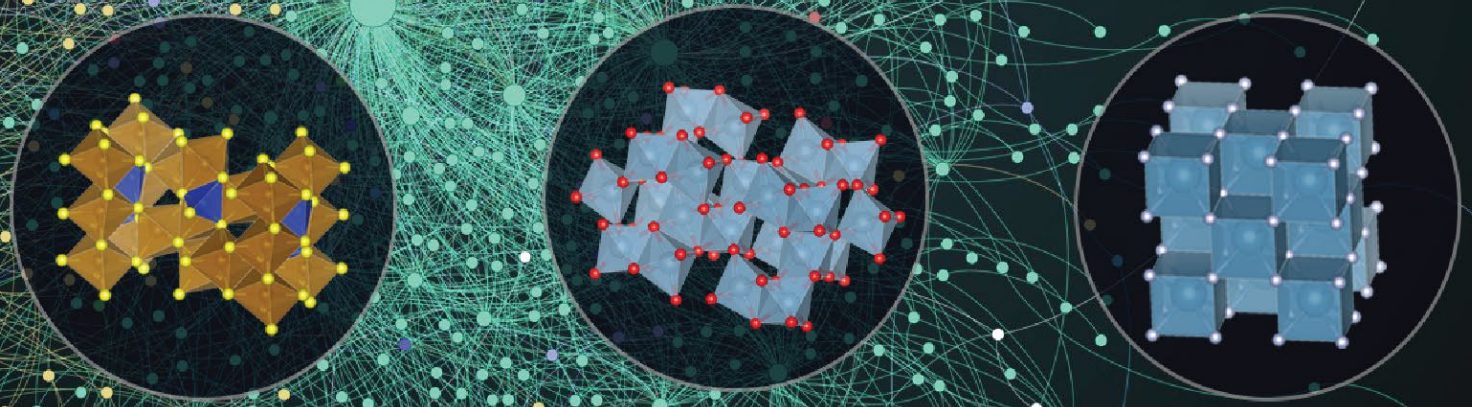
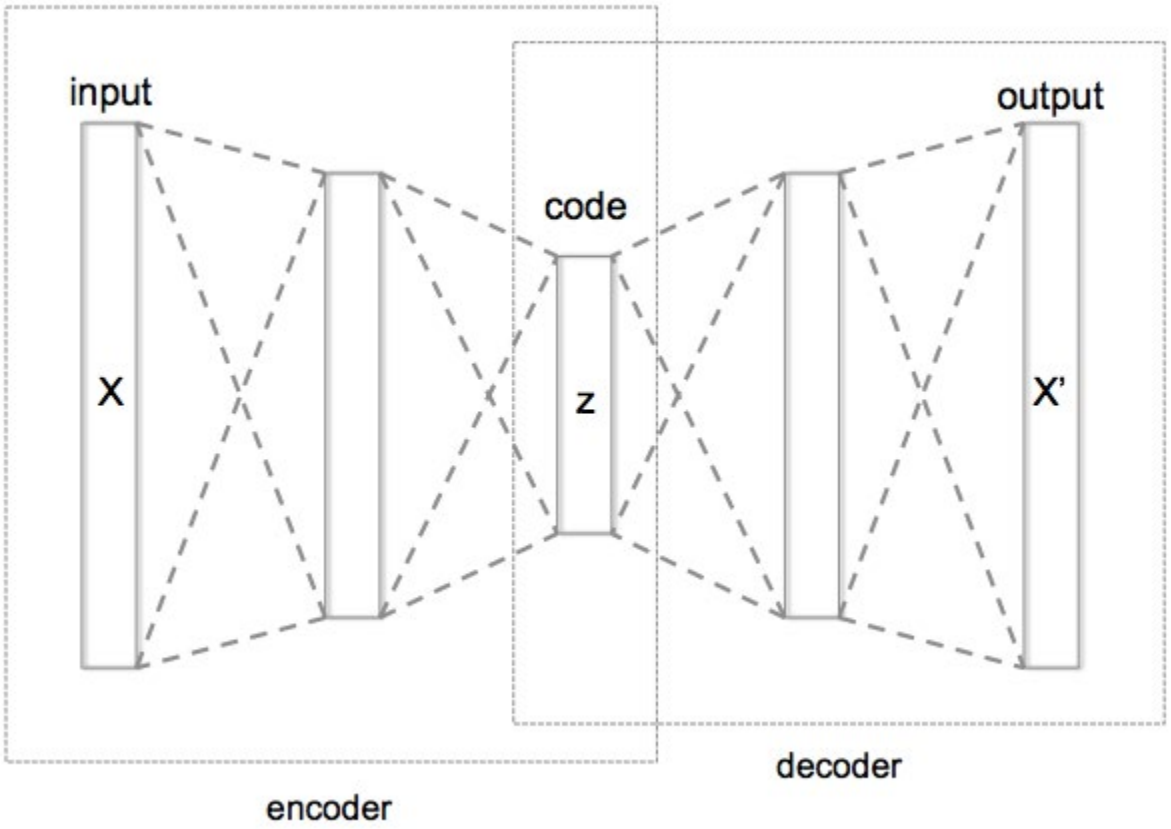
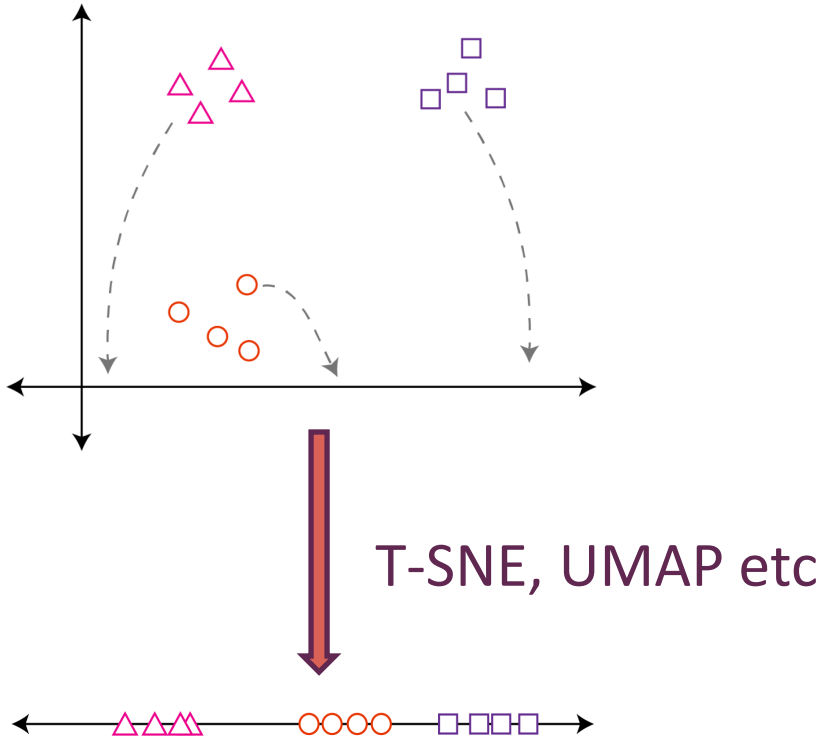
variational autoencoders



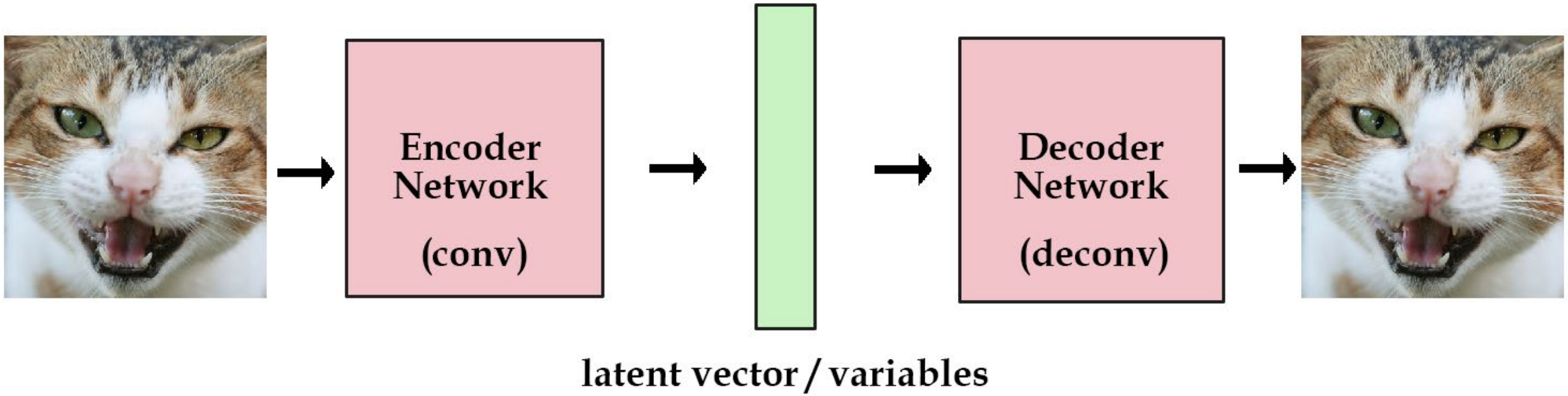
VAEs are modified versions of autoencoders that efficiently encode data





https://en.wikipedia.org/wiki/Autoencoder

Autoencoders are made up of two important steps



https://kvfrans.com/variational-autoencoders-explained/

Technically this is an entirely new type of compression mechanism!

Left: Original image (1419 KB PNG) at ~1.0 MS-SSIM. **Center:** JPEG (33 KB) at ~0.9 MS-SSIM. **Right:** Residual GRU (24 KB) at ~0.9 MS-SSIM. This is 25% smaller for a comparable image quality

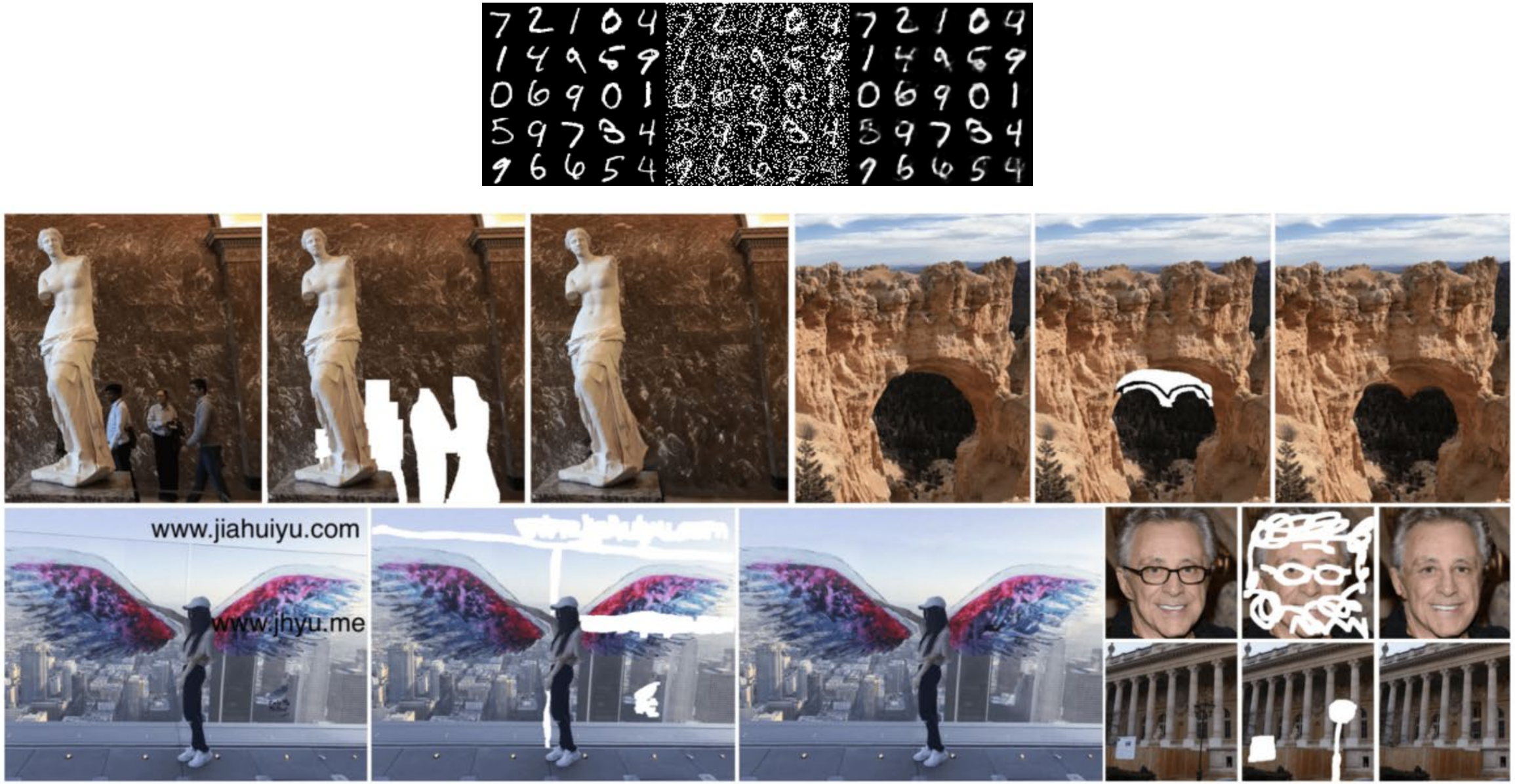
Taking a look around his nose and mouth, we see that our method doesn't have the magenta blocks and noise in the middle of the image as seen in JPEG. This is due to the blocking artifacts produced by JPEG, whereas our compression network works on the entire image at once. However, there's a tradeoff -- in our model the details of the whiskers and texture are lost, but the system shows great promise in reducing artifacts.

Left: Original. **Center:** JPEG. **Right:** Residual GRU.

While today's commonly used codecs perform well, our work shows that using neural networks to compress images results in a compression scheme with higher quality and smaller file sizes. To learn more about the details of our research and a comparison of other recurrent architectures, check out our paper. Our future work will focus on even better compression quality and faster models, so stay tuned!

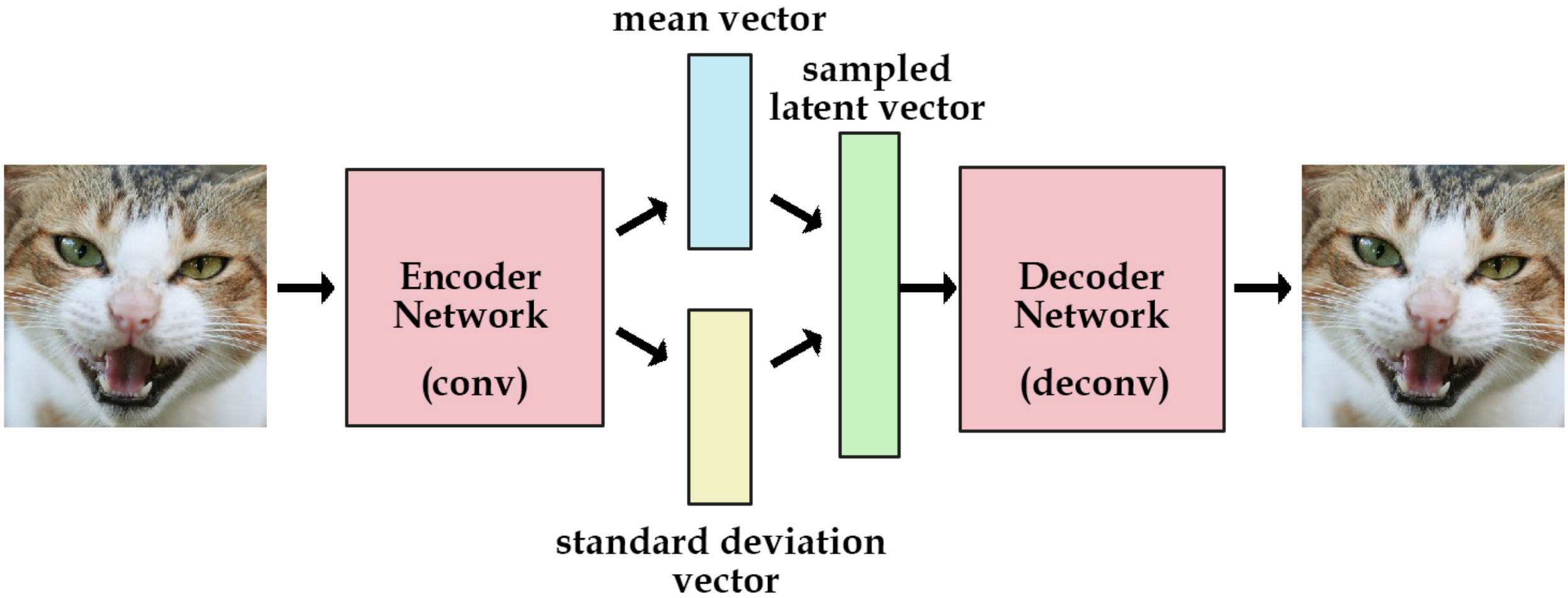
https://ai.googleblog.com/2016/09/image-compression-with-neural-networks.html

Autoencoders can also be used for denoising and neural inpainting



https://towardsdatascience.com/10-papers-you-must-read-for-deep-image-inpainting-2e41c589ced0

So how is a variational autoencoder different?



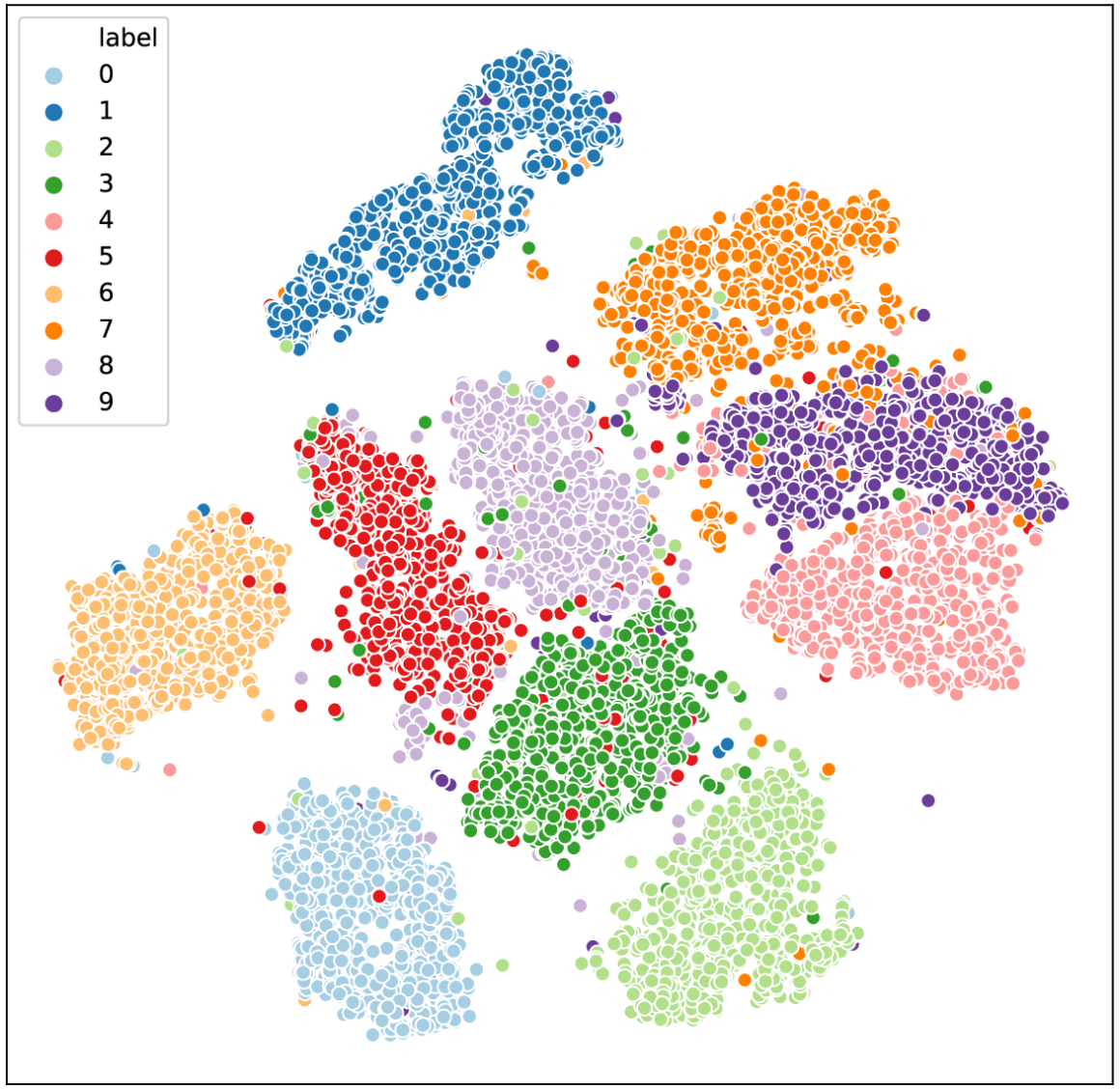
https://kvfrans.com/variational-autoencoders-explained/

To make model generative, we need to sample from latent space

Backpropagation would not be able to work through a simple sampling operator

Fix this with “reparameterization trick”

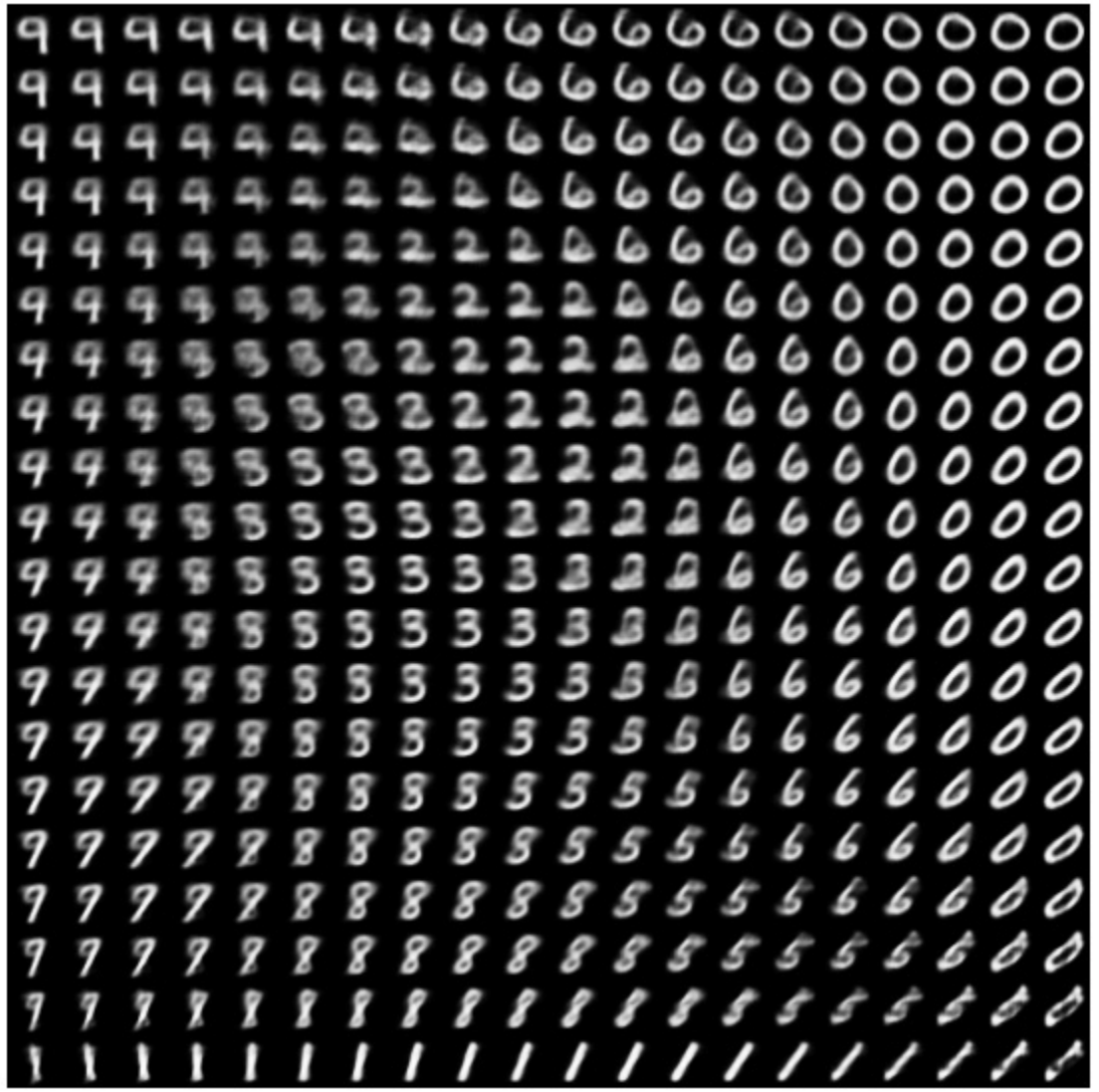
$$vector = \mu + \sigma\epsilon$$



Why can’t we just give latent vector some random numbers?

We would get garbage out. It needs to come from a normal distribution

Done correctly, the latent space is continuous! So we can systematically alter vector to see how output image changes



Let's define the math of the VAE process

Encoder: given a dataset *X* consisting of N i.i.d. samples $\{x^{(i)}\}_{i=1}^{N}$, we aim to model each *x* as generated by some latent variables $z = \mu + \sigma \cdot \epsilon$ where ϵ is random noise

These variables define a Gaussian distribution $q_\phi(z|x)$ for the latent variable z given an input x where ϕ denotes the parameters of the encoder

Decoder: The decoder takes a point in the latent space z and attempts to reconstruct the original input x. The decoder defines the probability distribution $p_\theta(x|z)$, where θ denotes the parameters of the decoder. The goal is to generate outputs that closely match the original inputs, thus learning a good representation of the data in the latent space.

Training VAEs and the objective function

Training involves optimizing the parameters ϕ and θ for encoder and decoder, respectively

The objective function is Evidence Lower Bound (ELBO) which does two things

$$ELBO = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\theta(z|x)||p_\theta(z))$$

Reconstruction loss encourages the reconstructed outputs to be as close as possible to the original inputs, promoting accurate representation of the data. It measures the likelihood of the data given the latent variables, aiming to maximize this likelihood.
KL divergence regularizes the encoder by penalizing deviations of the latent variable distribution $q_\theta(z|x)$ from the prior distribution $p_\theta(z)$. In simple terms, it keeps the representations of different data points distinct and encourages the latent variables to follow the prior distribution, which is typically chosen to be a standard normal distribution $N(0, I)$ for simplicity.

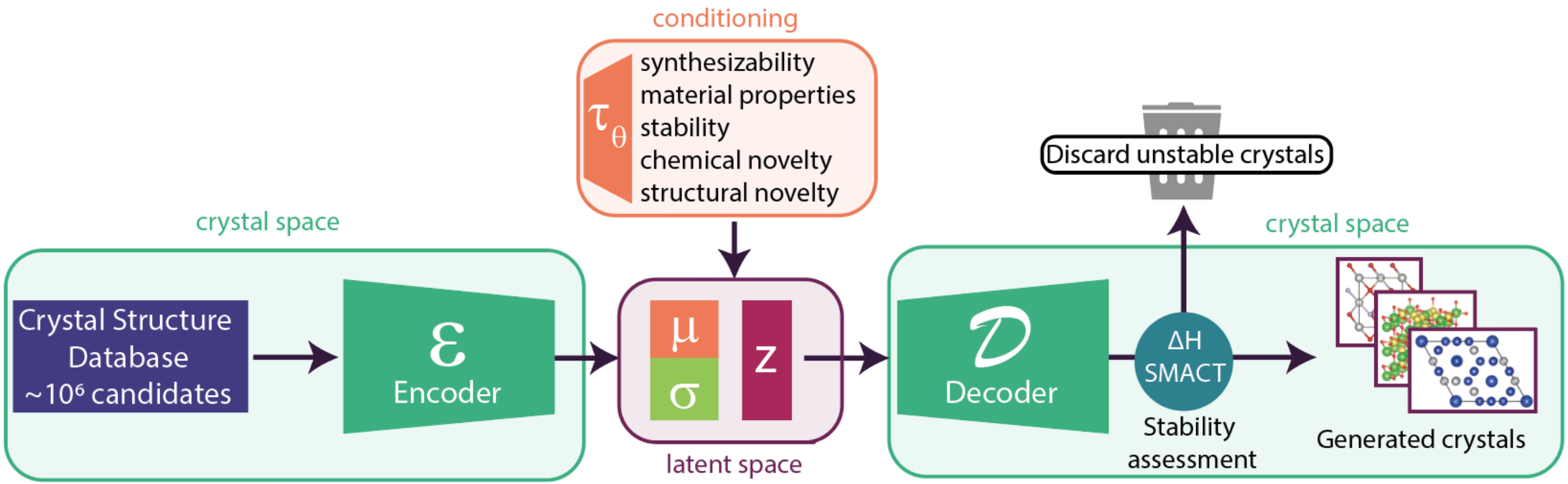
Generating new samples

The generative process can be described as follows:
1.For each *x*, draw a latent variable $z \sim p_\theta(z)$, where $p_\theta(z)$ is the prior over the latent variables.
2.Given z, generate x from $p_\theta(x|z)$, the conditional probability of x given z parameterized by θ

The goal is to maximize the marginal likelihood of the observed data $p_\theta(x)$, (which is the probability of observing x without considering z). Mathematically, we get marginal likelihood by integrating out the latent variable z from the joint distribution $p_\theta(x, z) = p_\theta(x|z)p_\theta(z)$ which is usually intractable due the complexity of integration over all possible z values.

VAEs tackle this by introducing an approximate posterior distribution $q_\phi(z|x)$, which is just the output of the encoder

Conditional variational autoencoders also exist



To understand conditional VAEs we have to introduce Bayesian inference and probability theory! (much more on these topics coming soon)

So how do we make these VAEs conditional?

The main difference between VAE and CVAEs is that we condition both the encoder and the decoder on some additional information c such as class labels

1.The encoder now approximates $q_\theta(z|x,c)$, the conditional distribution of z given x and c
2.The decoder models $p_\phi(x|z,c)$, the conditional distribution of x given z and c

The ELBO for CVAEs thus becomes:

$$ELBO_{CVAE} = \mathbb{E}_{q_\phi(z|x,c)}[\log p_\theta(x|z,c)] - D_{KL}(q_\theta(z|x,c)||p_\theta(z,c))$$

Notice that the prior over the latent variables can also be conditioned on c i.e. $p_\theta(z|c)$ which allows the model to generate data that is condition-specific.

Disentangled autoencoders are also available

We add one parameter in loss function, β, that measures how much the scaled divergence is present in the loss function

If β is too small, then you get overfitting (high variance)
If β is too large, then you get poor performance (high bias)

$$\mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}, \beta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta \, D_{KL}\big(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\big)$$

You can see that increasing β causes disentangled to only use 5 instead of all 10

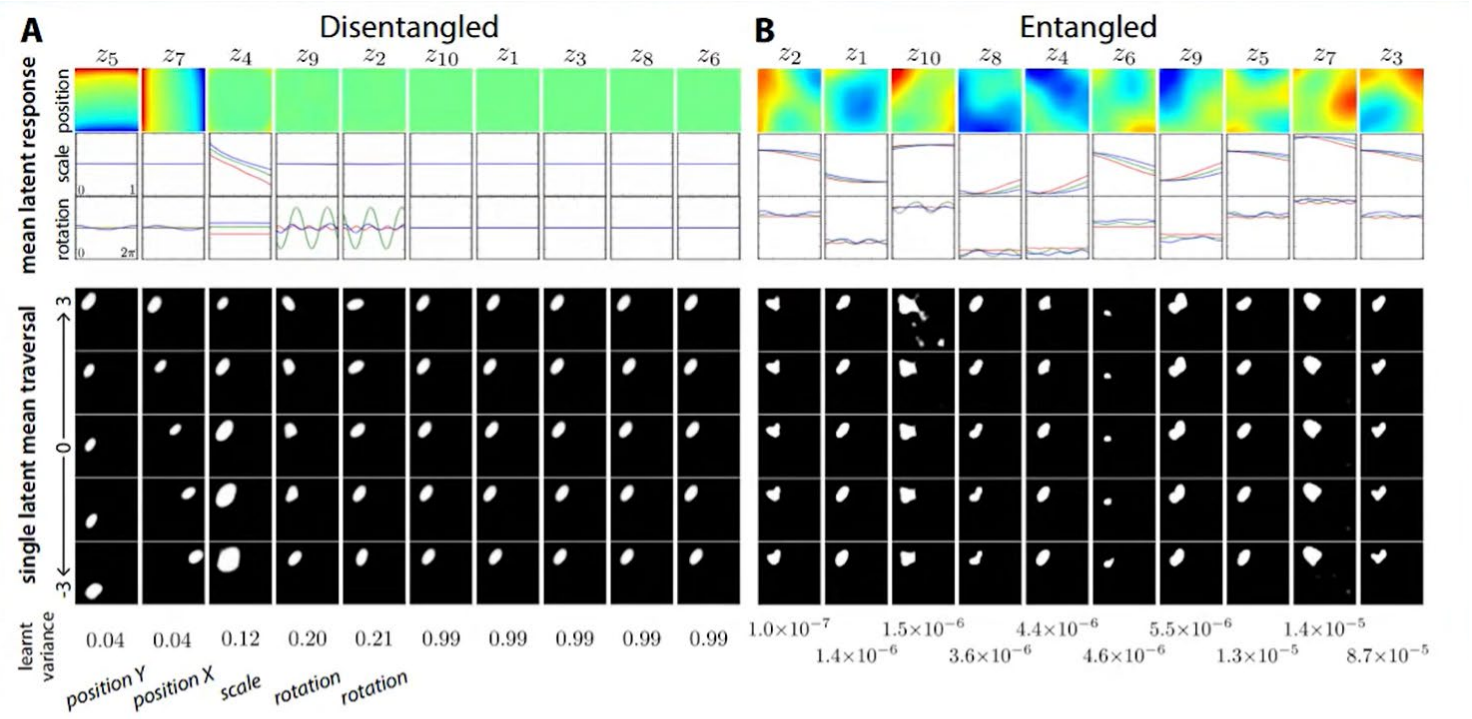


Figure 2: **A:** Disentangled representation learnt with $\beta = 4$. Each column represents a latent z_i, ordered according to the learnt Gaussian variance (last row). Row 1 (position) shows the mean activation (red represents high values) of each latent z_i as a function of all 32x32 locations averaged across objects, rotations and scales. Row 2 (scale) shows the mean activation of each unit z_i as a function of scale (averaged across rotations and positions). Row 3 (rotation) shows the mean activation of each unit z_i as a function of rotation (averaged across scales and positions). *Square* is red, *oval* is green and *heart* is blue. Rows 4-8 (second group) show reconstructions resulting from the traversal of each latent z_i over three standard deviations around the unit Gaussian prior mean while keeping the remaining 9/10 latent units fixed to the values obtained by running inference on an image from the dataset. After learning, five latents learnt to represent the generative factors of the data, while the others converged to the uninformative unit Gaussian prior. **B:** Similar analysis for an entangled representation learnt with $\beta = 0$.

http://jck.bio/disentangling_a_latent_space/

We can see that specific latent factors do specific things

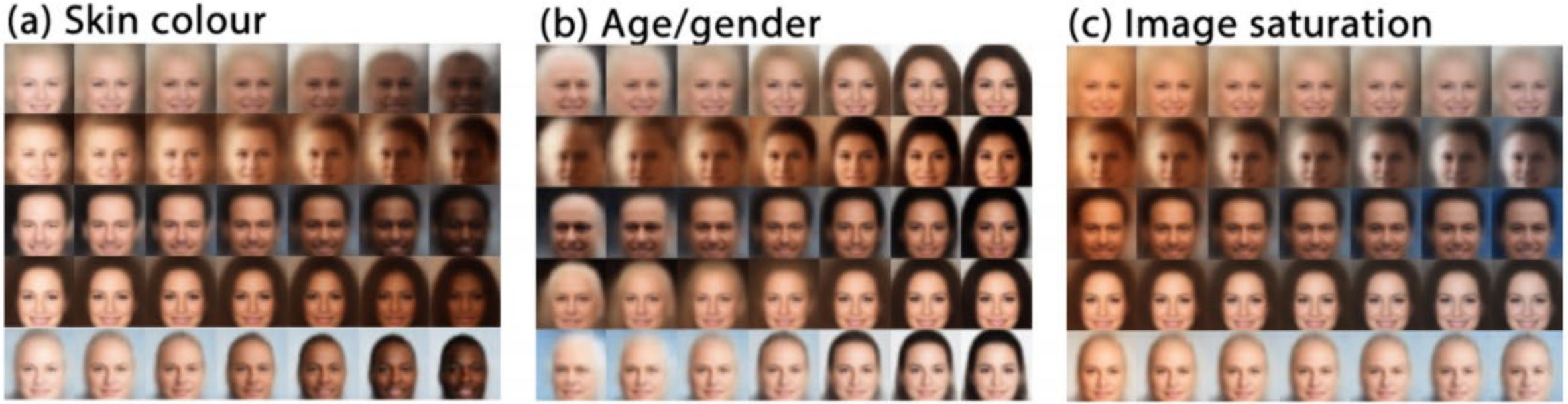


Figure 4: **Latent factors learnt by β-VAE on celebA:** traversal of individual latents demonstrates that β-VAE discovered in an unsupervised manner factors that encode skin colour, transition from an elderly male to younger female, and image saturation.

Comparing GANs with VAEs

Similarities: they both have two components. GAN has generator and discriminator. VAE has encoder and decoder

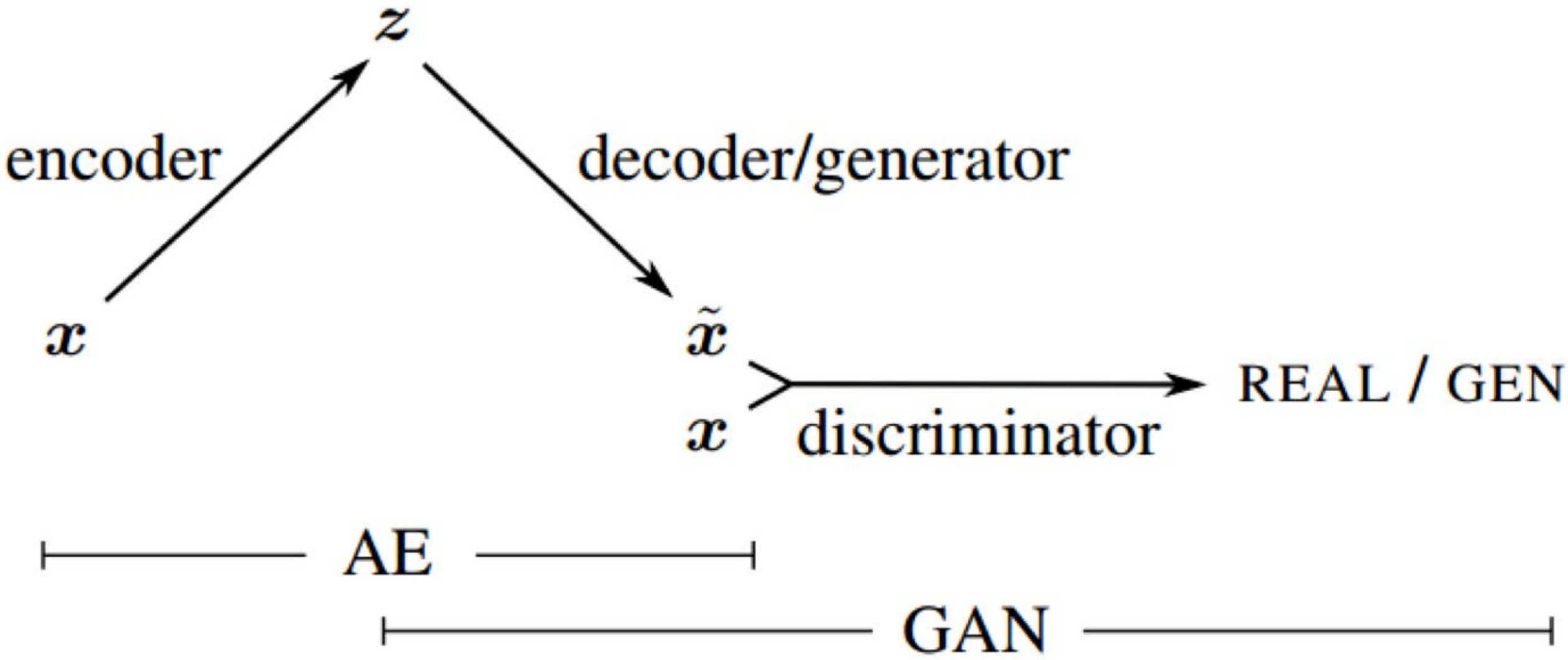
GAN plays minimax game whereas VAE tries to minimize reconstruction and latent loss

GAN not very stable because it requires finding “Nash Equilibrium” or end of game point and we don’t know how to do that

VAEs have a closed form end of training solution (minimize loss)

How good are they? VAEs are ok in theory, but make blurry images, GANs are much sharper because training is more empirical (trial and error)

Recently people have combined VAEs and GANs to create VAE-GANs



https://arxiv.org/abs/1512.09300

Comparing GANs with VAEs

Diffusion models