top 5 results are animal, four-legged animal, mammal, wolf, fox. We can find wolf are the first concrete class we can obtain, which achieved our open-world zero-shot classification.

To show the quantitative results of our method, we realize that, although the first 3 results (animal, four-legged animal, mammal) in the previous wolf example is somewhat correct, we want to get a more concrete answer which is a wolf. Thus, we design a fixed candidate which excludes these broad words and only contains unseen classes and some disturbances. In this way, we can get a quantitative accuracy without the influence of those broad words like animal, fruit, bird.

Here, we provide the experiment setting of our seen class list and our candidate pool. Our seen classes list totally contain 36 classes, they are horse, tiger, panda, penguin, piano keys, cheetah, hyena, dog, lemon, koala, fur, squirrel, fox, rabbit, goose, sea lion, elephant, otter, duck, cock, chimpanzee, goat, orange, ball, bull, tomato, cherry, pear, turkey, seal, porpoise, alpaca, pigeon, lion, donkey, bear. Our candidate pool totally contains 20 classes which are fowl, zebra, bear, wolf, husky, swan, giraffe, jackal, peach, sheep, seal, apple, banana, train, bag, balloon, car, pen, table, eagle. The results are shown in the main paper Table 5.

C.2 Detail of Prototypical network

As described in the main paper Sec.5.1, we conduct one-shot learning using the prototypical network [49]. Here, we provide more details about the experiment. Prototypical network use the same training and test set with HVE. It is not clear for a prototypical networks to direct conduct zero-shot in the released code, so we provide a easier mode, one-shot for each test class, and use its one-shot setting. To train the prototypical network, we set the input image size as $3 \times 224 \times 224$, hidden dimension as 64, learning rate as $1e - 3$. We use their official code and follow the same training strategy as [49]. During testing, we use 5-way 1-shot setting, table.5 in the main paper provide the final result of the prototypical network.

C.3 The GAN model in Cross Feature Imagination

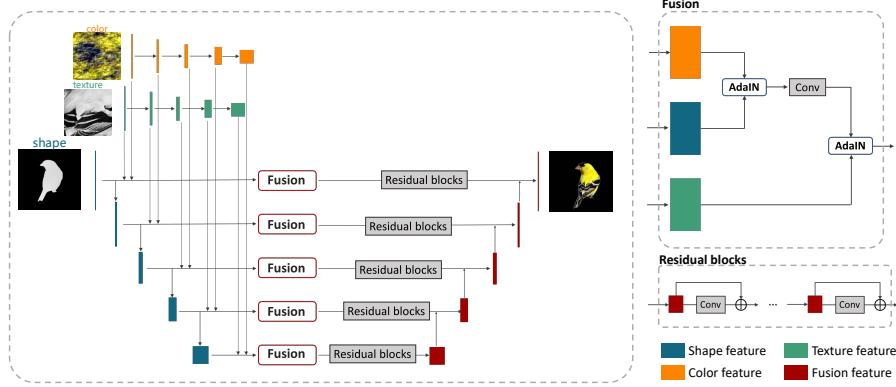


Fig. 17. The architecture of the GAN generator in cross feature imagination.

As introduced in main paper Sec. 5.2, we design a cross-feature pixel2pixel GAN model to generate the final image.

Fig. 17 shows the architecture of the generator. In order to ensure the quality of image generation, we set $k = 1$ when extracting the texture feature (the selection of k is introduced in Sec. A.2). Each feature, shape, texture, and color are encoded by a sequence of convolutions maps to $F_{s,i}, F_{t,i}, F_{c,i}, i = 1, 2, \dots, K$. These features are fused into F_i by K feature fusion modules. As for fusion modules, we apply AdaIN and L residual blocks to blend the different features. The outputs are de-convoluted into $H_i, i = 1, 2, \dots, K$, and then we get the final result. In our experiments, we use $K = 5$ and $L = 5$.

We use the same discriminator and loss function as [26]. In training the GAN model. We use Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ for both generator and discriminator, and set batch size to 16. The model is trained for 200 epochs on the training set, and we use the learning rate at $2e - 4$ in the first 100 epochs and $2e - 5$ from epoch 100 to 200. To compare our result, we use the original pix2pix GANs [] as our baseline model. The GAN model takes one type of feature as input and the original image as output. We trained three separate pix2pix GANs for each feature.

In Fig. 18, we show more results about the imagination on the test set when different types of features are given (denoted as sub-image a, b, c). With one feature (Line I), the retrieval model can find the other two plausible features (Line II), and then the GAN model can imagine the whole objects (Line III). Comparing the imagination results and the original images to which the input features belong (Line-IV), we can find that they have a similar input feature, while the other two features of the imagination images are decided by the retrieved features.

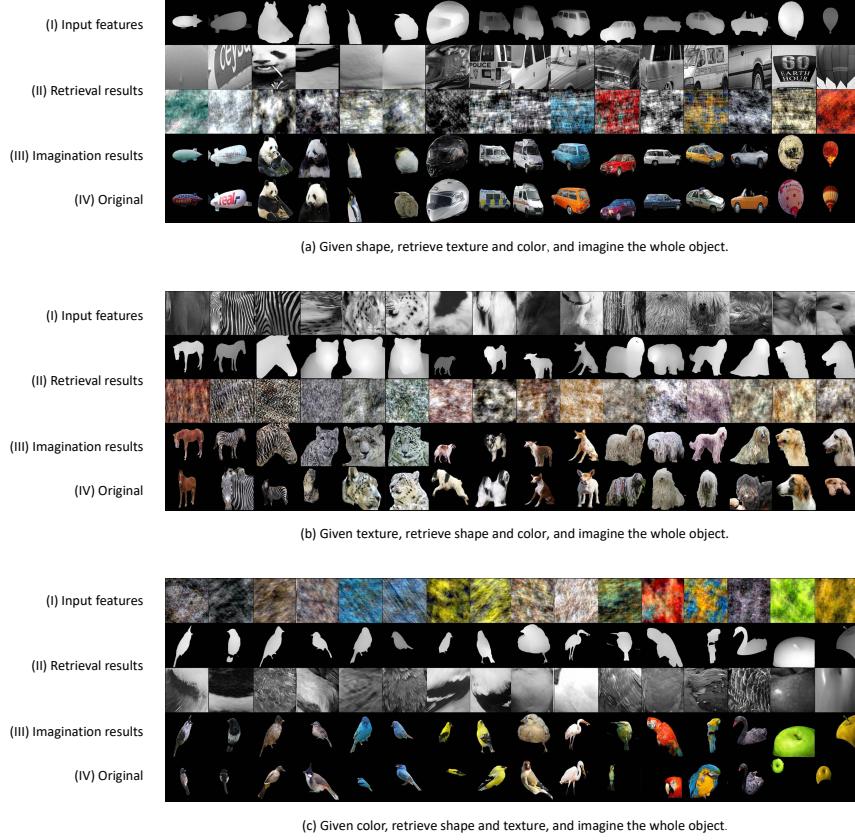


Fig. 18. Results of the imagination when different types of features are given (see sub-images a, b, c). Given one feature (Line I), the retrieval model can find the other two plausible features (Line II), and then the GAN model can imagine the whole objects (Line III). Line IV shows the original images of the input features.

D Discussion of Limitations and Future Work

As we use Grad-Cam as the foundation for selecting foreground parts, sometimes models could output the wrong saliency map (for example, the pre-trained models may regard water as the most important object for classifying a boat. As a result, our image parsing pipeline would take water as the foreground). So how to select accurate foreground parts is the first step we need to improve in the future.

After getting the foreground part of images, we use three feature extractors to mimic the process of humans perceiving this world. Our feature extractors are far away from perfect. In this work, we concatenate four square patches to get I_t , but

this introduced two lines in the I_t (the line between different squares). These lines may cause confusion to the neural networks. What's more, 4 patches may not be the best choice to represent the object's texture. Some classes are too complex to be represented by only 4 patches. For example, cars have windows, metal, rubber, paint, and plastic. Different parts of cars could have different textures. 4 square patches can represent the texture of cars to some degree, but if we want to get the representation of all texture information without any omission, we need to improve the work process of the texture extractors. However, the main goal of our model design is to summarize the contributions by end-to-end learning, with minimal human-introduced bias and assumptions in the architecture design. We mainly provide the first fully objective, data-driven, and indeed first-order, measure of the respective contributions.

In the open-world zero-shot learning experiment, we use a prototype dataset that contains limited classes, we will explore this direction with a larger dataset.

We think our HVE take a small but important step towards, from a humanoid perspective, better understanding the contributions of shape, texture, color to classification, zero-shot learning, imagination, and beyond.

Dataset copyright. We used publically available data. ImageNet, CUB, iLab-20M.

Imagenet license: Researcher shall use the Database only for non-commercial research and educational purposes We use a subset of the ILSVRC2012 dataset (ImageNet) and iLab-20M. The object in iLab-20M is toy vehicles, under Creative Commons CC-BY license. They do not contain any personally identifiable information offensive content.