DECONTEXTUALIZED LEARNING FOR INTERPRETABLE HIERARCHICAL REPRESENTATIONS OF VISUAL PATTERNS

R. Ian Etheredge[1, 2, 3, *], Manfred Schartl[4, 5, 6, 7], and Alex Jordan[1, 2, 3]

[1]Department of Collective Behaviour, Max Planck Institute of Animal Behavior, Konstanz, Germany

[2]Centre for the Advanced Study of Collective Behaviour, University of Konstanz, Konstanz, Germany

[3]Department of Biology, University of Konstanz, Konstanz, Germany

[4]Centro de Investigaciones Científicas de las Huastecas Aguazarca, A.C., Calnali, Hidalgo, Mexico

[5]Developmental Biochemistry, Biocenter, University of Würzburg, Würzburg, Bavaria, Germany

[6]Hagler Institute for Advanced Study, Texas A&M University, College Station, TX, USA

[7]Xiphophorus Genetic Stock Center, Texas State University San Marcos, San Marcos, TX, USA

August 25, 2020

SUMMARY

Apart from discriminative models for classification and object detection tasks, the application of deep convolutional neural networks to basic research utilizing natural imaging data has been somewhat limited; particularly in cases where a set of interpretable features for downstream analysis is needed, a key requirement for many scientific investigations. We present an algorithm and training paradigm designed specifically to address this: decontextualized hierarchical representation learning (DHRL). By combining a generative model chaining procedure with a ladder network architecture and latent space regularization for inference, DHRL address the limitations of small datasets and encourages a disentangled set of hierarchically organized features. In addition to providing a tractable path for analyzing complex hierarchal patterns using variation inference, this approach is generative and can be directly combined with empirical and theoretical approaches. To highlight the extensibility and usefulness of DHRL, we demonstrate this method in application to a question from evolutionary biology.

Keywords Generative Modeling · Interpretable AI · Disentangled Representation Learning · Hierarchical Features · Image Analysis · Small Data

*Corresponding author: ietheredge@ab.mpg.de

1 Introduction

The application of deep convolutional neural networks (CNNs[1]) to supervised tasks is quickly becoming ubiquitous, even outside of standardized visual classification tasks.[2] In the life sciences, researchers are leveraging these powerful models for a broad range of domain-specific discriminative tasks such as automated tracking of animal movement,[3–6] the detection and classification of cell lines,[7–9] and mining genomics data.[10]

A key motivation for the expanded use of deep feed-forward networks lies in their capacity to capture increasingly abstract and robust representations. However, outside of the objective function they have been optimized on, building interpretability into these representations is often difficult as networks naturally absorb all correlations found in the sample data and the features which are useful for defining class boundaries can become highly complex (Figure S1). For many investigations the main objective falls outside of a clearly defined detection or classification task, e.g. identifying a set of descriptive features for downstream analysis, and interpretability and generalizability is much more important. Because of this, in contrast to many traditional computer vision algorithms,[11–14] the application of more expressive approaches built on CNNs and other deep networks to research has been limited[15] (Figure 2).

Unsupervised learning, a family of algorithms designed to uncover unknown patterns in data without the use of labeled samples, offers an alternative for compression, clustering, and feature extraction using deep networks. Generative modeling techniques have been especially effective in capturing the complexity of natural images, i.e. generative adversarial networks (GANs[16]) and variational autoencoders (VAEs,[17,18]). VAEs in particular offer an intuitive way for analyzing data. As an extension of variational inference, VAEs combine an inference model, which performs amortized inference (typically a CNN) to approximate the true posterior distribution and encode samples into a set of latent variables ($q_\phi(z|x)$), and a generative model which generates new samples from those latent variables ($p_\theta(x|z)$). Instead of optimizing on a discriminative task, the objective function in VAEs is less strictly defined but typically seeks to minimize the reconstruction error between inputs x and outputs $p_\theta(q_\phi(x))$ (reconstruction loss) as well as the divergence between the distribution of latent variables $q_\phi(z|x)$ and the prior distribution $p(z)$ (latent regularization).

1.1 Overcoming Hurdles to Application

In VAEs, two problems often arise which are of primary concern to researchers using natural imaging data. *1)* The mutual information between x and z can become vanishingly small, resulting in an uninformative latent code and overfit to sample data, the information preference problem;[22,24] this is particularly true when using powerful convolutional decoders which are needed to create realistic model output.[20,23,24] *2)* In contrast to the hierarchical representations produced by deep feed-forward networks used for discriminative tasks, in generative models local feature contexts become emphasized at the cost of large-scale spatial relationships. This is a product of the restrictive mean-field assumption of pixel-wise comparisons and produces generative models capable of reproducing complex image features while using only local feature contexts without capturing higher-order spatial relationships within the latent encoding.[22]

Table 1: Desired characteristics of an integrative tool for investigations of natural image data and general representation learning meta-prior enforcement strategies.

Desired Characteristic	Representation Learning Meta-Prior[19]	Example Approach
Disentangling factors of variation	Limited number of shared factors of variation	Latent regularization[20,21]
Capturing spatial relationships	Hierarchical organization of representation	Hierarchical model architecture[22]
Incorporating existing knowledge	Local variation on manifolds	Structured latent codes[23]
Connect analyses and experiments	Local variation on manifolds	Generative models[16,17]
Inference	Probability mass and local variation on manifolds	Variational inference[17]

The basis of a more expressive and robust approach for investigating natural image data has some key requirements: *1)* provide a useful representation which disentangles factors of variation along a set of interpretable axes; *2)* capture feature contexts and hierarchical relationships; *3)* incorporate existing knowledge of feature importance and relationships between samples; *4)* allow for statistical inference of complex traits; and *5)* provide direct connections between analytical, virtual and experimental approaches. Here we integrate meta-prior enforcement strategies taken from represnetation learning[19] to specifically address the requirements of researchers using natural image data (Table 1).

Here we propose to address the limitations of existing approaches and incorporate the specific requirements of researchers using a combination of meta-prior enforcement strategies. VAEs with a ladder network architecture has been show to better capture a hierarchy of feature by mitigating the explain away problem of lower level feature, allowing for bottom-up and top-down feedback.[22] Additionally, combining pixel-wise error with a perceptual loss function[25] adapted from neural style transfer,[26,27] may also reduce the restrictive assumptions of amortized inference and pixel-wise reconstruction error by balancing them against abstract measures of visual similarity.

In terms of the latent regularization, a disentangled representation of causal factors requires an information-preserving latent code. Choosing a regularization techniques which mitigate the trade off between inference and data fit[21] can encourage the disentanglement of generative factors along a set of variables in an interpretable way. We also propose a novel training paradigm inspired by GAN chaining that further relaxes the natural covariances in the data: *decontextualized learning* and actually uses the restrictive assumptions of GAN generator networks to our advantage to overcome the limitations of small datasets, typical for many studies in the natural sciences and further increase the disentanglement of generative factors (Figure 1, Methods 4.2).

While several metrics have been proposed for assessing interpretability and disentanglement,[28–30] these metrics rely heavily on the associated labels, well defined features or stipulations from classification of detection competitions, e.g.[31] In addition to being highly domain specific, for most practical investigations in the natural sciences, these types of

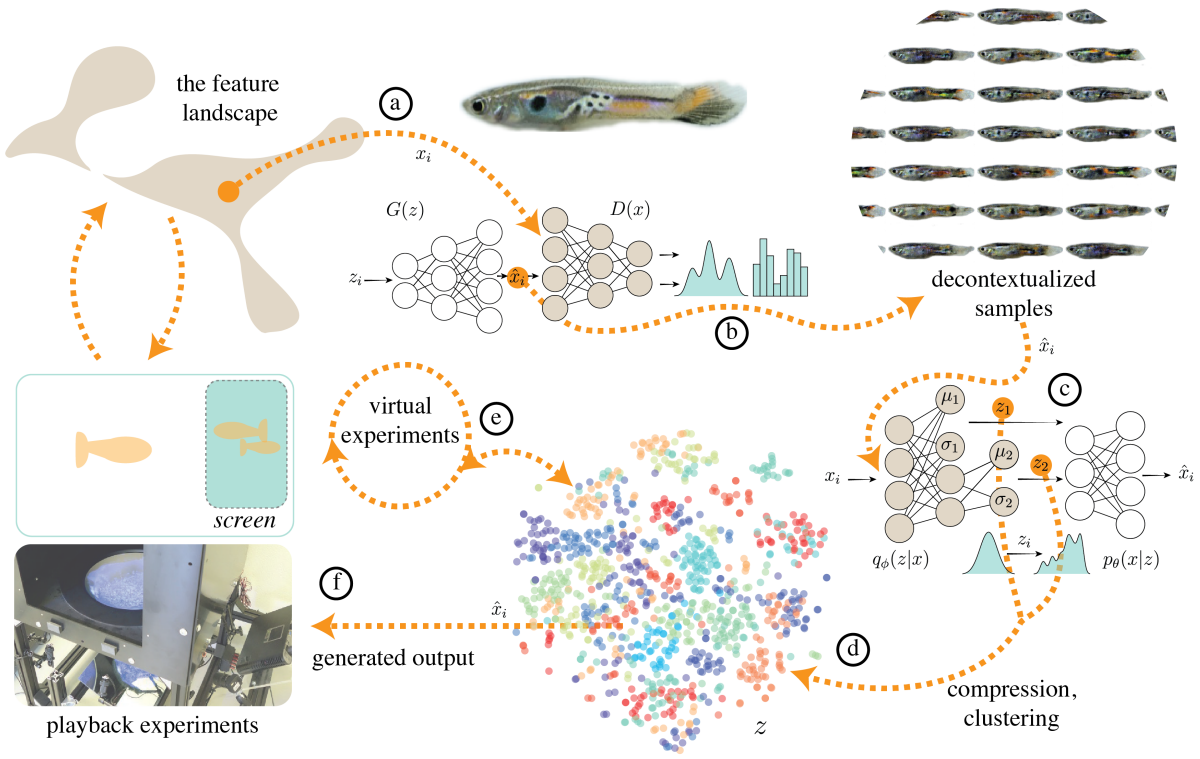

Figure 1: *Overview*. a) Many patterns (e.g. male guppy ornaments) consist of combinations of several elements which have hierarchical relationships, spatial dependence, and feature contexts which may hold distinct biological importance. In our proposed framework, small sample sizes are supplemented using a generative (GAN) model which learns images statistics sufficient to produce novel out of sample examples (b). This model can be used to produce an unlimited number of novel samples and also reduces the covariances within sample data, which is advantageous for disentangling generative features. We use these "decontextualizd" generated outputs as input (c) to a variational auto encoder (VAE). Via a specific combination of meta-prior enforcement strategies and network architectures, we capture the hierarchical structure, disentangling factors of variation across multiple scales of increasing abstraction (z_1 through z_n). Using the learned distribution over these variables, the latent representation, parameterized by a mean and variance term, we (d) define a color-pattern space. Using this low dimensional representation we can (e) interface with downstream models such as evolutionary algorithms and (f) produce photo-realistic outputs to be used in playback experiments and immersive VR. Interpolations through the color-pattern space with animated models and VR allows researchers to manipulate generated output for experimental tests. Techniques represented with dashed lines: (d) capturing a hierarchy of visual features, (e) combining a low-dimensional latent representation with virtual experiments and (f) playback experiments represent the current gaps in our analytical and experimental framework. Our approach directly address these shortcomings to span these gaps, creating a robust, integrated framework for investigating natural visual stimuli.

labels do not exist and we must often rely on fundamentally qualitative assessments. In many cases, labeled data is not available and interpreting traversals of the latent code (Figure S2) may introduce our own perceptual biases. Here, we adapt an approach from explainable AI: integrated gradients[32] in application to latent variable exploration too provide a direct assessment of latent variables, quantifying latent feature attributions without the necessity of labeled data and allows for exploring latent variables without adding additional human biases (Methods 4.3).

We demonstrate the proposed framework using two example datasets: male guppy ornamentation and butterfly wing patterns from the discipline of sensory ecology and evolution (see Appendix A for motivation and background on existing approaches).

2 Results

While biological datasets are typically small, they are usually highly structured and standardized compared to large classification datasets (e.g. ImageNet[33]). This provides an advantage for controlling noise and uninformative covariates in the data. Using a modified infoGAN[23] architecture, we incorporate prior knowledge about the structure of our sample data to generate realistic samples from the complex image distribution conditioned on a set of latent variables. Here, we incorporate prior knowledge about our samples of male guppy ornamentation images by providing a 32-class categorical latent code (Figure 3b, top right). These 32-classes represent the 32 individual tanks, unique subsets of the overall sample, with shared traits related to guppy ornamentation patterns. The categories learned by the trained model posses unique features which also covary in the sample data, e.g. a distinct black bar and orange stripe which characterizes one guppy species, *P. wengei* (Figure S2, a). While generated samples share characteristics and even resemble known varieties, generated samples posses decontextualized combinations of features across examples (Figure S2, a). We use these, decontextualized samples as input to our variational (VAE) model for our "decontextualized" training paradigm.

GAN training and VAE training are performed in separate steps so that models are not jointly optimized. The generated samples from the trained GAN model are used as training data to a variational model (Figure 1) with a hierarchical model architecture[22] which consists of 10 latent variables across four codes $(z_1, \ldots, z_4)$ with increasing expressivity, (Methods 4.2.2). We observe distinct clusters in the latent space of the trained model which correspond to sample categories and differs qualitatively from two existing method (raw pixel and perceptuall loss embeddings using tSNE,[34,35] Figure 2). The unique latent space of the four latent encodings capture unique factors of variation in the sample data in a scale-dependent way (Figure S2, Figure S3). In this model, z_1, the latent code with the lowest capacity captures local traits such as the color and intensity of discrete patches, e.g. z_{1_1} encodes variation in the intensity of an orange spot (S2 4b, left). At higher levels $(z_2, \ldots, z_4)$, latent variables encode complex traits which combine multiple elements, (S2 4b, right). We use this same latent representation to describe the relationship between samples and calculate likelihood estimates. Samples with rare traits, e.g. such as the “Tr5” strain in our sample data which are distinctly melanated, cluster together in the embedded space, and have a low sample likelihood (3).

Embedding the 4, 10-dimensional, latent codes reveals scale-dependent relationships between elements. In z_1 (Figure S3, left) color values and local features dominate the relationship between points (Figure S3, left). Nearest neighbor samples (Minkowski distance[36] in the 10-dimensional space, Figure S3, b) show color similarity whereas higher order features, e.g. patterning and morphology, determine the relationships between samples in the more expressive latent spaces $(z_2, \ldots, z_4)$. Though we find strong covariance between features across scales, in some cases the nearest neighbors samples differ greatly depending on the scale and feature context (Figure S3, b).

We assess the level of disentanglement of our trained variational models using the metric established in[30] using known class labels as attribute classes (butterfly species, learned class from infoGAN pre-training, and guppy strain varieties). Across models, we find the most expressive latent codes (z_4) provide the highest degree of disentanglement between known classes with the highest disentanglement score overall using our decontextualized, DHRL method (see Table 2).

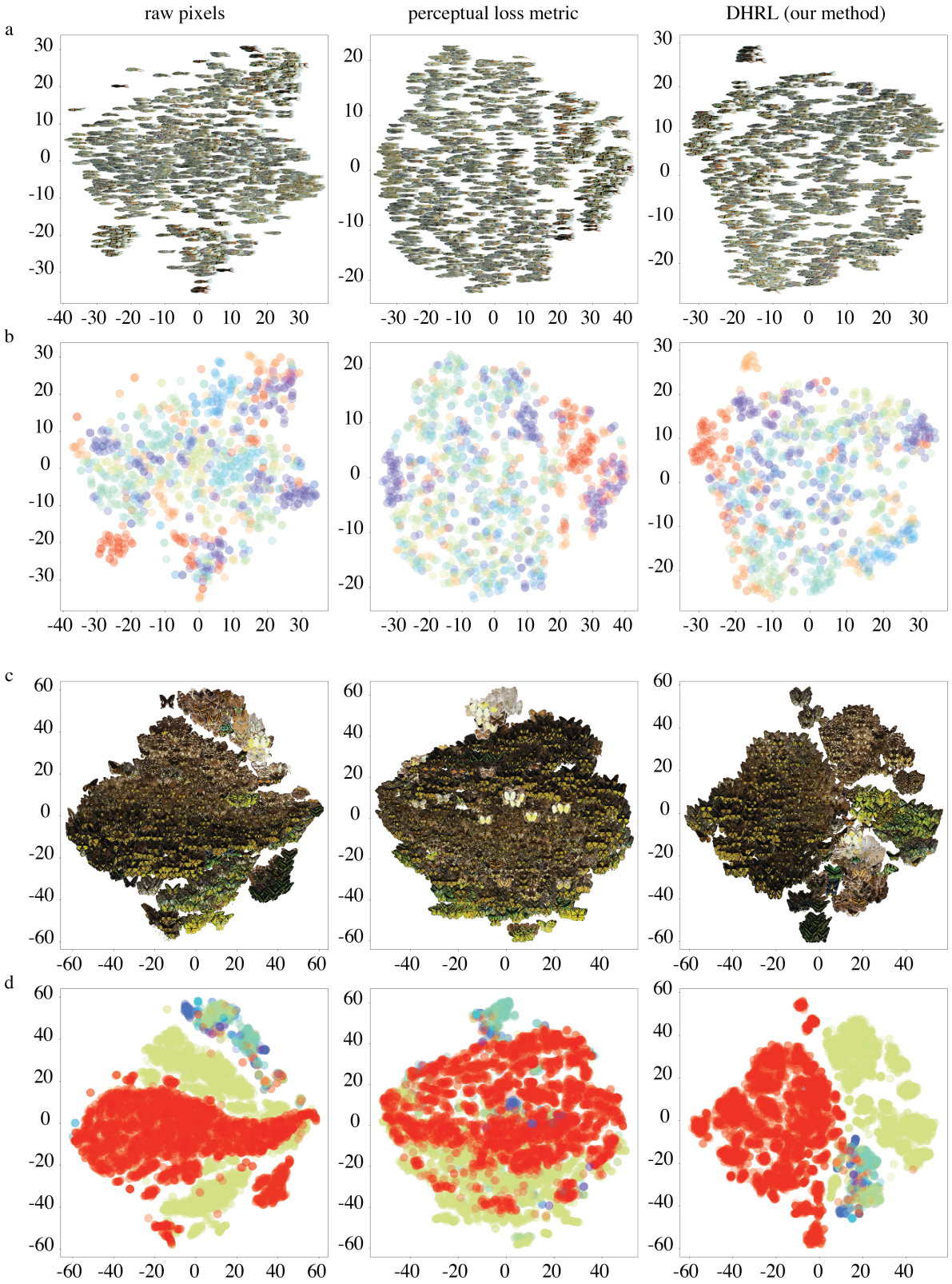

Figure 2: *Comparing with existing techniques*. 2-dimensional embedding of (left) raw pixel distributions, (middle) using a perceptual similarity score ([15,25]), and (right) our framework. a) guppies, b) butterflies. Colors indicate unique sub groups for each sample (guppy variety and butterfly species).

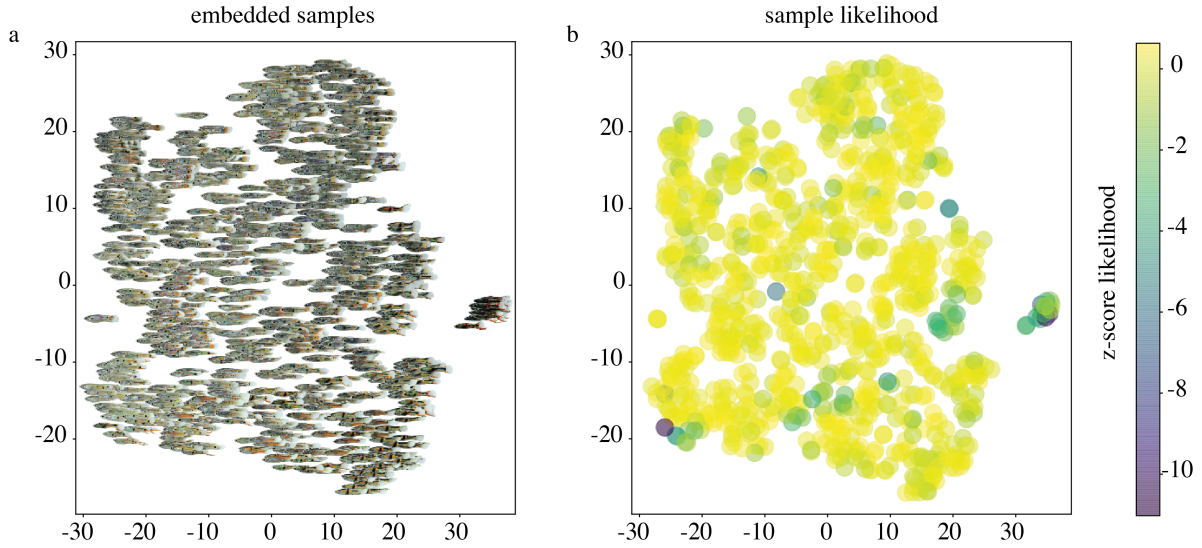

Figure 3: *Sample likelihood estimates*. a) Embedded samples b) Normalized (standard score) likelihood estimates for each sample.

Table 2: Disentanglement and completeness metrics for of VLAE inference network across datasets and when using our decontextualized learning approach (DHRL).

VLAE Training Data	$D(z_1), C(z_1)$	$D(z_2), C(z_2)$	$D(z_3), C(z_3)$	$D(z_4), C(z_4)$
Butterflies (n=9531)	0.64, 0.60	0.67, 0.55	0.63, 0.51	0.88, 0.60
Guppies (n=987)	0.29, 0.32	0.12, 0.13	0.13, 0.16	0.56, 0.66
Guppies (gen., n=19k)	0.12, 0.16	0.13, 0.18	0.32, 0.42	0.62, 0.75
Guppies **DHRL**	0.14, 0.16	0.18, 0.23	0.31, 0.39	**0.90, 0.95**

We also provide a qualitative approach for attributing latent variables to image features using network gradients (Methods 4.3); when labels are unknown. In Figure 4, a-d we visualize one variable of z_1, the least expressive latent variable space (z_{13}) of the DHRL-trained guppy latent variable model. We find that the same latent variable controls the relative intensity of green color patches across individuals. Looking at a single variable of more expressive latent codes z_{27} of the trained butterfly model (4, e-h) we find that this latent variable controls the size of yellow patches on the lower wings relative to the size of yellow patches on the upper wings (when patches are not present this variable has no effect (4, f). Further investigation of latent variables can be performed using the provided tool (`https://github.com/ietheredge/VisionEngine/notebooks/IntegratedGradients.ipynb`).

Using the latent representation, z, of our DHRL trained variational model of guppy ornaments as input, we apply an evolutionary algorithm (Figure 5), defined by a fitness function from the guppy literature: oranger, higher contrast males are preferred by females.[37] Starting from a parent population initialized by our sample embedding (900 samples), we simulate 500 generations under these selective forces. We observe exaggerated and more numerous orange and black

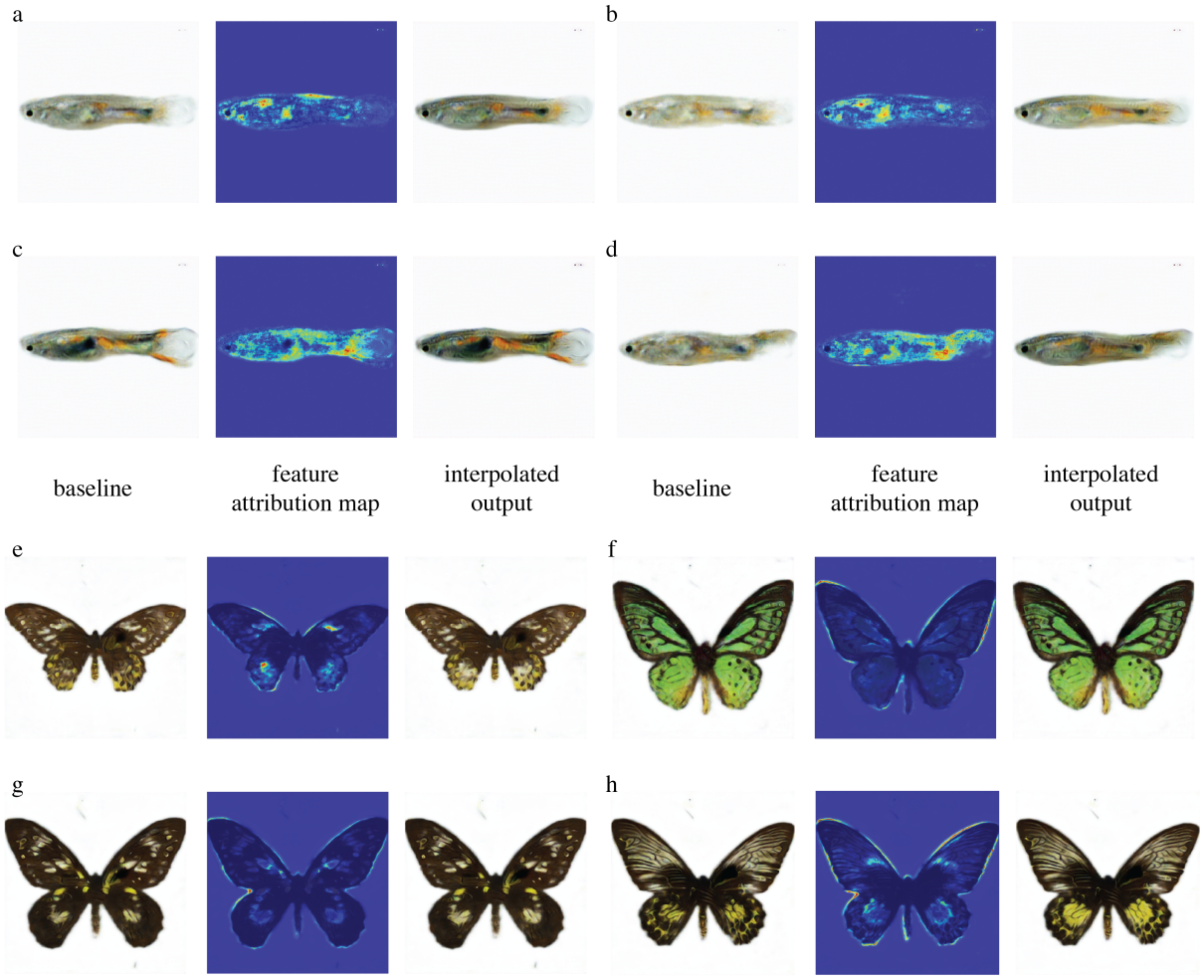

Figure 4: *Latent variable feature attribution*. a-d) Four samples of genereated guppy images performing integrated gradeints feature attribution (see Methods 4.3). a-d) guppy images visualizing the feature attributions of the latent variable $z_1 3$ of the DHRL trained variational model. e-h) butterfly images using latent feature $z_3 7$. Heatmap values have been normalized using a standard score. Images to the left are generated with the latent feature set to its lowest value in the sample, to the right with the highest value in the sample.

patches in novel configurations compared to the initial population (Figure 5, b). Projecting the latent representation of generations 1, 250, and 500, we find that instead of a single peak, after several generations, many novel solutions are optimized (Figure 5 a). Investigating the values of the latent variables over generations reveals two distinct latent factors driven to fixation in the population under these selective forces (S4). We also observe to population optimization of latent factors over time in Movie S5. Using a single Titan Xp GPU with 12GB memory we could simulate a population size of 1000 individuals in an average of 19.5 seconds per generation.

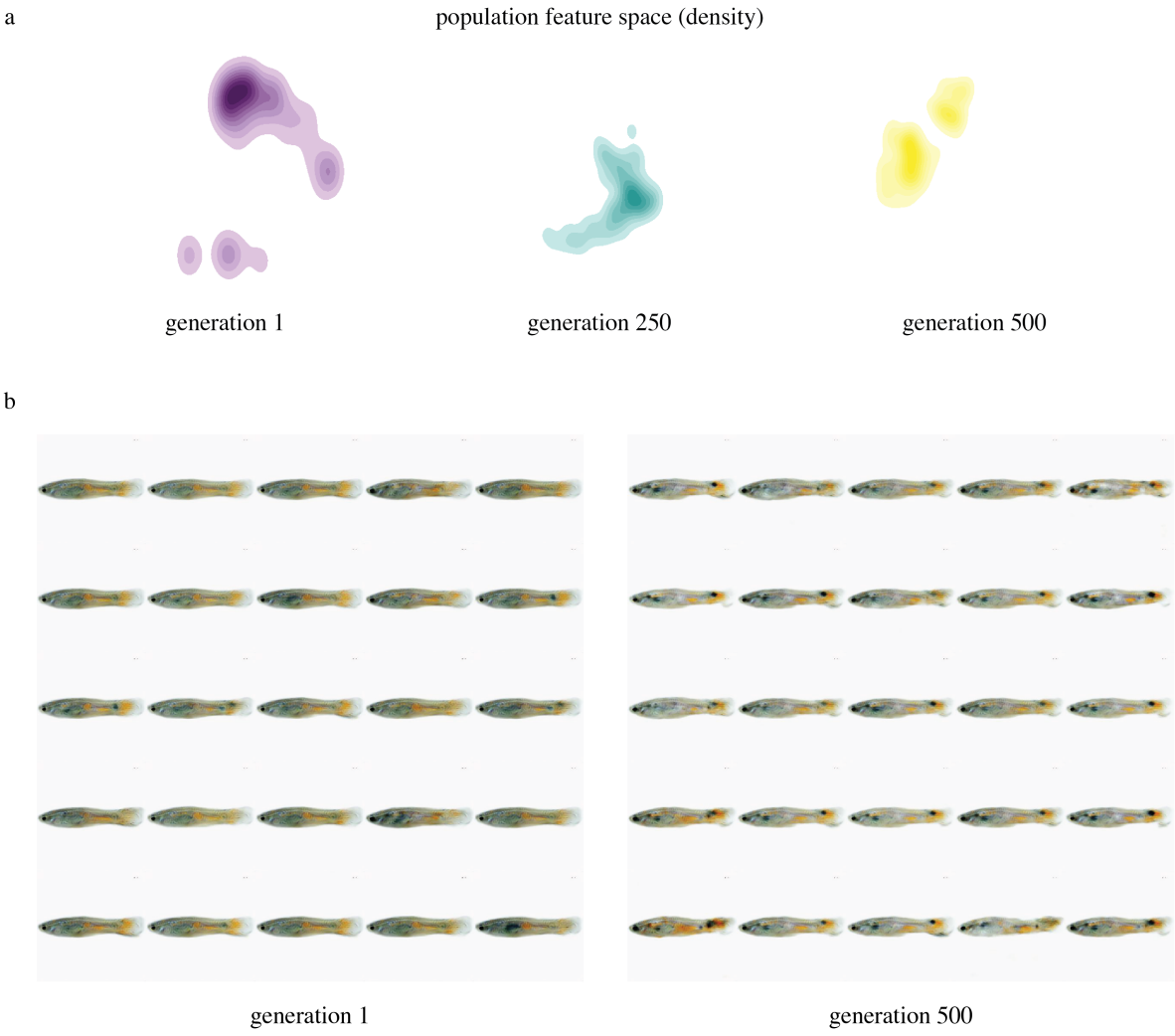

Figure 5: *Virtual Experiments*. a) Kernel density plot of samples over generations 1, 250, 500 selecting orange ornaments and contrast. After 500 generations the population has shifted from the initial sample distribution, finding two peaks which maximize the fitness function. b) Samples of initial parent population, left, with the highest fitness, compared to those with the highest fitness after 500 generations. Samples in later generations show higher numbers of brighter orange and dark melanated patches and increased within body contrast.

3 Discussion

Supervised discriminative learning algorithms are already becoming an integral tool for researchers across disciplines whereas unsupervised generative modeling approaches remain a relatively young and active area of machine learning research. Already, the highly expressive generative models like the ones presented here are transforming the way we interact with image data. By solving problems in a more general way, generative modeling approaches provide more direct connections to hypothesis testing and connecting observations. Here, we demonstrate how these approaches may serve as an engine for more integrative studies of animal coloration patterns, and natural image data more generally, directly connecting approaches.

Analytically, our approach captures important hierarchical features across spatial scales that existing approaches do not account for (Figure 2, Figure S3, Appendix A.1), it removes the inherent biases of predefined filters by learning features directly from the sample data, and it disentangles complex factors of variation into a useful, meaningful representation (Figure S2, Figure 4). More than compressing data into a low dimensional space, this approach is generative and can create novel out-of-sample examples with high fidelity. This is a potentially transformative extension for researchers in the natural sciences which is not offered by existing approaches, allowing researchers to test analytical results with virtual experiments, and empirically, by using virtual reality playback experiments or observational studies (see Movie S6).

These techniques can be adapted to many domain specific questions (see A.1 for a specific discussion regarding the potential impact of this approach on the study of color pattern evolution). As the latency between input and output decreases in video playback experiments, integrating instantaneous behavioral feedback and in-the-loop methods for hypothesis testing may be used to design complex real-time assays. More sophisticated virtual experiments may also incorporate agent based models and evolutionary algorithms working directly on the latent representation to create complex simulations (e.g as in,[38] Figure 5). In our demonstration, we are able to simulate 1000 individuals in under 20 seconds per generation with very little optimization and asynchronous approaches may already be possible. Analytically, as research in machine learning aimed at understanding how information is organized and used by algorithms advances, a growing theoretical framework with a basis in statistical mechanics[39] and information theory[40] may provide additional avenues for investigating the statistical properties of color pattern spaces and their evolution.

4 Experimental Procedures

4.1 Materials Availability

Guppy images were collected from a maintained stock at the University of Wuerzburg under authorization 568/300-1870/13 of the Veterinary Office of the District Government of Lower Franconia, Germany, in accordance with the German Animal Protection Law (TierSchG). Individuals were imaged on a white background with fixed lighting

conditions[41] using a Cannon D600 digital camera. Images were down sampled and center cropped to final size of 256 x 256 pixels. The dataset consists of 977 standardized RGB images across three species and 13 individual strains.

Butterfly images were downloaded from the Natural History Museum, London under a creative commons license (DOIs: https://doi.org/10.5519/qd.gvq3p7xq, https://doi.org/10.5519/qd.pw8srv43). This dataset consists of 9531 RGB images.

For each dataset, we segmented samples from the background using a customized object segmentation network adapted from.[42] For each dataset we annotated 8 samples to train the segmentation network. All samples were cropped and resized to 256 x 256 and placed on a transparent background (RGBA). For calculating the perceptual loss during training, images were translated to 3-channel images with a white background using alpha blending.

Updated links to original data repositories can be accessed here: `https://github.com/ietheredge/VisionEngine/README.md`.

4.1.1 Data and Code Availability

All models were implemented using Tensorflow 2.2 and can be accessed here: `https://github.com/ietheredge/VisionEngine`, including installation and evaluation scripts to reproduce our results. Instructions for creating new data loaders for training new datasets using this method can be found at `https://github.com/ietheredge/VisionEngine/data_loaders/datasets/README.md`.

4.2 Key Methods

DHRL relies on a three-step process of sequential training where first a generative adversarial network is trained to transform a noise sample into realistic out of sample examples. Next, a variational autoencoder is pre-trained on the generated samples. Then finally, the pretrained variational model is fine-tuned on the original samples.

4.2.1 InfoGAN

We use an unsupervised approach to disentangle discrete and continous latent factors adapted from[23] (InfoGAN) which modifies the minimax game typically used for training GANs such that:

$$\min_{G,Q} \max_{D} V_I(D, G, Q) = V(D, G) - \lambda L_I(G, Q) \tag{1}$$

where $V(D, G)$ is the original GAN objective introduced in[16] and $L_I(G, Q)$ approximates the lower bound of the mutual information $I(c; G(z, c))$ using Monte Carlo sampling such that $L_I(G, Q) \leq I(c; G(z, c))$.[23] Like the generator G and discriminator D, Q is parameterized as a neural network and shares all convolutional layers with D.

Both discrete $Q(c_d|x)$ and continuous latent codes $Q(c_c|x)$ are provided with continuous latent codes treated as a factored Gaussian distributions. Importantly, InfoGAN does not require supervision and no labels are provided, e.g.[29]

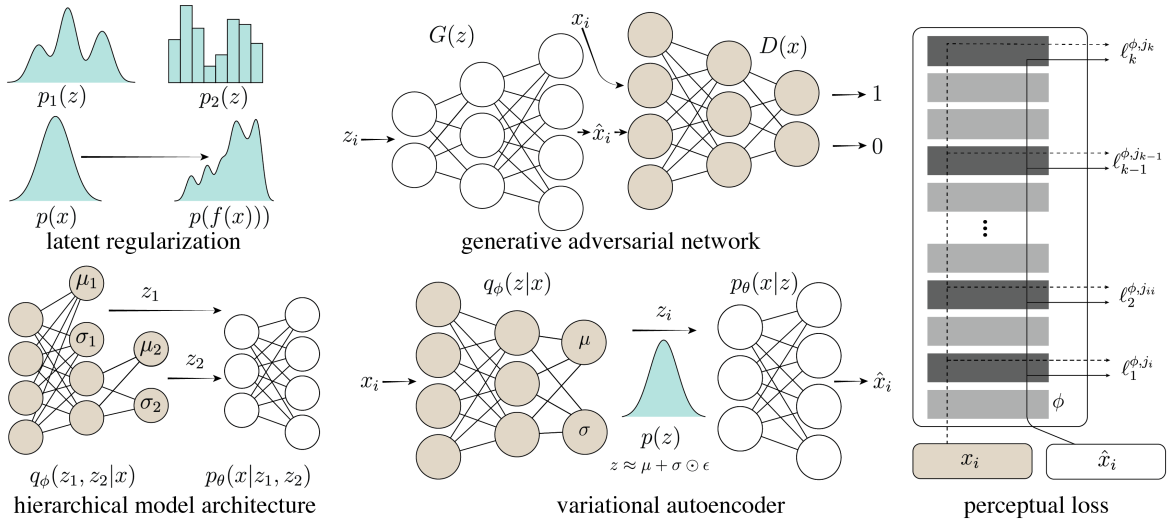

Figure 6: *Key Methods Top left*: the distributions of our latent representation may be parameterized by a number of continuous or discrete variables. In infoGAN, a categorical latent code is combined with continuous latent codes which allows for disentangling substructures in the sample data without labeled samples. *Top right*, example structure of a generative adversarial network. Here, a noise vector, z_i is input to the generator network $G(z)$ which produces a reconstructed output $\hat{x_i}$. A real sample, x_i, and generated sample $\hat{x_i}$ are subsequently passed through a separate discriminator network $D(x)$ which determines if the sample is real (1) or generated (0). In infoGAN, the latent encoding of generated samples is optimized an additional network Q which shares all convolutional layers with D. *Bottom left*, the generic architecture of a variational ladder autoencoder (VLAE). Multiple latent spaces $(z_1, z_2, ..., z_k)$ are learned with each successive input layer having increasing expressivity and abstraction (abilitiy to combine features across spatial scales). *Bottom middle*, structure of a variational autoencoder (VAE). x_i and $\hat{x_i}$ are an example input and its reconstructed output, the probabilistic encoder or inference model, $q_\phi(z|x)$, performs posterior inference learning shared model parameters, ϕ, across samples, approximating the true posterior distribution. The probabilistic decode, $p_\theta(Z|X)$, $p_\theta(X|Z)$, learns a joint distribution of the encoded space, Z, and the data space x. The low dimensional bottleneck, Z is a distribution of latent variables capable of reconstructing sample inputs, parameterized by a vector of means μ and standard deviations σ. The noise term ϵ allows for the parameters of this multivariate distribution to be optimized using back propagation, known as the reparametrization trick. *Right*, Perceptual loss models use a pretrained network, ϕ, e.g. VGG-16.[43] Two samples, the original input and reconstructed output, are input to the model and the maxpooling layer activations for each are used as outputs. The distance between these functions emphasizes higher-level similarity than standard pixel-wise differences. Outputs from the shallow layer ℓ_1 represents low-level local features where as output from the deeper layer ℓ_k contains information from across spatial scales and more abstract representations. The euclidean distance between these activation outputs gives a metric for the similarity of the two inputs as "perceived" by a network pre-trained on a much broader dataset. Perceptual loss functions can be used as a stand-alone transfer-learning approach to finding perceptual differences between samples or as part of any network as an additional or alternative reconstruction loss (see 2).

We substitute the original generator and discriminator models from[23] with the architecture described in[44] and increase the flexibility of the latent code, providing additional continuous and discrete latent codes. For guppy experiments, we provide two continuous and 19 discrete codes (samples were drawn from 19 paternal lines). For the basis noise vector input to the generator, we used 100-unit random noise vector.

4.2.2 Variational Ladder Autoencoder

In contrast to hierarchical architectures, e.g.,[45,46] we learn a hierarchy of features by using multiple latent codes with increasing levels of abstraction from,[22] i.e. $q_\phi(z_1, \ldots, z_L|x)$. The expressivity of z_i is determined by its depth. The encoder $q_\phi(z_1, \ldots, z_L|x)$ consists of four blocks such that:

$$H_\ell = G_\ell(H_{\ell-1}) \tag{2}$$

$$z_\ell \sim \mathcal{N}(\mu_\ell(H_\ell), \mathbf{I}) \tag{3}$$

where H_ℓ, G_ℓ, and μ_ℓ are neural networks. For our encoder model, G_ℓ is a stack of convolutional, batch normalization, and leaky rectified linear unit activation (Conv-BN-LeakyReLU), we stack four Conv-BN-LeakyReLU blocks for each G_ℓ with increasing number of channels for each subsequent convolutional layer, i.e. N-channels/2, N-channels, N-channels, N-channels*2 where N-channels is 16, 64, 256, 1024 for G_1, G_2, G_3, G_4 respectively. We apply spectral normalization to all convolutional layers (see below). Because we want to preserve feature localization, we use average pooling followed by a squeeze-excite block to apply a context-aware weighting to each channel (see below).

Similarly, the decoder, $p_\theta(x|z_1, \ldots, z_L)$, is composed of blocks such that:

$$\tilde{z}_\ell = U_\ell([\tilde{z}_{\ell+1}; V_\ell(z_\ell)]) \tag{4}$$

where $[.;.]$ denotes channel-wise concatenation. Parallel to G_ℓ, blocks in the encoder: U_ℓ are composed of Conv-BN-ReLU blocks (note the use of ReLU and not LeakyReLU in the decoder) with decreasing number of channels in each convolutional layer, i.e. N-channels*2, N-channels, N-channels, N-channels/2 where N-channels is 1024, 256, 64, 16. No spectral normalization wrappers or squeeze-excite layers are applied in the decoder.

4.2.3 Squeeze-Excite Layers

Squeeze-and-Excitation Networks[47] were proposed to improve feature interdependence by adaptively weighting each channel within a feature map based on the filter relevance by applying a a channel-wise recalibaration. Here we apply squeeze-excite (SE) layers prior to the variational layer such that each embedding z_i captures features with cross-channel dependencies. Each SE layer consists of a global average pooling layer which averages channel-wise features followed by two fully connected layers with relu activations, the first with size channels/16 and the second with

the same size as the number of input channels. Finally a sigmoid, "excite," layer assigns channel wise probabilities which are then multiplied channel wise with the original inputs.

4.2.4 Reconstruction Loss

We minimize the negative log likelihood of the sample data by minimizing the mean squared error between input and output, jointly optimizing the reconstruction loss for each sample x:

$$\begin{aligned}\mathcal{L}_{\text{pixel-wise}} &= \mathbb{E}_{p_{\text{data}}}(x)\mathbb{E}_{q_\phi}(z \mid x)\left[\log p_\theta(x \mid z)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(x_i - p_\theta(q_\phi(x_i))\right)^2\end{aligned} \tag{5}$$

To relax the restrictive mean-field assumption which is implicit in minimizing the pixel-wise error, we jointly optimize the similarity between inputs and outputs using intermediate layers of a pretrained network, VGG16,[43] as feature maps.[25–27] Here we calculate the Gram matrices of feature maps, which match the feature distributions of real and generated outputs for each layer as:

$$G^\ell_{ab} = \frac{\sum_{cd} F^\ell_{cda}(x) F^\ell_{cdb}(x)}{CD} \tag{6}$$

$$\mathcal{L}_{\text{perceptual}} = \sum_{\ell=1}^{L} \frac{\frac{1}{n}\sum_{i=1}^{n}\left(G^\ell_{ab}(x_i) - G^\ell_{cd}(p_\theta(q_\phi(x_i)))\right)^2}{L} \tag{7}$$

for feature maps F_a and F_b in layer ℓ across locations c, d. This measures the correlation between image filters and is equivalent to minimizing the distance between the distribution of features across feature maps, independently of feature position.[48]

The combined reconstruction loss is a weighted sum of the perceptual loss and pixel-wise error:

$$\mathcal{L}_{\text{reconstruction}} = \alpha\mathcal{L}_{\text{perceptual}} + \beta\mathcal{L}_{\text{pixel-wise}} \tag{8}$$

where α and β are Lagrange multipliers controlling the influence of each loss term. Here we set α = 1e-6 and β = 1e5 to balance the contribution of reconstruction terms with variational loss (see below).

4.2.5 Maximum Mean Discrepancy

We use the maximum mean discrepancy approach (MMD)[21] to maximize the similarity between the statistical moments of $p(z)$ and $q_\phi(x)$ using the kernel embedding trick:

$$\mathrm{MMD}(p(z)\|q_\phi(z)) = \mathbb{E}_{p(z),p(z')}\left[k\left(z,z'\right)\right] + \mathbb{E}_{q_\phi(z),q_\phi(z')}\left[k\left(z,z'\right)\right] - 2\mathbb{E}_{p(z),q_\phi(z')}\left[k\left(z,z'\right)\right] \tag{9}$$

using a Gaussian kernel, $k\left(z,z'\right)$, such that

$$k\left(z,z'\right) = e^{-\frac{\|z-z'\|^2}{2\sigma^2}} \tag{10}$$

to measure the similarity between $p_\theta(z)$ and $q_\phi(z)$ in Euclidean space. We measured similarity using multiple kernels with varying degrees of smoothness, controlled by the value of σ^2, i.e. multi-kernel MMD,[49] with varying bandwidths: σ^2 = 1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 5, 10, 15, 20, 25, 30, 35, 100, 1e3, 1e4, 1e5, 1e6.

Weighing the influence of MMD kernel differences on the combined objective function is controlled by the Lagrange multiplier λ applied across each latent code. Giving the combined objective:

$$\mathcal{L}_{\text{total}} = \left(\sum_{i}^{L} \lambda\, \text{MK-MMD}\left(q_\phi(z_i)\|p(z_i)\right)\right) + \mathcal{L}_{\text{reconstruction}} \tag{11}$$

where L is the number of hierarchical latent codes and z_i is the n-dimensional latent code and the prior, $p(z_i) = \mathcal{N}(0, \mathbf{I})$ and $\mathcal{L}_{\text{reconstruction}}$ define above. Here, we set λ = 1.

4.2.6 Denoise Training

In addition to further relaxing the contribution of pixel-wise error, adding a denoising criterion has been shown to yield better sample likelihood by learning to map both training data and corrupted inputs to the true posterior, providing more robust training for out of sample data.[50] We implement this with the addition of noise layer which samples a corrupted input $\tilde{x}$ from input x before passing $\tilde{x}$ to the encoder $q_\phi(z|\tilde{x})$. We use apply random binomial noise (salt and pepper) to ten percent of pixels.

4.2.7 Spectral Normalization

Spectral normalization has been proposed as a method to prevent exploding gradients when using rectified linear units to stabilize GAN training via a global regularization on the weight matrix of each layer as opposed to gradient clipping to provide bounded first derivatives (the Lipschitz constraint).[51]

4.3 Latent Feature Attribution and Disentanglement

Understanding the importance of features for model predictions is an active area of research. Integrated gradients, introduced by,[32] assigns feature importance, determining causal relationships between predictions and image features by summing the gradients along paths between x' and x.

$$\text{IG}_i(x) ::= (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial \mathcal{P}\left(x' + \alpha \times (x - x')\right)}{\partial x_i} d\alpha \tag{12}$$

We adapt this procedure to investigate the contribution of each latent variable parameter z_i where we use a baseline z, an encoding of a singe sample x and iterate z_j while holding all other z_l constant and summing the gradients of the decoder $p_\theta(x|z)$ such that:

$$\text{IG}_i^{approx}(p_\theta(x|z^j)) ::= \left(p_\theta(x|z^j)_i - p_\theta(x|z^{j\prime})_i\right) \times \sum_{k=1}^{m} \frac{\partial \mathcal{P}\left(p_\theta(x|z^{j\prime}) : {z^{j\prime}}_j = {z^{j\prime}}_j + \frac{k}{m} \times (z_j^j - {z^{j\prime}}_j^j)\right)}{\partial p_\theta(x|z^j)_i} \times \frac{1}{m} \tag{13}$$

where j is the axis of latent code being interpolated, i is the individual feature (pixel), $p_\theta(x|z)$ is the reconstructed output, $p_\theta(x|z')$ is the baseline reconstructed output, k is the perturbation constant, and m is the number of steps in the approximation of the integral. We use the Riemann sum approximation of the integral over the interpolated path $\mathcal{P}$ which involves computing the gradient in a loop over the inputs for $k = 1, \ldots, m$. Here, we use m = 300 and $k = 2\max(|z|)$ for each z^j starting from a baseline $p_\theta(x|z^{j\prime}) : z_j = -\max(|z|)$.

We use the technique developed in[30] for assessing disentanglement, measuring the relative entropy of latent factors for predicting class labels. We measure disentanglement of D_i of each latent code is measured by $D_i = (1 - H_K(P_i))$ where H_K is the entropy and P_i is the relative importance of the generative factor. We also include a metric of completeness C_i, approximating the degree to which the generative factor is captured by a single latent variable, where $C_j = (1 - H_D(P_j))$ where P_j is the unweighted contribution of generative factors.[30] Here, in the absence of labeled features, we use species (butterflies), breeding line variants (guppies), and predicted class of the generative model (generated guppies, 4.2.1, above) for each model as approximate class labels (one class). This approximation naturally overestimates D_i and underestimates C_j as there is some overlap between classes in terms of visual features (see Figure 2, Figure S3). While[30] proposes a third term to evaluate representations I to measure the relative informativeness, we found that this value was highly coupled to the choice of the Lagrange multiplier λ used for latent regularization (above).

4.4 Simulating Evolution on the Latent Space

For demonstrating an example virtual experiment, we use a genetic algorithm, with a parent population of 1000 random samples, evolved over 500 generations. Parent samples are random initialized across the the latent variables of each

latent code. Fitness was calculate as an equally weighted sum of the total percentage of pixels within two ranges (orange rgb(0.9, 0.55, 0.) > rgb(1., 0.75, 0.1) and black rgb(0., 0., 0.) < rgb(0.2, 0.2, 0.2)) measured on the generated output, a simplification of empirical results from the literature.[37,52] During each generation predicted fitness for each sample in the population was measured by the fitness of the nearest neighboring value in the reference table (for processing speed). To simulate weak selective pressure on the fitness function, we drew 500 random parent subsamples weighted by their proportional fitness. An additional 200 samples were drawn, without the proportional fitness weighting. Together, from the 700 subsamples in each generation we drew 300 random pairs, the "alleles" from each sample (the specific latent variable values) were chosen randomly with equal probability to create a combined offspring between the two samples. Each combined offspring then had two alleles randomly mutated, one by drawing from a random normal distribution and the other by replacing an existing value with zero (similar to destabilizing and stabilizing mutations). The next generation thus consisted of 100 samples, 700 parent samples + 300 offspring. This process was repeated for 500 generations.

5 Acknowledgments

We would like to thank members of the Dept. of Collective Behavior, Max Planck Institute of Animal Behavior and Centre for the Advanced Study of Collective Behaviour, University of Konstanz for comments on earlier versions of the manuscript as well as the Max Planck Computing and Data Facility for use of computational resources.

5.1 Author contributions

RIE conceived the approach and designed the methodology; MS and RIE collected sample data. RIE wrote the manuscript. AJ secured funding. All authors contributed to editing and approving the manuscript.

5.2 Declaration of Interests

The Authors have no financial or non-financial competing interest.

References

[1] Yann LeCun, Koray Kavukcuoglu, and Clément Farabet. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE international symposium on circuits and systems*, pages 253–256. IEEE, 2010.

[2] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[3] Katarzyna Bozek, Laetitia Hebert, Alexander S Mikheyev, and Greg J Stephens. Towards dense object tracking in a 2d honeybee hive. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4185–4193, 2018.

[4] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie W. Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 2018.

[5] Talmo D. Pereira, Diego E. Aldarondo, Lindsay Willmore, Mikhail Kislin, Samuel S. H Wang, Mala Murthy, and Joshua W. Shaevitz. Fast animal pose estimation using deep neural networks. *Nature Methods*, 16(1):117–125, 2019.

[6] Jacob M Graving, Daniel Chae, Hemal Naik, Liang Li, Benjamin Koger, Blair R Costelloe, and Iain D Couzin. Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning. *Elife*, 8:e47994, 2019.

[7] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al. U-net: deep learning for cell counting, detection, and morphometry. *Nature methods*, 16(1):67–70, 2019.

[8] Julian Riba, Jonas Schoendube, Stefan Zimmermann, Peter Koltay, and Roland Zengerle. Single-cell dispensing and 'real-time'cell classification using convolutional neural networks for higher efficiency in single-cell cloning. *Scientific reports*, 10(1):1–9, 2020.

[9] Claire McQuin, Allen Goodman, Vasiliy Chernyshev, Lee Kamentsky, Beth A Cimini, Kyle W Karhohs, Minh Doan, Liya Ding, Susanne M Rafelski, Derek Thirstrup, et al. Cellprofiler 3.0: Next-generation image processing for biology. *PLoS biology*, 16(7):e2005970, 2018.

[10] Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T Afshar, et al. A universal snp and small-indel variant caller using deep neural networks. *Nature biotechnology*, 36(10):983–987, 2018.

[11] Paul VC Hough. Method and means for recognizing complex patterns, December 18 1962. US Patent 3,069,654.

[12] Christopher G Harris, Mike Stephens, et al. A combined corner and edge detector. In *Alvey vision conference*, volume 15, pages 10–5244. Citeseer, 1988.

[13] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157 vol.2, Sep. 1999.

[14] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, November 2004.

[15] Briana D Ezray, Drew C Wham, Carrie E Hill, and Heather M Hines. Unsupervised machine learning reveals mimicry complexes in bumblebees occur along a perceptual continuum. *Proceedings of the Royal Society B*, 286(1910):20191501, 2019.

[16] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *CoRR*, abs/1406.2661, 2014.

[17] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[18] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China, 22–24 Jun 2014. PMLR.

[19] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, Aug 2013.

[20] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Balancing learning and inference in variational autoencoders. 33:5885–5892, 2019.

[21] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520, 2007.

[22] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Learning hierarchical features from deep generative models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4091–4099. JMLR. org, 2017.

[23] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

[24] Tian Qi Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 2610–2620, 2018.

[25] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

[26] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 262–270, Cambridge, MA, USA, 2015. MIT Press.

[27] Leon Gatys, Alexander Ecker, and Matthias Bethge. A neural algorithm of artistic style. *Journal of Vision*, 16(12):326–326, Sep 2016.

[28] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[29] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2649–2658, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[30] Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

[31] Grégoire Mesnil Yann Dauphin, Xavier Glorot, Salah Rifai, Yoshua Bengio, Ian Goodfellow, Erick Lavoie, Xavier Muller, Guillaume Desjardins, David Warde-Farley, Pascal Vincent, et al. Unsupervised and transfer learning challenge: a deep learning approach. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 97–110, 2012.

[32] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 2017.

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[34] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[35] Dmitry Kobak and Philipp Berens. The art of using t-sne for single-cell transcriptomics. *Nature communications*, 10(1):1–14, 2019.

[36] Richard Dubes and Anil K. Jain. Clustering methodologies in exploratory data analysis. In *Advances in computers*, volume 19, pages 113–228. Elsevier, 1980.

[37] Anne E Houde. Mate choice based upon naturally occurring color-pattern variation in a guppy population. *Evolution*, 41(1):1–10, 1987.

[38] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems 31*, pages 2451–2463. Curran Associates, Inc., 2018. `https://worldmodels.github.io`.

[39] Yasaman Bahri, Jonathan Kadmon, Jeffrey Pennington, Sam S Schoenholz, Jascha Sohl-Dickstein, and Surya Ganguli. Statistical mechanics of deep learning. *Annual Review of Condensed Matter Physics*, 2020.

[40] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. pages 368–377, 1999.

[41] Darrell J Kemp. Female mating biases for bright ultraviolet iridescence in the butterfly eurema hecabe (pieridae). *Behavioral Ecology*, 19(1):1–8, 2008.

[42] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 221–230, 2017.

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 2015.

[44] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 779–788. IEEE Computer Society, 2016.

[45] Philip Bachman. An architecture for deep, hierarchical generative models. In *Advances in Neural Information Processing Systems*, pages 4826–4834, 2016.

[46] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *Advances in neural information processing systems*, pages 3738–3746, 2016.

[47] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[48] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 2230–2236. ijcai.org, 2017.

[49] Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in neural information processing systems*, pages 1205–1213, 2012.

[50] Daniel Im Jiwoong Im, Sungjin Ahn, Roland Memisevic, and Yoshua Bengio. Denoising criterion for variational auto-encoding framework. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[51] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[52] John A Endler and Anne E Houde. Geographic variation in female preferences for male traits in poecilia reticulata. *Evolution*, 49(3):456–468, 1995.

[53] Henry Walter Bates. *The naturalist on the river Amazons*. London; Toronto: JM Dent; New York: EP Dutton,[1910, reprinted 1921], 1863.

[54] Charles Darwin. *The descent of man: and selection in relation to sex*. John Murray, Albemarle Street., 1871.

[55] Alfred Russel Wallace. The colors of animals and plants. *The American Naturalist*, 11(11):641–662, 1877.

[56] Fritz Müller. *Über die vortheile der mimicry bei schmetterlingen*. 1878.

[57] Edward Bagnall Poulton. *The colours of animals: their meaning and use, especially considered in the case of insects*. D. Appleton, 1890.

[58] Gerald Handerson Thayer. *Concealing-coloration in the animal kingdom: an exposition of the laws of disguise through color and pattern: being a summary of Abbott H. Thayer's discoveries*. Macmillan Company, 1918.

[59] Hugh Bamford Cott. Adaptive coloration in animals. 1940.

[60] John Maynard Smith. Natural selection and the concept of a protein space. *Nature*, 225(5232):563–564, 1970.

[61] Jakob Von Uexküll. A stroll through the worlds of animals and men: A picture book of invisible worlds. *Semiotica*, 89(4):319–391, 1992.

[62] Cedric P van den Berg, Jolyon Troscianko, John A Endler, N Justin Marshall, and Karen L Cheney. Quantitative colour pattern analysis (qcpa): A comprehensive framework for the analysis of colour patterns in nature. *Methods in Ecology and Evolution*, 11(2):316–332, 2020.

[63] Rafael Maia, Hugo Gruson, John A. Endler, and Thomas E. White. pavo 2: New tools for the spectral and spatial analysis of colour in r. *Methods in Ecology and Evolution*, 0(0), 2019.

[64] Mary Caswell Stoddard, Rebecca M. Kilner, and Christopher Town. Pattern recognition algorithm reveals how birds evolve individual egg pattern signatures. *Nature Communications*, 5:4117 EP –, Jun 2014. Article.

[65] Felipe M. Gawryszewski. Color vision models: Some simulations, a general n-dimensional model, and the colourvision r package. *Ecology and Evolution*, 8(16):8159–8170, 2018.

[66] Cynthia Tedore and Sönke Johnsen. Using RGB displays to portray color realistic imagery to animal eyes. *Current Zoology*, 63(1):27–34, 06 2016.

[67] John A. Endler. Variation in the appearance of guppy color patterns to guppies and their predators under different visual conditions. *Vision Research*, 31(3):587 – 608, 1991.

[68] John A. Endler and Paul W. Mielke JR. Comparing entire colour patterns as birds see them. *Biological Journal of the Linnean Society*, 86(4):405–431, 2005.

[69] John A. Endler. A framework for analysing colour pattern geometry: adjacent colours. *Biological Journal of the Linnean Society*, 107(2):233–253, 2012.

[70] Jolyon Troscianko and Martin Stevens. Image calibration and analysis toolbox – a free software suite for objectively measuring reflectance, colour and pattern. *Methods in Ecology and Evolution*, 6(11):1320–1331, 2015.

[71] Eleanor M. Caves and Sönke Johnsen. Acuityview: An r package for portraying the effects of visual acuity on scenes observed by an animal. *Methods in Ecology and Evolution*, 9(3):793–797, 2018.

[72] John A. Endler, Gemma L. Cole, and Alexandrea M. Kranz. Boundary strength analysis: Combining colour pattern geometry and coloured patch visual properties for use in predicting behaviour and fitness. *Methods in Ecology and Evolution*, 9(12):2334–2348, 2018.

[73] Mary Caswell Stoddard and Daniel Osorio. Animal coloration patterns: Linking spatial vision to quantitative analysis. *The American Naturalist*, 193(2):164–186, 2019.

[74] H.F. Nijhout. Elements of butterfly wing patterns. *Journal of Experimental Zoology*, 291(3):213–225, 2001.

[75] G Th Fechner. Ueber die subjectiven nachbilder und nebenbilder. *Annalen der Physik*, 126(7):427–470, 1840.

[76] Ray Fuller and Jorge A Santos. *Human factors for highway engineers*. Pergamon Amsterdam, The Netherlands, 2002.

[77] Gemma L Cole and John A Endler. Male courtship decisions are influenced by light environment and female receptivity. *Proceedings of the Royal Society B: Biological Sciences*, 283(1839):20160861, 2016.

[78] Andreas Nieder, David J Freedman, and Earl K Miller. Representation of the quantity of visual items in the primate prefrontal cortex. *Science*, 297(5587):1708–1711, 2002.

[79] S Fujita, T Kitayama, N Mizoguchi, Y Oi, N Koshikawa, and M Kobayashi. Spatiotemporal profiles of transcallosal connections in rat insular cortex revealed by in vivo optical imaging. *Neuroscience*, 206:201–211, 2012.

[80] Scott Cheng-Hsin Yang, Mate Lengyel, and Daniel M Wolpert. Active sensing in the categorization of visual patterns. *Elife*, 5:e12215, 2016.

[81] Laura A Kelley and Jennifer L Kelley. Animal visual illusion and confusion: the importance of a perceptual perspective. *Behavioral Ecology*, 25(3):450–463, 2014.

[82] Sami Merilaita, Nicholas E. Scott-Samuel, and Innes C. Cuthill. How camouflage works. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1724):20160341, 2017.

[83] Clelia Gasparini, Giovanna Serena, and Andrea Pilastro. Do unattractive friends make you look better? context-dependent male mating preferences in the guppy. *Proceedings of the Royal Society B: Biological Sciences*, 280(1756):20123072, 2013.

[84] David Marr. Vision: A computational investigation into the human representation and processing of visual information, henry holt and co. *Inc., New York, NY*, 2(4.2), 1982.

[85] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154, 1962.

[86] Diego E Pafundo, Mark A Nicholas, Ruilin Zhang, and Sandra J Kuhlman. Top-down-mediated facilitation in the visual cortex is gated by subcortical neuromodulation. *Journal of Neuroscience*, 36(10):2904–2914, 2016.

[87] Spencer J Ingley, Mohammad Rahmani Asl, Chengde Wu, Rongfeng Cui, Mahmoud Gadelhak, Wen Li, Ji Zhang, Jon Simpson, Chelsea Hash, Trisha Butkowski, et al. anyfish 2.0: an open-source software platform to generate and share animated fish models to study behavior. *SoftwareX*, 3:13–21, 2015.

[88] John R Stowers, Maximilian Hofbauer, Renaud Bastien, Johannes Griessner, Peter Higgins, Sarfarazhussain Farooqui, Ruth M Fischer, Karin Nowikovsky, Wulf Haubensak, Iain D Couzin, et al. Virtual reality for freely moving animals. *Nature methods*, 14(10):995, 2017.

[89] H. Naik, R. Bastien, N. Navab, and I. D. Couzin. Animals in virtual environments. *IEEE Transactions on Visualization and Computer Graphics*, 26(5):2073–2083, 2020.

[90] Kuo-Hua Huang, Peter Rupprecht, Thomas Frank, Koichi Kawakami, Tewis Bouwmeester, and Rainer W Friedrich. A virtual reality system to analyze neural activity and behavior in adult zebrafish. *Nature Methods*, pages 1–9, 2020.

[91] John R. Stowers, Anton L. Fuhrmann, Maximilian Hofbauer, Martin Streinzer, Axel Schmid, Michael H. Dickinson, and Andrew D. Straw. Reverse engineering animal vision with virtual reality and genetics. *Computer*, 47:38–45, 2014.

[92] Charles Darwin. *On the origin of species*. John Murray, 1859.

[93] Ronald A Fisher. The evolution of sexual preference. *The Eugenics Review*, 7(3):184, 1915.

[94] Russell Lande. Models of speciation by sexual selection on polygenic traits. *Proceedings of the National Academy of Sciences*, 78(6):3721–3725, 1981.

[95] Mark Kirkpatrick. Sexual selection and the evolution of female choice. *Evolution*, 36(1):1–12, 1982.

[96] Yoh Iwasa and Andrew Pomiankowski. The evolution of mate preferences for multiple sexual ornaments. *Evolution*, 48(3):853–867, 1994.

[97] Richard O Prum. The lande–kirkpatrick mechanism is the null model of evolution by intersexual selection: implications for meaning, honesty, and design in intersexual signals. *Evolution: International Journal of Organic Evolution*, 64(11):3085–3100, 2010.

[98] Eleanor M. Caves, Nicholas C. Brandley, and Sönke Johnsen. Visual acuity and the evolution of signals. *Trends in Ecology Evolution*, 33(5):358 – 372, 2018.

[99] Mathieu Joron and James LB Mallet. Diversity in mimicry: paradox or paradigm? *Trends in Ecology & Evolution*, 13(11):461–466, 1998.

[100] William A Searcy and Stephen Nowicki. *The evolution of animal communication: reliability and deception in signaling systems*. Princeton University Press, 2005.

[101] Derek A Roff. The evolution of mate choice: a dialogue between theory and experiment. *Annals of the New York Academy of Sciences*, 1360(1):1–15, 2015.

[102] John A Endler. Predation, light intensity and courtship behaviour in poecilia reticulata (pisces: Poeciliidae). *Animal Behaviour*, 35(5):1376–1385, 1987.

Appendix A Example Application to the Evolution of Color Patterns: Background

The incredible variety of color patterns seen in nature evolved under the selective forces imposed by the environment, and the visual experience of their receivers.[53–59] Quantifying this diversity, and reliably testing the functional significance of these traits is fundamental to understanding fitness landscapes[60] and underlies many subdisciplines of sensory ecology, cognitive neuroscience, collective behavior, and evolution.

Creating quantitative descriptions of color patterns which take into account the unique sensory and semiotic worlds of their receivers[61] has been a central challenge in visual ecology. Many tools have been developed: Quantitative

Colour Pattern Analysis,[62] PAVO,[63] Natural Pattern Match,[64] among others.[65–73] Each of these tools uses one or an ensemble of complimentary metrics from image analysis and computer vision, e.g. image statistics, edge detection, and landmark-based filters.[14]

Still, fundamental gaps remain. One of these gaps is the difficulty in building quantitative descriptions of complex features with multiple subelements. Most existing approaches fail to capture the full complexity of many of color patterns; the algorithms themselves are insufficiently expressive. This is particularly true when spatial or scale dependent relationships between features exist, e.g. the irregular patterns of male guppy ornamentation or butterfly wing patterns where similar sets of elements are arranged in species-specific configurations.[74] Recently, researchers have begun employing machine learning algorithms such a as non-linear dimensionality reduction, e.g. t-distribute stochastic neighbor embedding (t-SNE,[34,35] Figure 2), and deep neural networks (Figure 2,[15] Figure S1). Still, while these techniques can better represent more complex relationships between pixel values within an image, current implementations do not disentangle features across scales or provide extensions to downstream experiments.

While complex trait may be difficult to quantify, they are nonetheless biologically relevant in terms of feature context[75–80] and the perceptions of shape, motion, and attention.[57–59,81–83] And in the brain, we know that perception is hierarchically organized,[84] and representations made at higher levels of the visual cortex and its homologs heavily influence the perception of low-level features.[85,86] While measuring local features across an image provides important insight on regularity and the nature of wide-field variation, a collection of local feature descriptions across space is fundamentally different to a feature description built across scales.

Another gap is in building direct connections between approaches. Establishing spectral sensitivity, acuity, and feature importance is typically done using stimulus playback experiments or behavioral assays. However, beyond using statistical descriptions of features to guide researchers in the creation of stimuli there are few explicit connections between analysis and experiment. The current state of the art: immersive virtual reality (VR) and low-latency playback experiment—with fully animated, photo-realistic, 3D models, provide a rich experimental basis for investigating the relationship between visual inputs, neural activity, and behavior.[87–90] VR systems are also beginning to better account for species-specific sensory biases including photoreceptor sensitivity, flicker fusion rate, acuity, and depth perception.[89,91] Still, currently these approaches rely on human-in-the-loop interventions for creating stimulus with even moderate complexity.

Additionally, because color pattern traits have evolved under selective pressure from multiple receivers, establishing these types of evolutionary trade-offs is important to our understanding. However, experimental approaches often require large, highly disruptive manipulations such as translocation experiments or large scale crossbreeding experiments. Simulations and virtual experiments may better allows researchers to be explicit about the stimulus that is being tested and greatly reduce the number of subjects needed (Methods 4.4).

A.1 The potential impacts of this approach on the study of evolution

This platform may be used to address many outstanding questions regarding the functional significance of color pattern traits; here, we discuss some of these questions. *1)* What are the constraints on the evolvability of a given trait? By identifying the topographical relationship between different traits within the color pattern space we can test predictions about the selective forces acting on them related to their geometric relationships, e.g. the axes of variation in traits meant to communicate viability should show increased orthogonality compared to co-occurring traits which have evolved under a Fisherian process.[92–97] *2)* Categorical perception is an important perceptual mechanism for understanding the evolution of color signals.[98] But in systems where color patterns are used for mimicry[53,55,99] or novelty, investigating the boundaries between complex traits is fundamental. By performing traversals across the distribution of the latent variables, interpolating between samples can allow for tests of continuous[100] versus categorical perception[101] of complex traits. *3)* Many color pattern traits have evolved under selective pressure from multiple receivers, e.g. both females and predators shape the diversity of male guppy ornaments.[102] Establishing these types of evolutionary trade-offs is difficult and often requires large, highly disruptive manipulations such as translocation experiments. Using evolutionary models similar to the ones presented here researchers can simulate multiple fitness landscapes and evolutionary trajectories simultaneously to perform a broad range of virtual experiments. Importantly, while each of these examples place either analytical, experimental, or virtual results at the center, by using the platform presented here, they maintain direct connections across approaches. Furthermore, they can incorporate existing techniques[67–73] as image preprocessing routines, during playback, or constraints on virtual experiments.

Appendix B Supplemental Figures

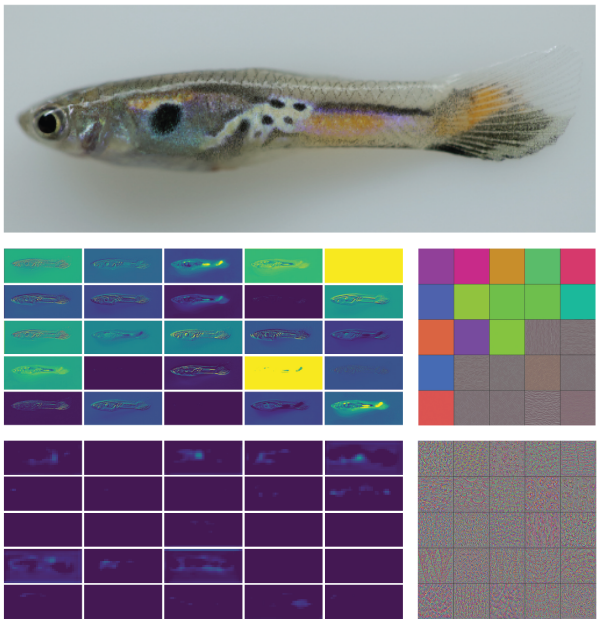

Figure S1: *Convolutional layers.* In typical *supervised* discriminative models, the objective being optimized is well defined, e.g. accurate classification or localization. As such, the representations provided by the downstream convolutional layers of deep networks take on characteristics optimized for task performance. At higher and higher network layers, the boundaries between classes can become complex and specialized to this objective because of the usefulness of such representations to identifying complex boundaries.Middle: Features learned at lower layers relate to color patches or gestures whereas at higher levels (bottom) fears become complex and interpretability can be difficult. Left, image pixels which are activated by pretrained image filters (yellow represents higher activations). Right, the maximally activating image feature for each filter.

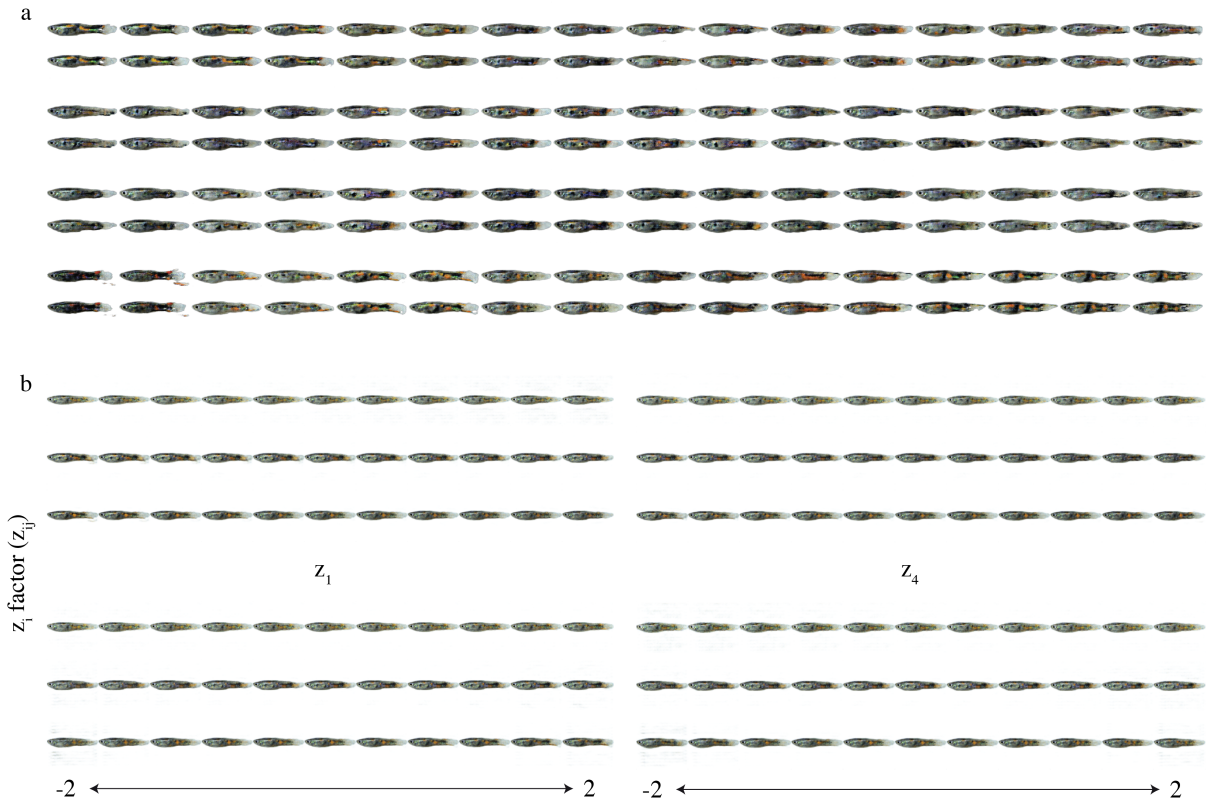

Figure S2: *Exploring latent variables*. a) We incorporated knowledge of our sample data by providing a 32-class categorical latent code and were able to generate examples from distinct classes learned by the model which capture meaningful combinations of features in our sample data. b) Latent traversal of 3 latent variables from $z_1, \ldots, z_4$. Top and middle are two embedded samples and bottom a latent code initialized at zero. For each latent variable (rows) we traverse values between -2 and 2 for the generated output. We see that each latent code has consistent effects. All samples in both a and b are generated from the generative models (b and c in Figure 1

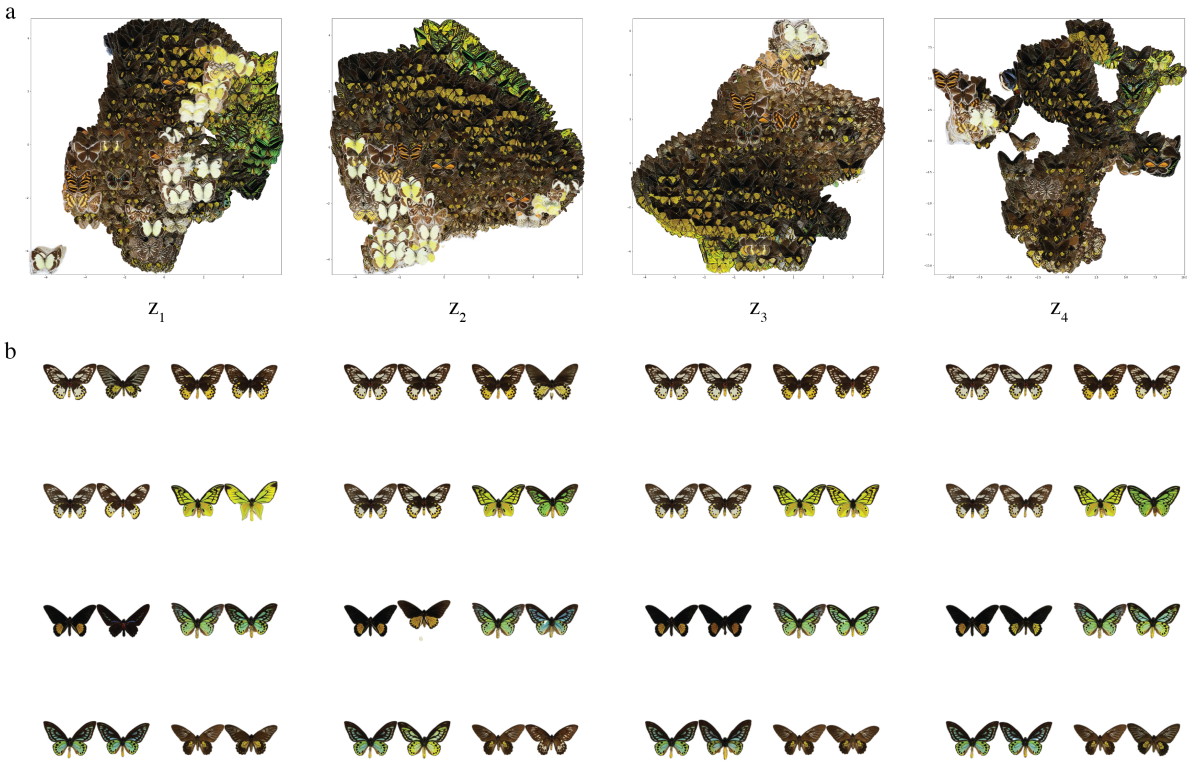

Figure S3: *Exploring latent representations*. a) 2D embedding (using tSNE[34,35]) of butterfly images using the 4 hierarchical latent encodings. The relationship between images at lower levels are dominated by color value similarities whereas at higher layers pattern elements at increasing spatial scales define the relationship between samples b) Nearest neighbors of 8 random samples based in the Minkowski distance[36] between the 10-dimensional space of each latent code $z_1, \ldots, z_4$

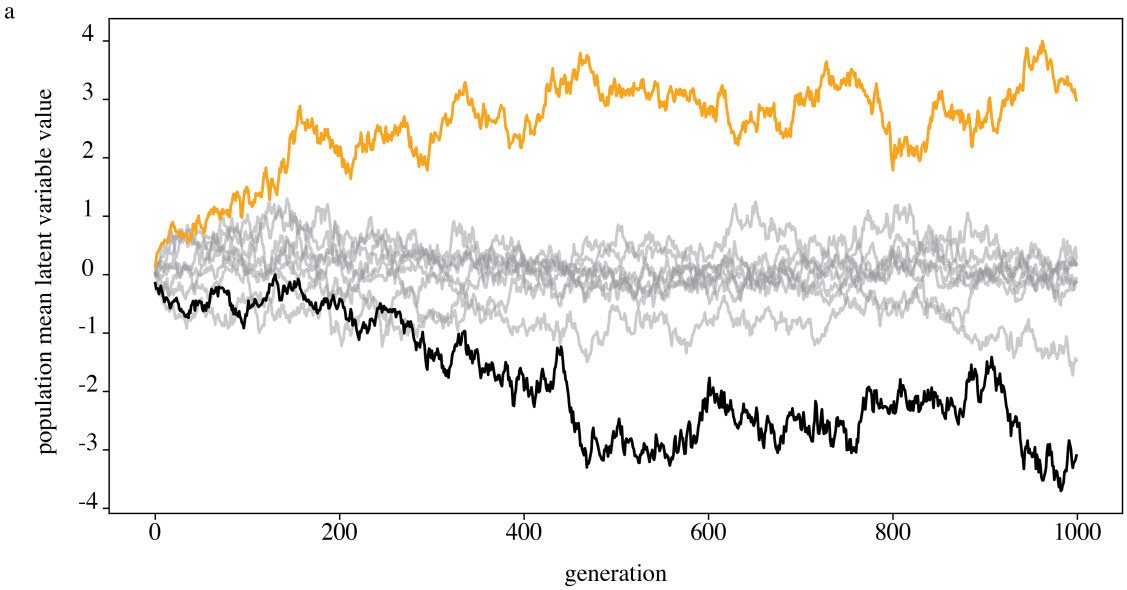

Figure S4: *Latent variable "alleles" over generations* a) We find two alleles are driven to fixation in the population after several generations selecting for oranger and higher contrast color patterns in guppies.

Figure S5: *Movie 1: The combined pattern space over 500 generations, visualized in 2D using tSNE*

Figure S6: *Movie 2: VR animation of learned coloration pattern models to an animated guppy for virtual playback.*