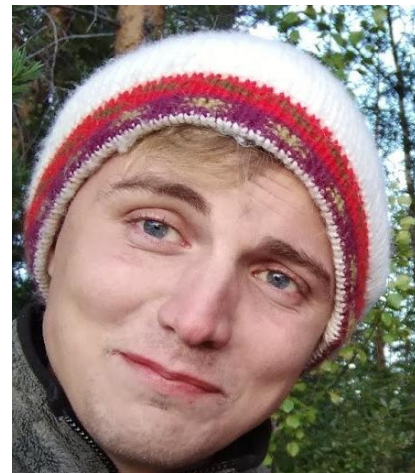


Network-to-Network Translation with Conditional Invertible Neural Networks

Robin Rombach*,



Patrick Esser*,

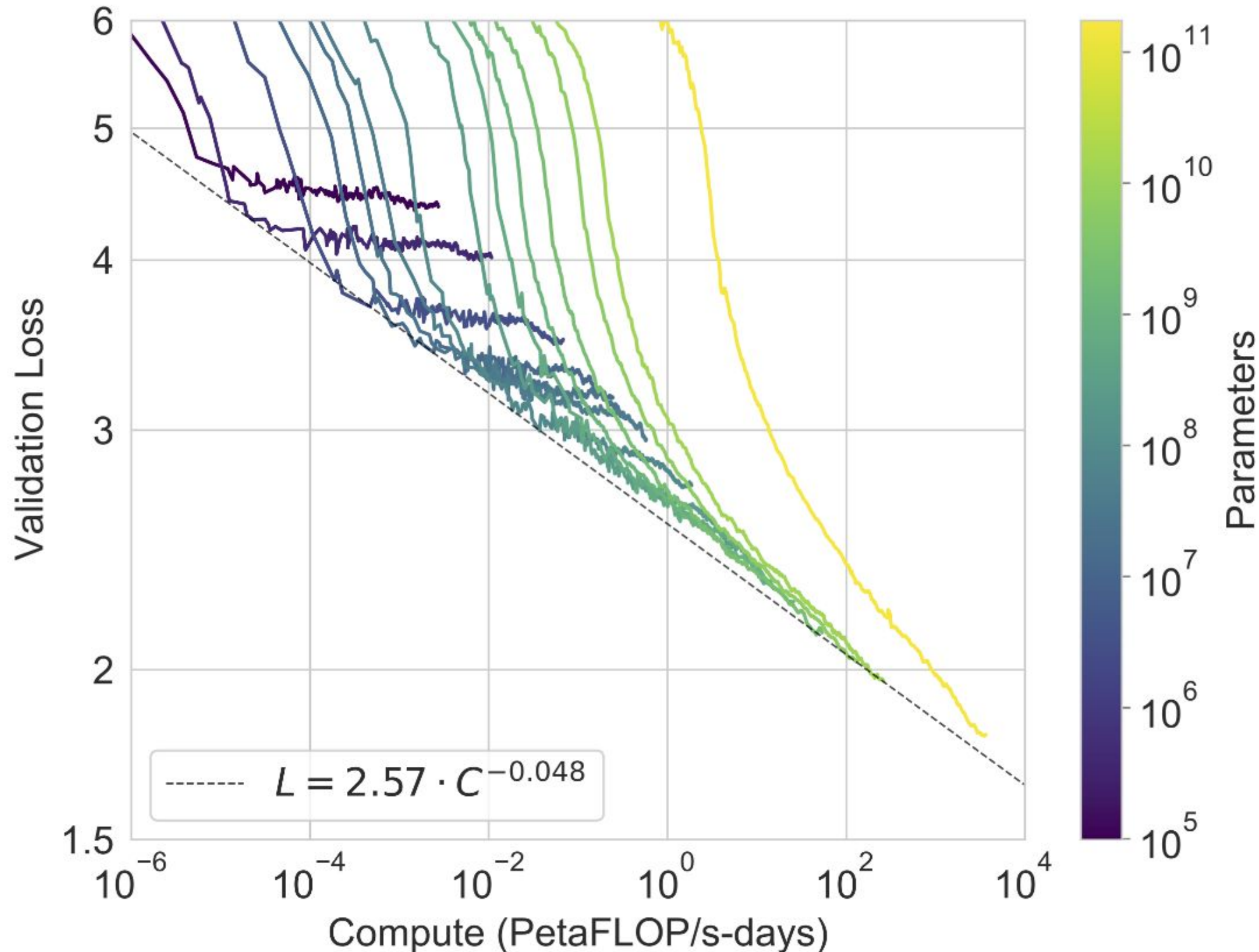


Björn Ommer





*equal contribution

The Bitter Lesson

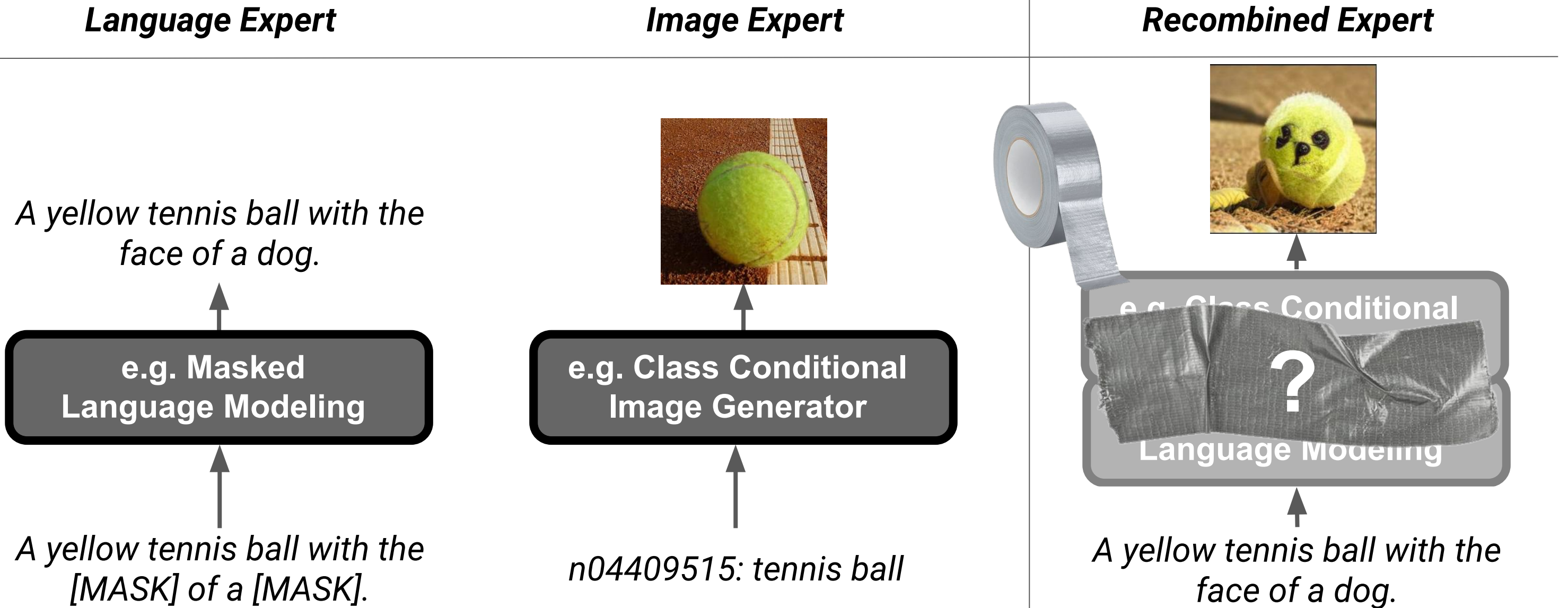


State-of-the-art
models are and
will be **huge**.

Infeasible to train and experiment with large models

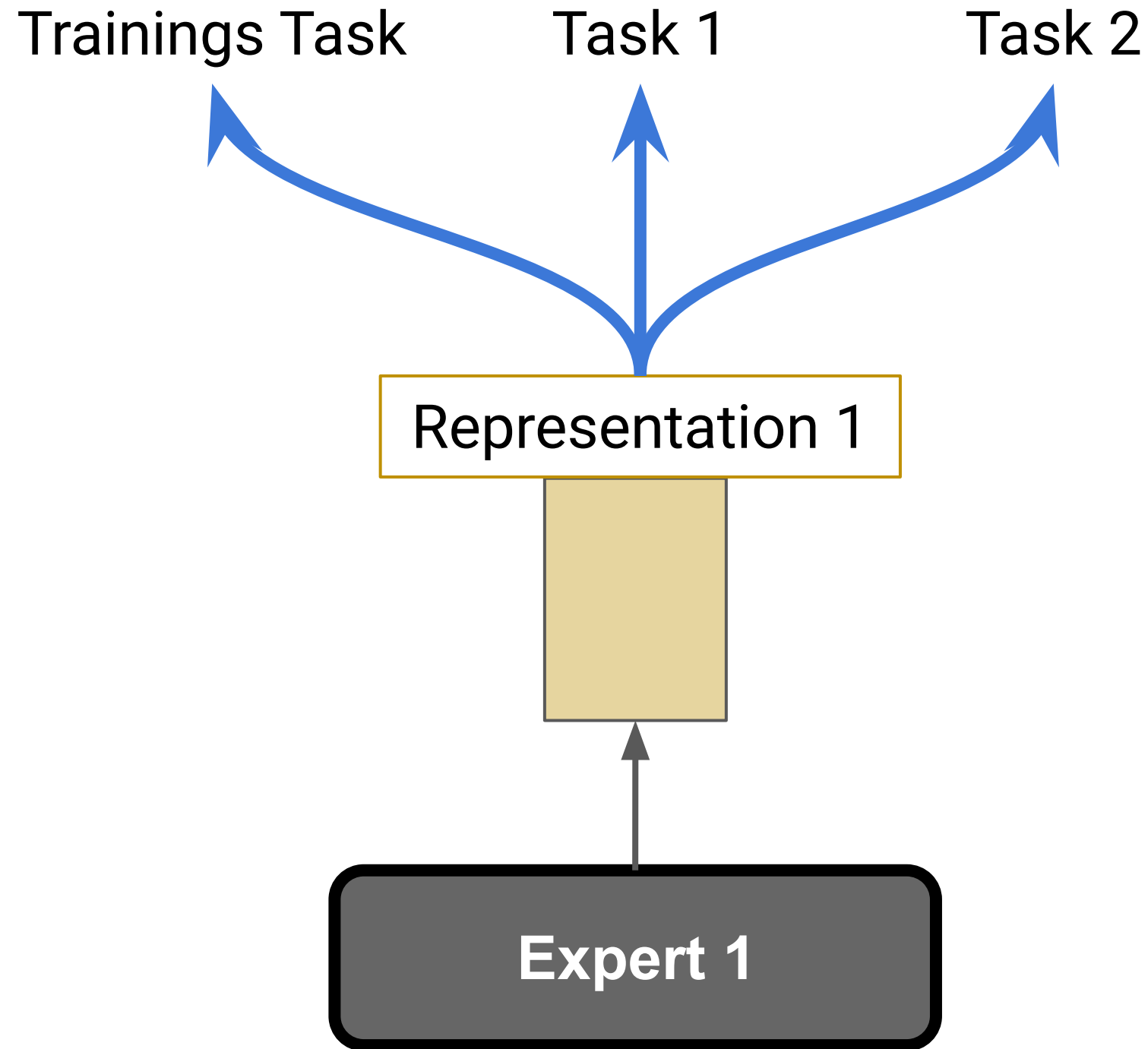
Estimated Training Costs on NVIDIA DGX-1	Model	Time	Cost	C02
	BigGAN	15 days	272.16 €	372.96 kg
	FUNIT	14 days	254.02 €	348.10 kg
<div><div><p>Input: The man went to the [MASK]₁ . He bought a [MASK]₂ of milk .</p><p>Labels: [MASK]₁ = store; [MASK]₂ = gallon</p><p>Sentence A = The man went to the store. Sentence B = He bought a gallon of milk. Label = IsNextSentence</p></div><div><p>Sentence A = The man went to the store. Sentence B = Penguins are flightless. Label = NotNextSentence</p></div></div>	BERT	10.3 days	186.88 €	256.10 kg

Must find ways to make optimal use of available models

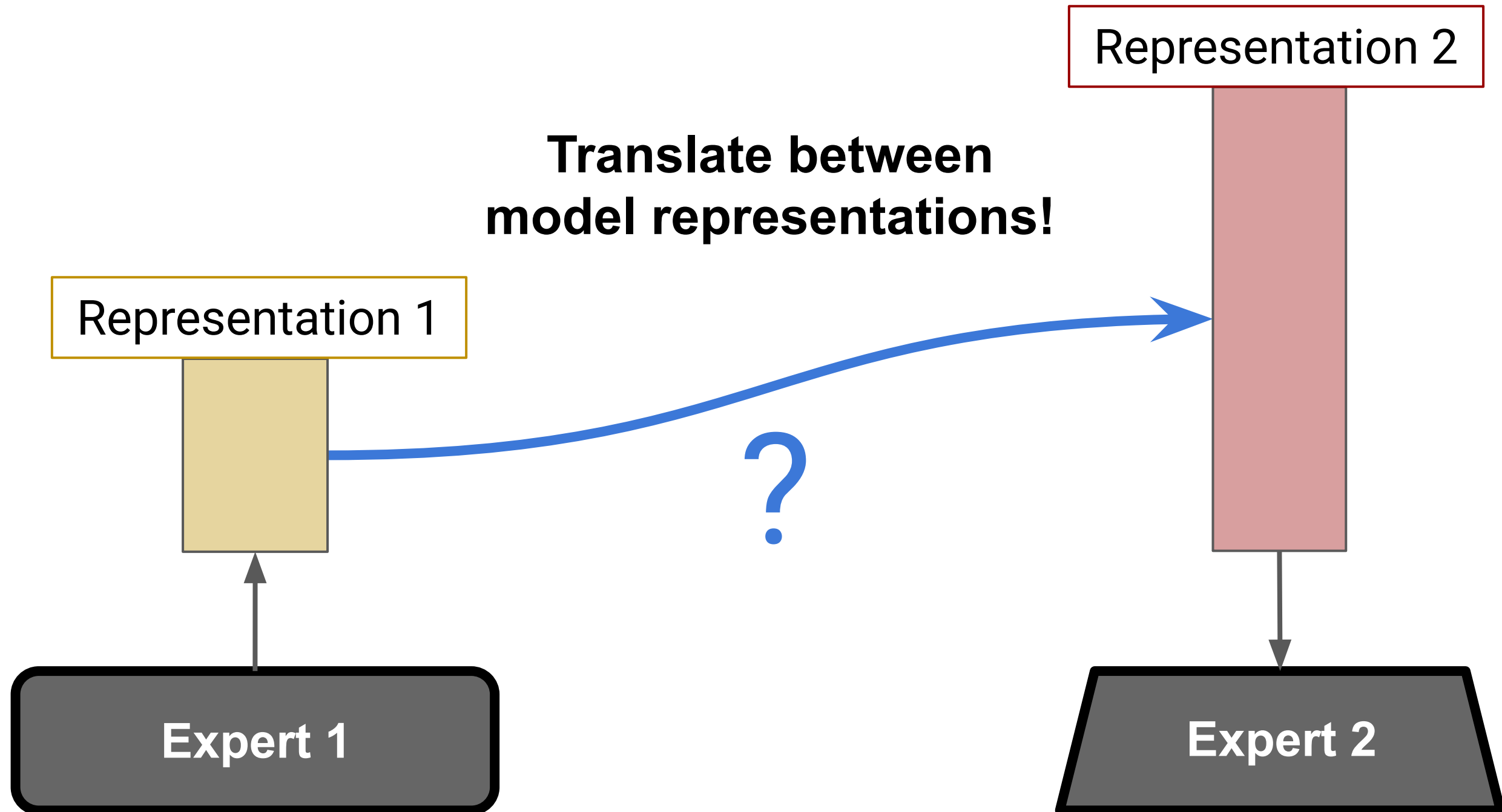


How can we **combine and reuse experts** to **solve new tasks** which neither of them can perform on its own?

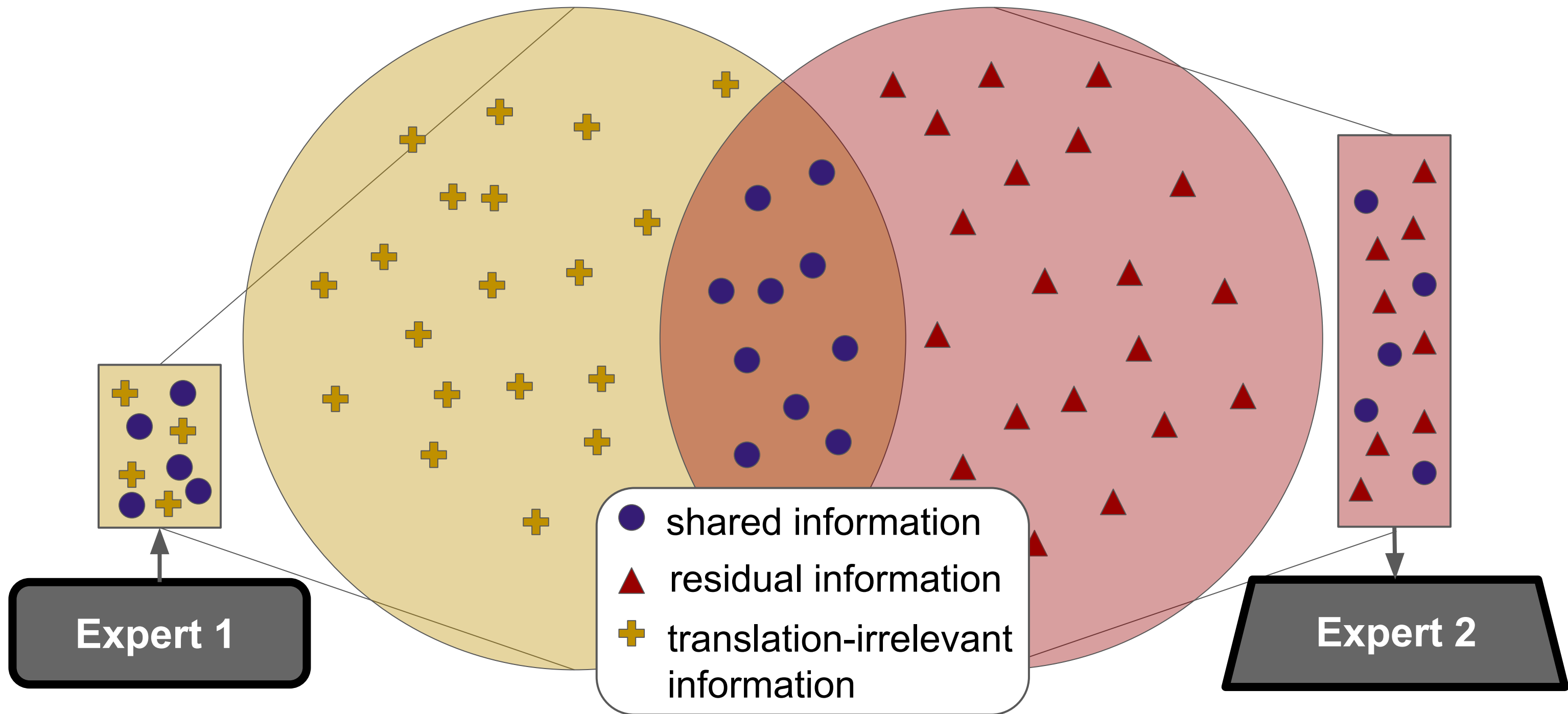
Hidden representations are awesome and reusable



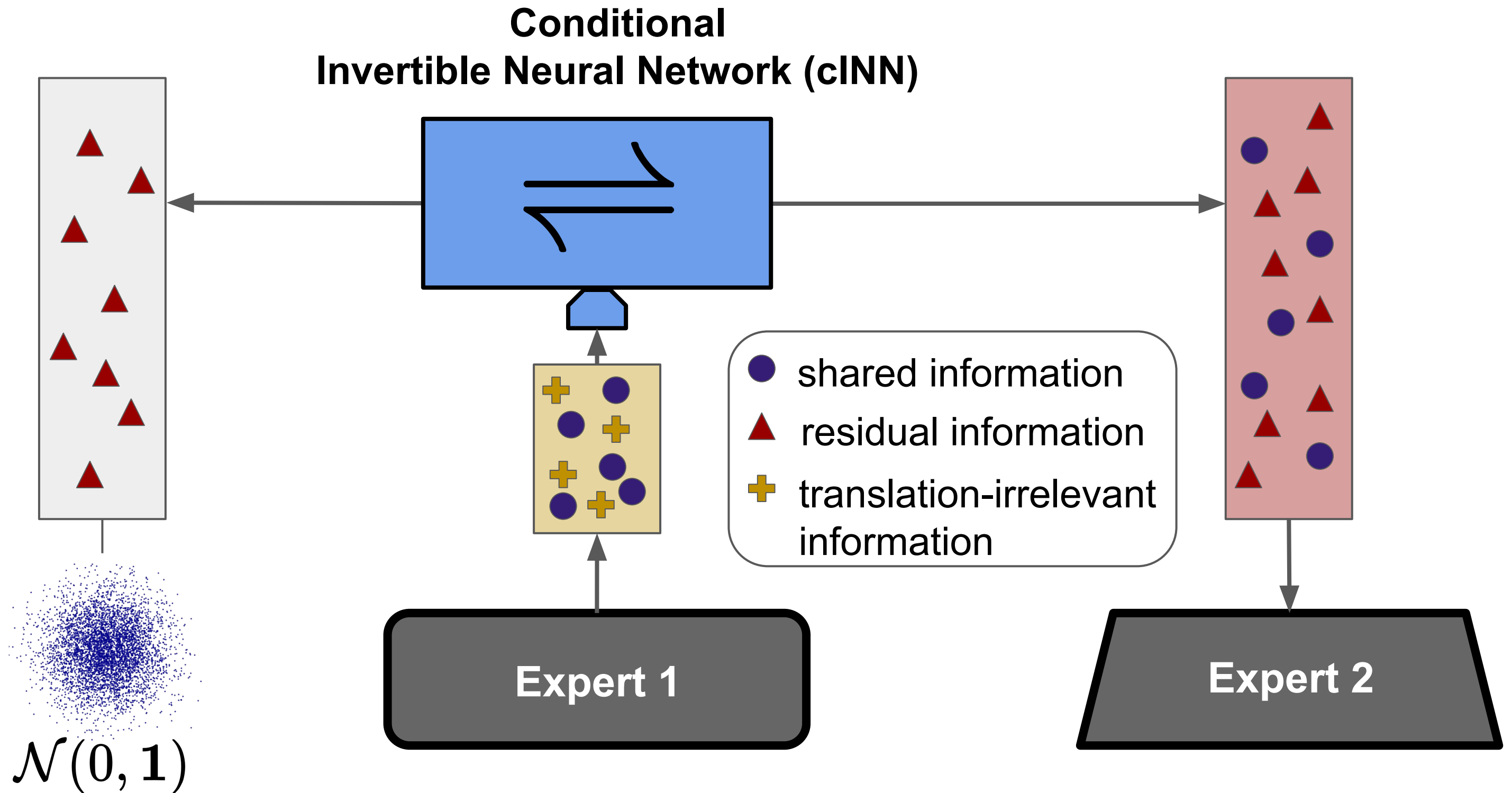
Hidden representations contain distilled expert knowledge



Shared and disjunct information in representations



Network-to-Network Translation



Training for Translation

Train in reverse direction

Disentangle v from z_1

$$\min I(v, z_1)$$

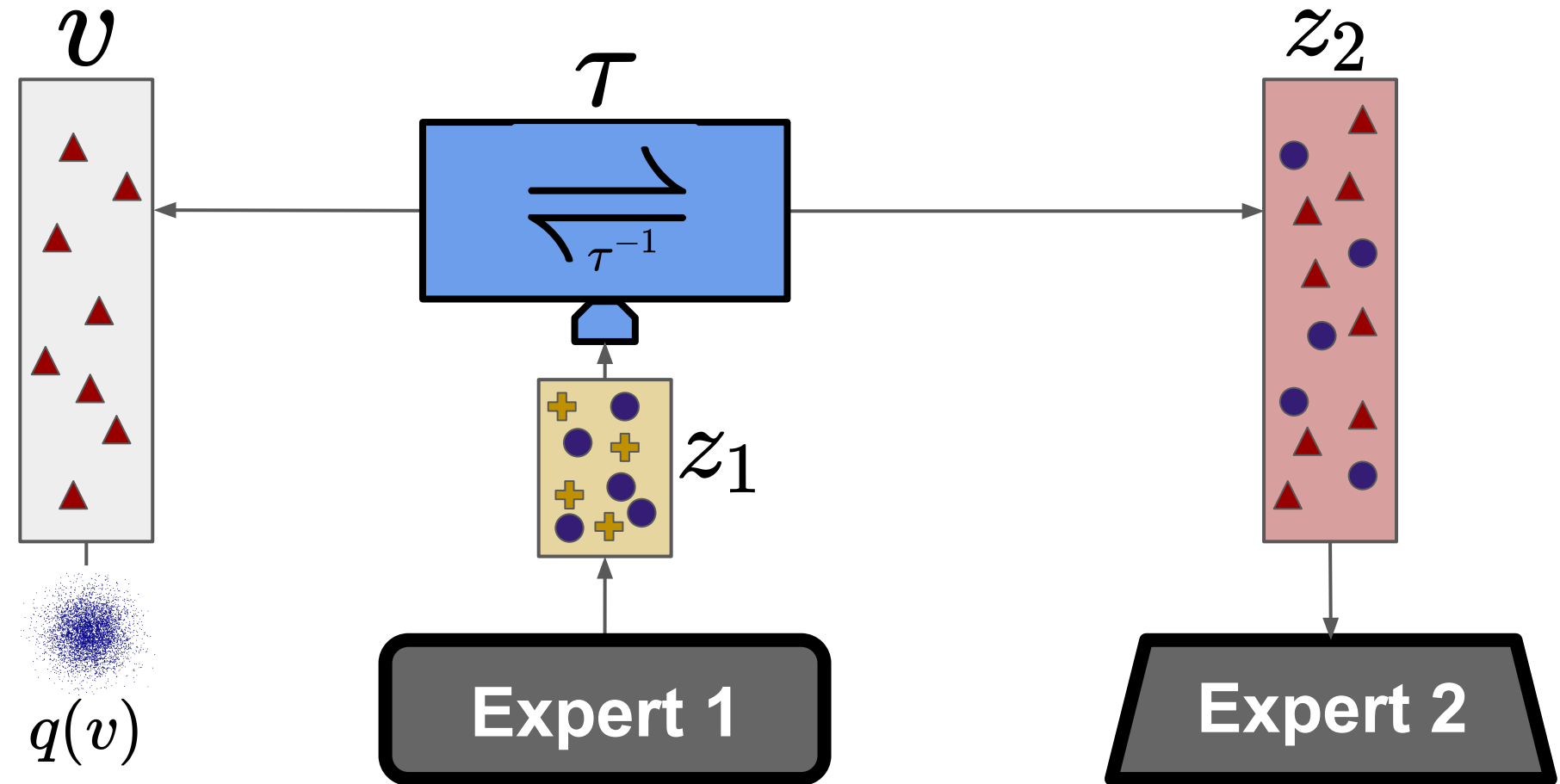
Translate in forward direction

Sample possible translations

$$z_2 \sim p(z_2|z_1)$$

via

$$z_2 = \tau(v|z_1), v \sim q(v)$$



Loss:

$$\begin{aligned} I(v, z_1) &\leq \mathbb{E}_{z_1} \text{KL} (p(v|z_1) || q(v)) \\ &= \mathbb{E}_{z_1, z_2} \left[-\log q(\tau^{-1}(z_2|z_1)) - |\det J_{\tau^{-1}}(z_2|z_1)| \right] - \mathbb{H}[z_2|z_1] \end{aligned}$$

Experiment #1: (S)BERT-to-BigGAN

Data: Image-Text Pairs

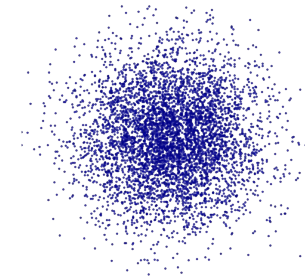
*A yellow and black
bird sitting in the
grass*

Captioning Model



Sentence-BERT













cINN















BigGAN
Generator

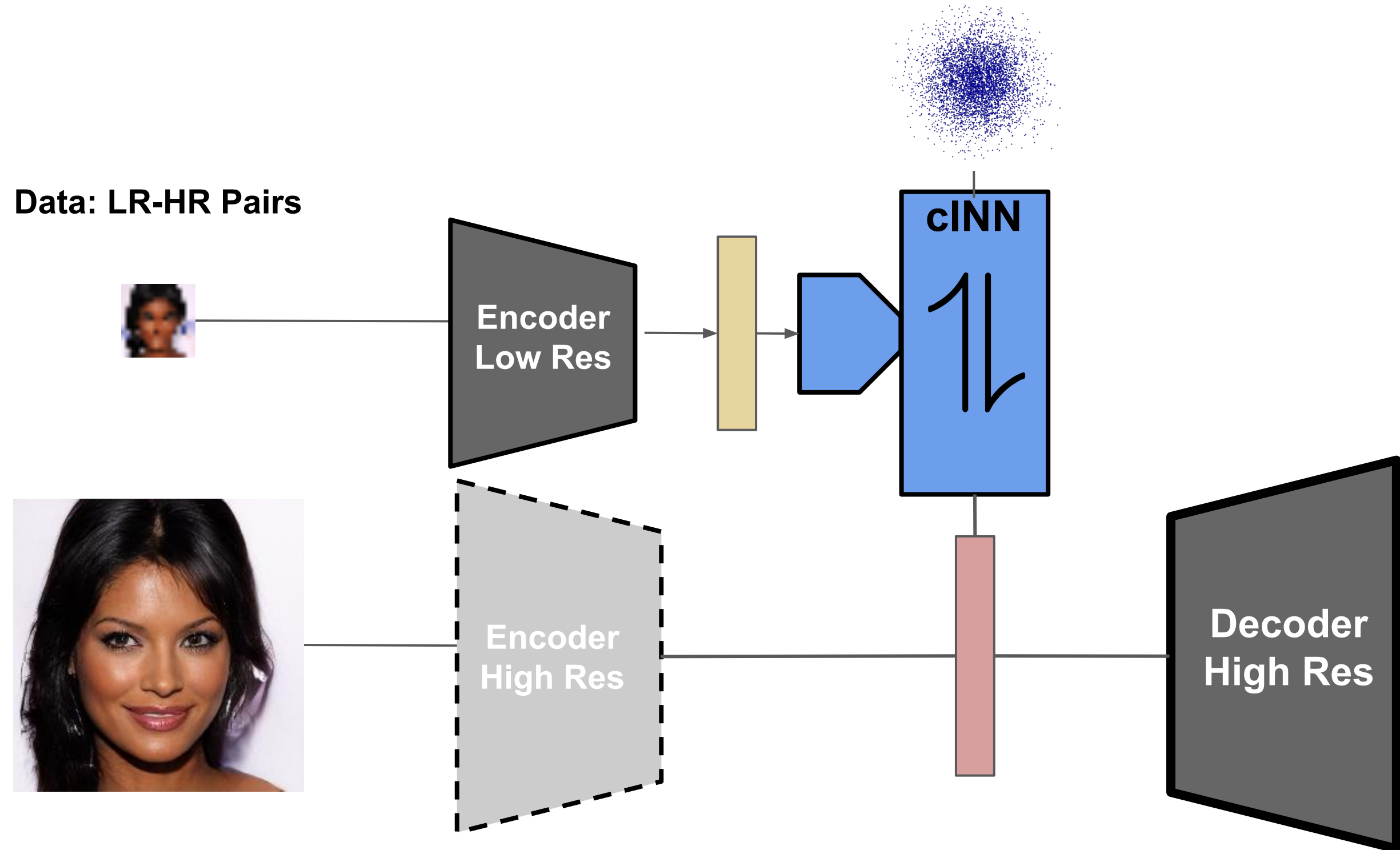
Results: (S)BERT-to-BigGAN

Text-to-Image translation between Sentence-BERT and BigGAN; utilize a captioning model to produce captions of BigGAN images during training.

<i>A blue bird sitting on top of a field</i>			
<i>A yellow bird is perched on a branch</i>			
<i>A school bus parked in a parking lot</i>			
<i>Two people on a paddle boat in the water</i>			

<i>A close up of a plant with broccoli</i>			
<i>A fighter jet flying through a cloudy sky</i>			
<i>A pizza sitting on top of a white plate</i>			
<i>A man riding skis down a snow covered slope</i>			

Experiment #2: Superresolution with AE-to-AE



Results: Superresolution with Net2Net

Animalfaces 16×16 to 256×256



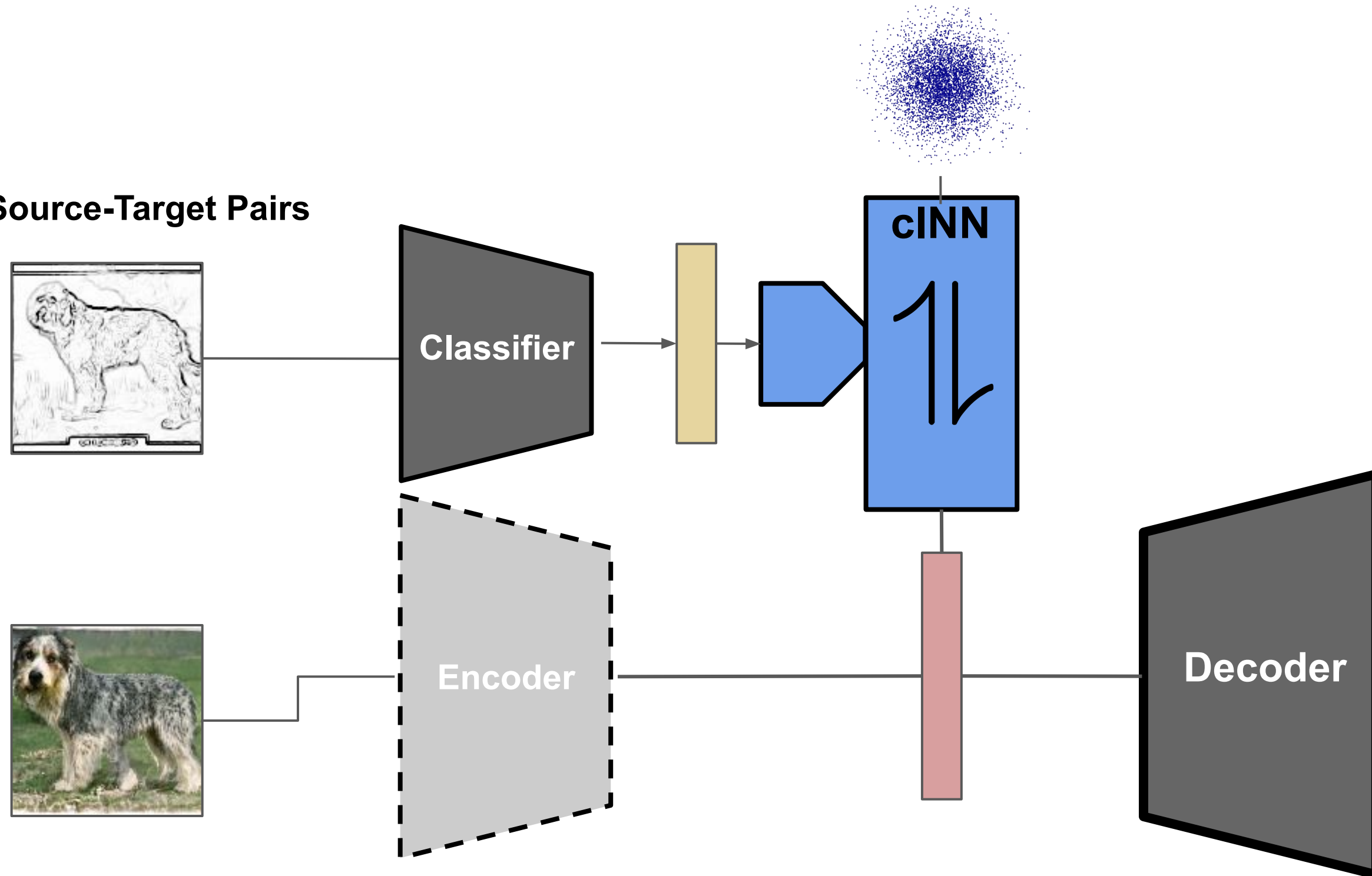
CelebA-HQ/FFHQ 32×32 to 256×256



→ **Combine experts from different scales**

Experiment #3: Image-to-Image with Suitable Classifiers

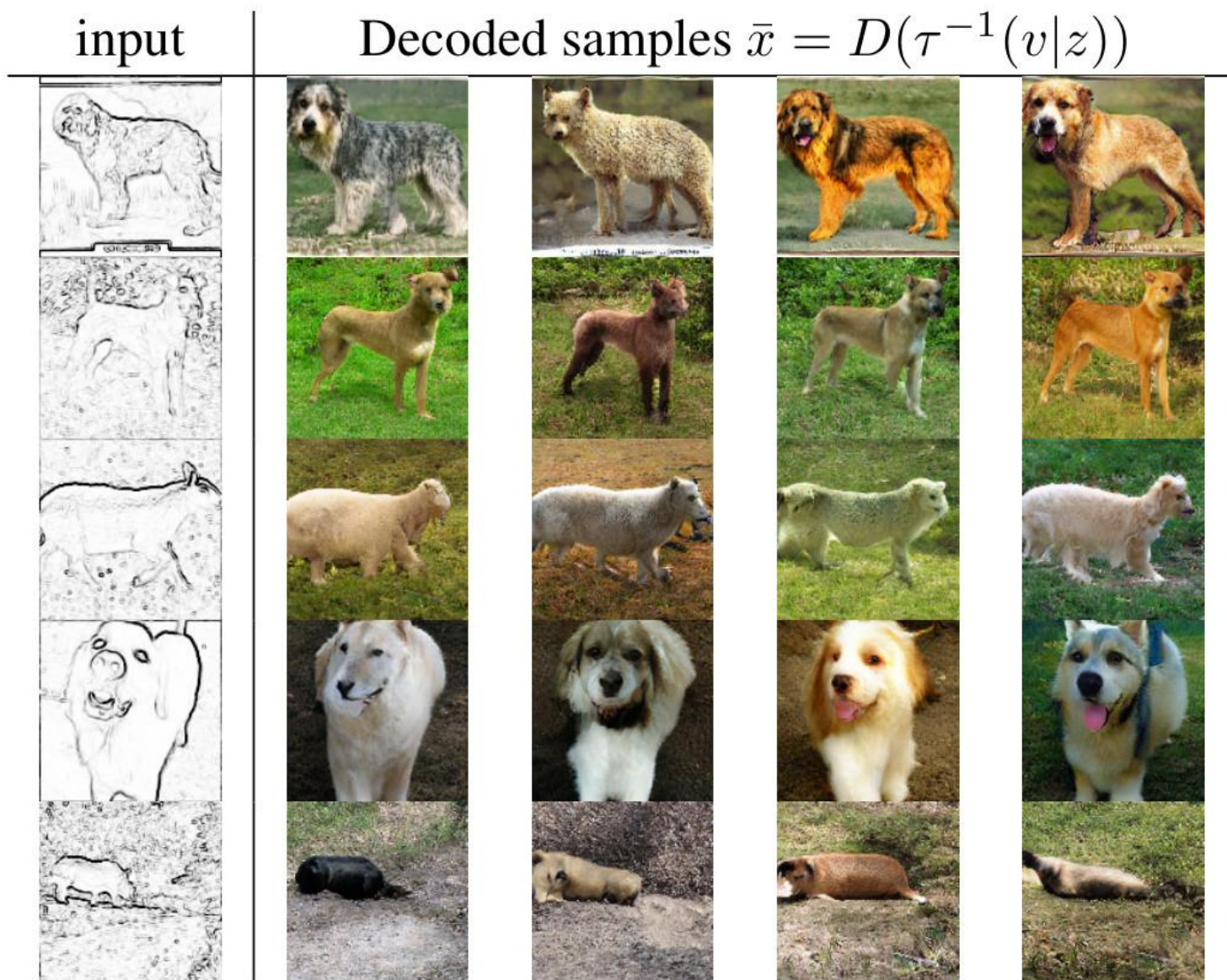
Data: Source-Target Pairs



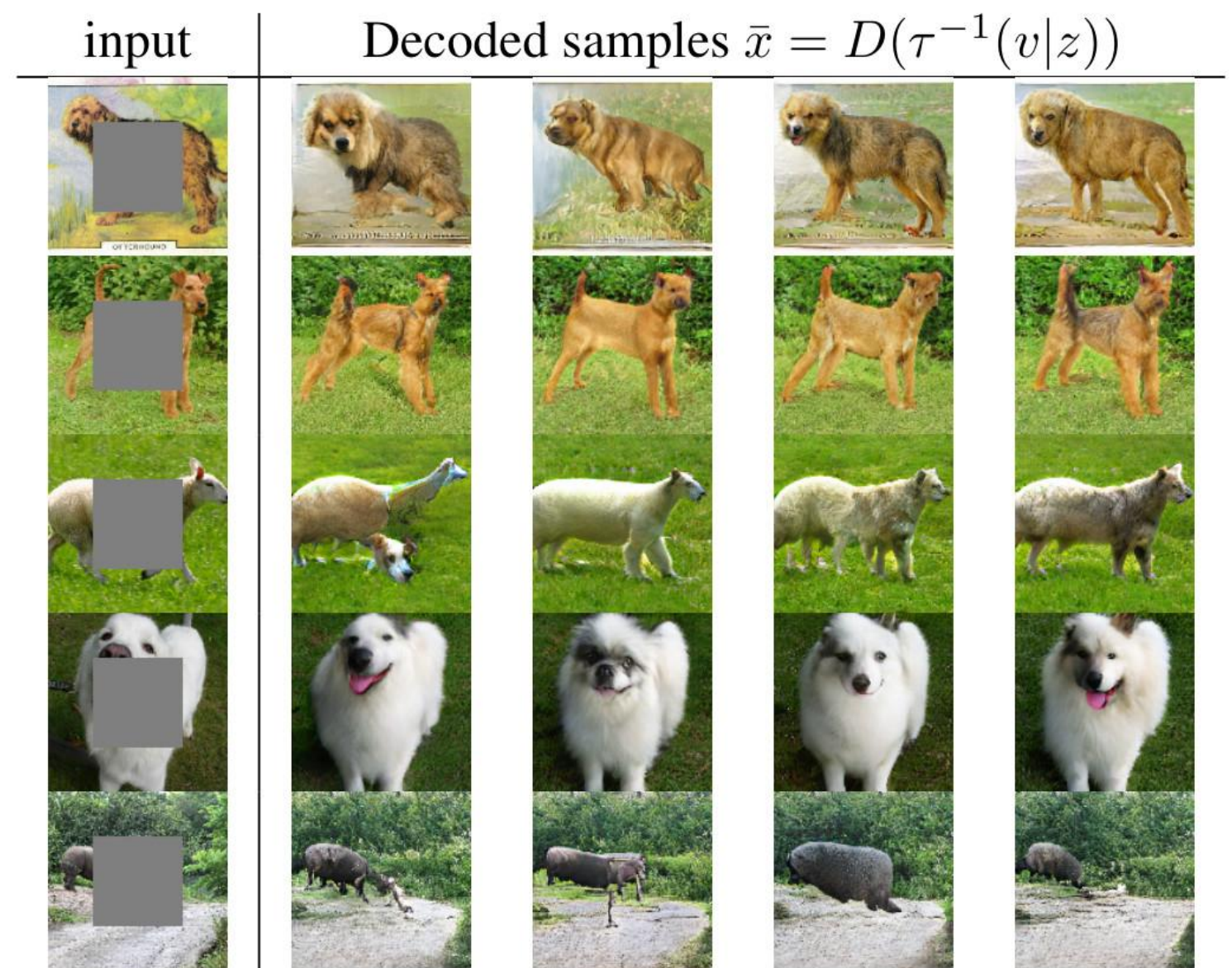
Results: Colorization and Inpainting with Net2Net

Use suitable experts for each task:

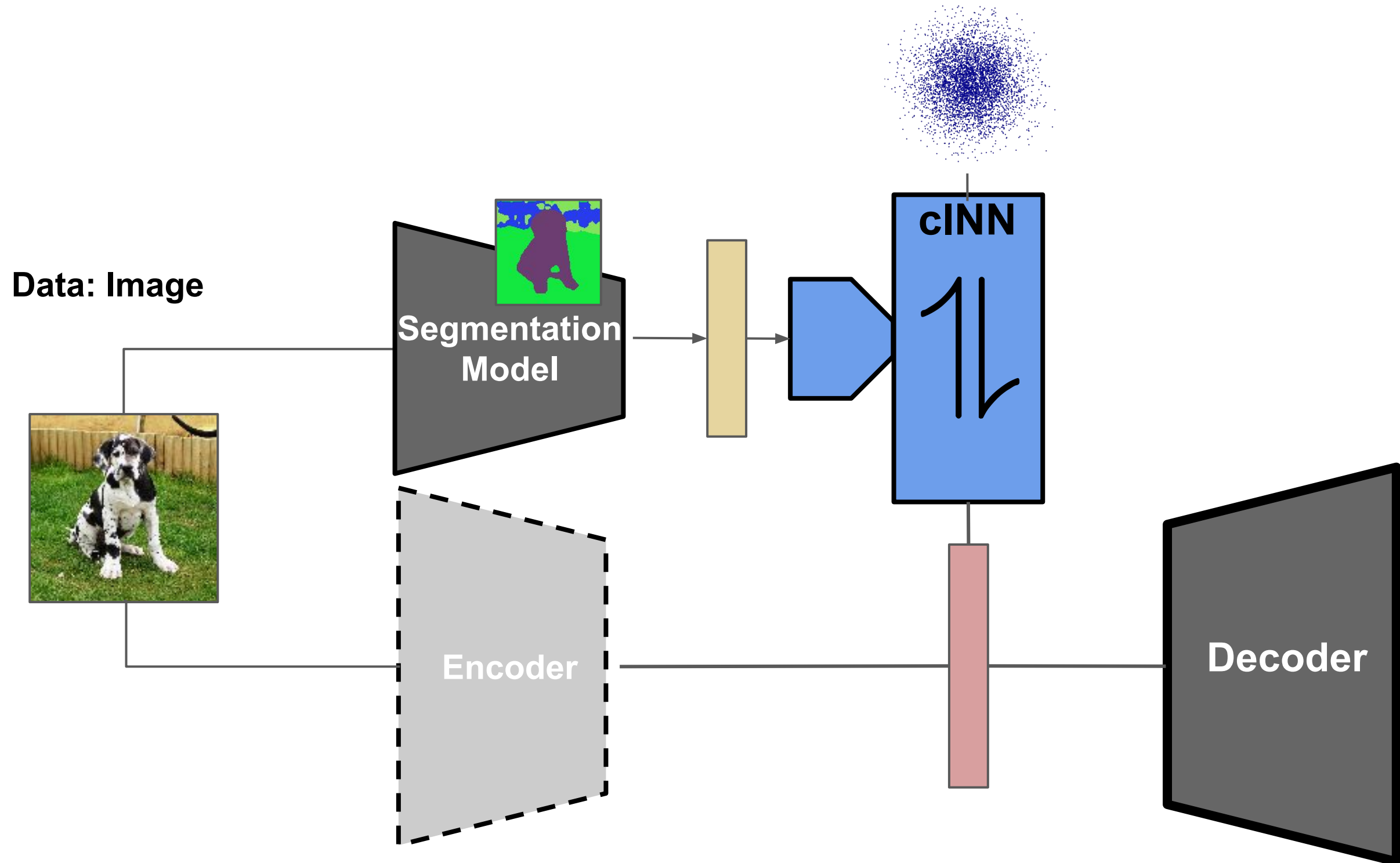
Stylized ResNet-50 for
Edge-Aware Edge-to-Image



Vanilla ResNet-50 for
Texture-Aware Inpainting

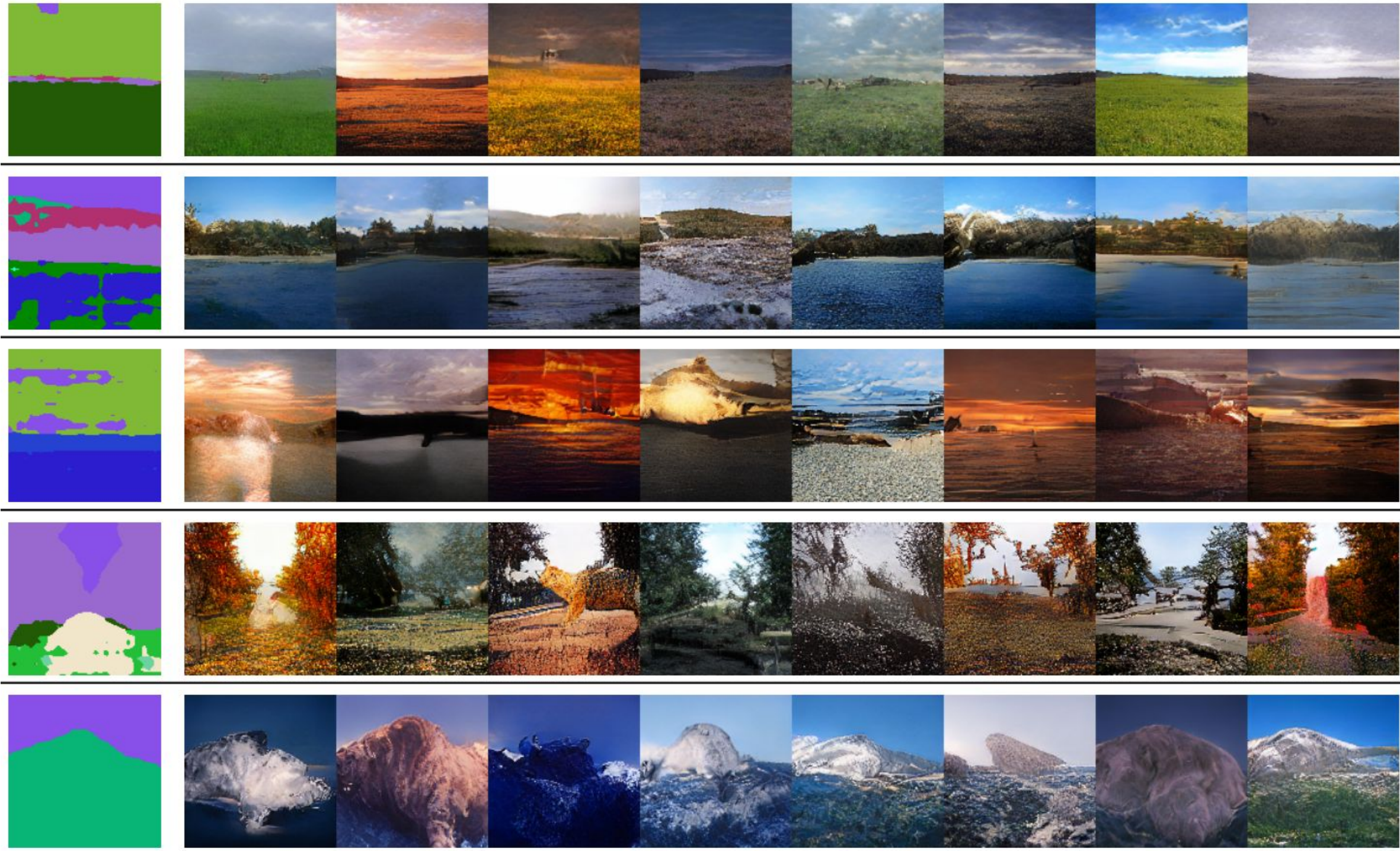


Experiment #4: Segmenter-to-Autoencoder



Results: Semantic Image Synthesis Synthesis with Net2Net

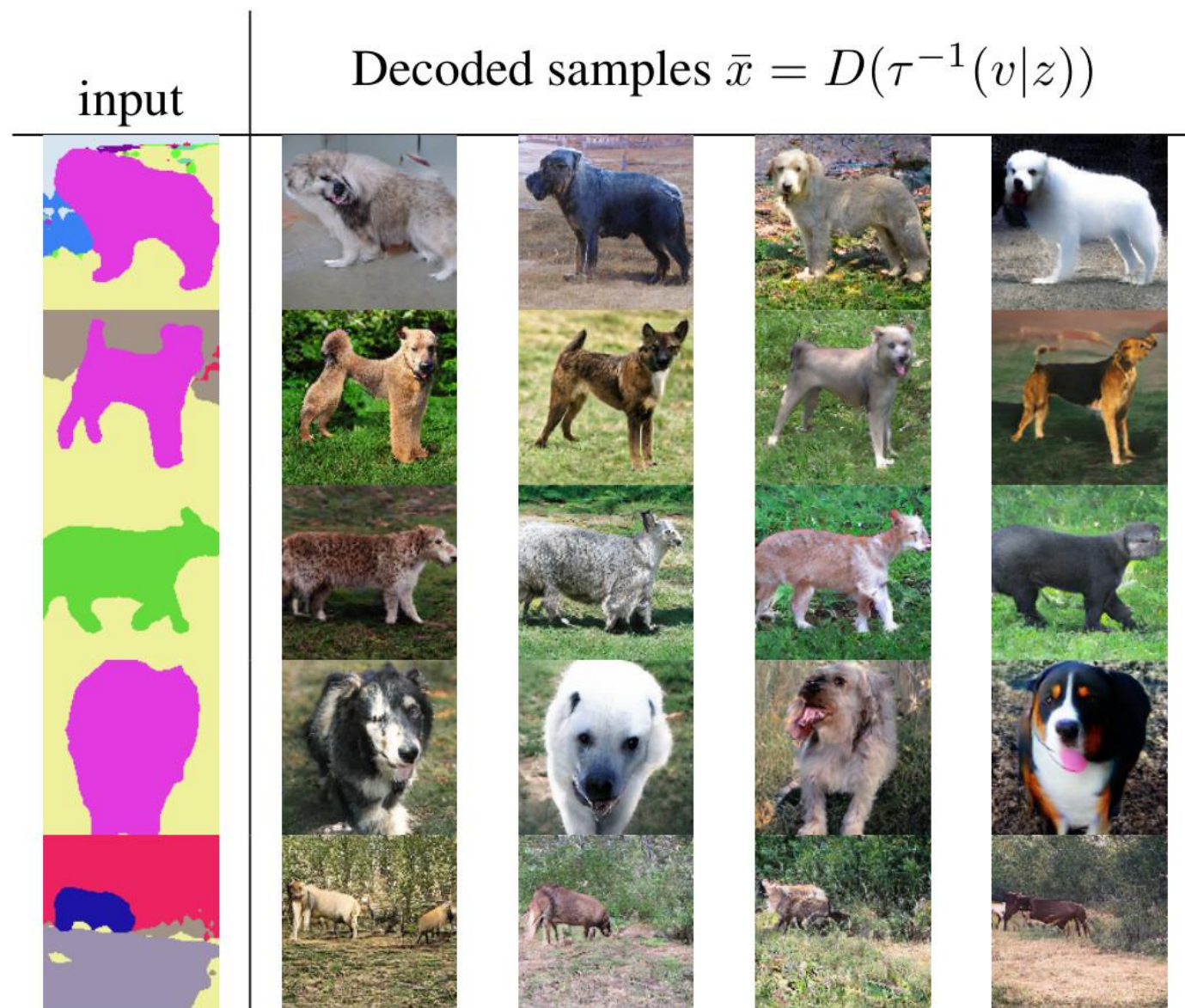
$\Phi(x)$ translating $\Phi(x)$ onto target domain of AE g with different samples $v \sim q(v)$



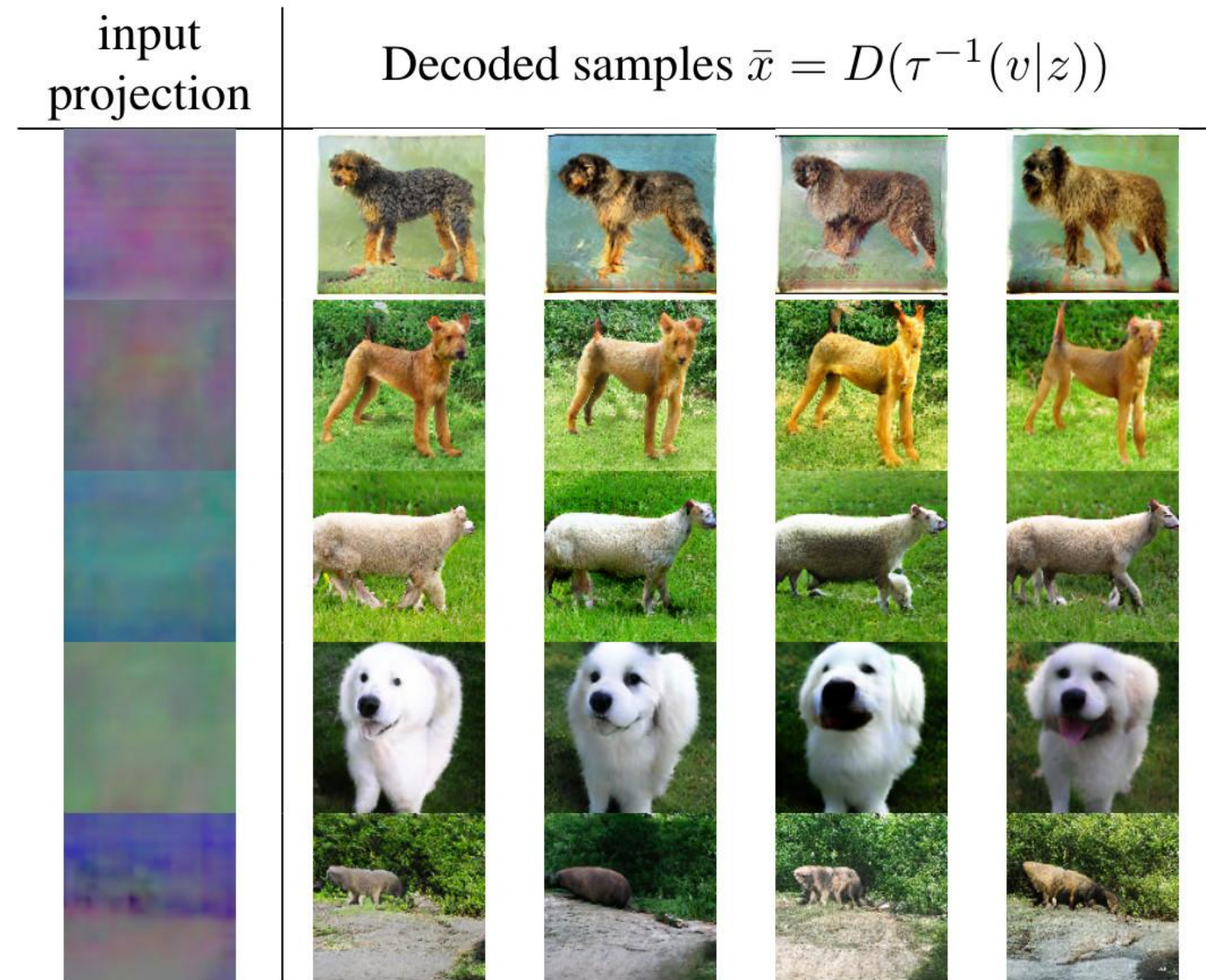
Results: Controlling Variability with Net2Net

Use suitable layers of experts to control variability:

Argmax of Segmentation Expert
for high variability



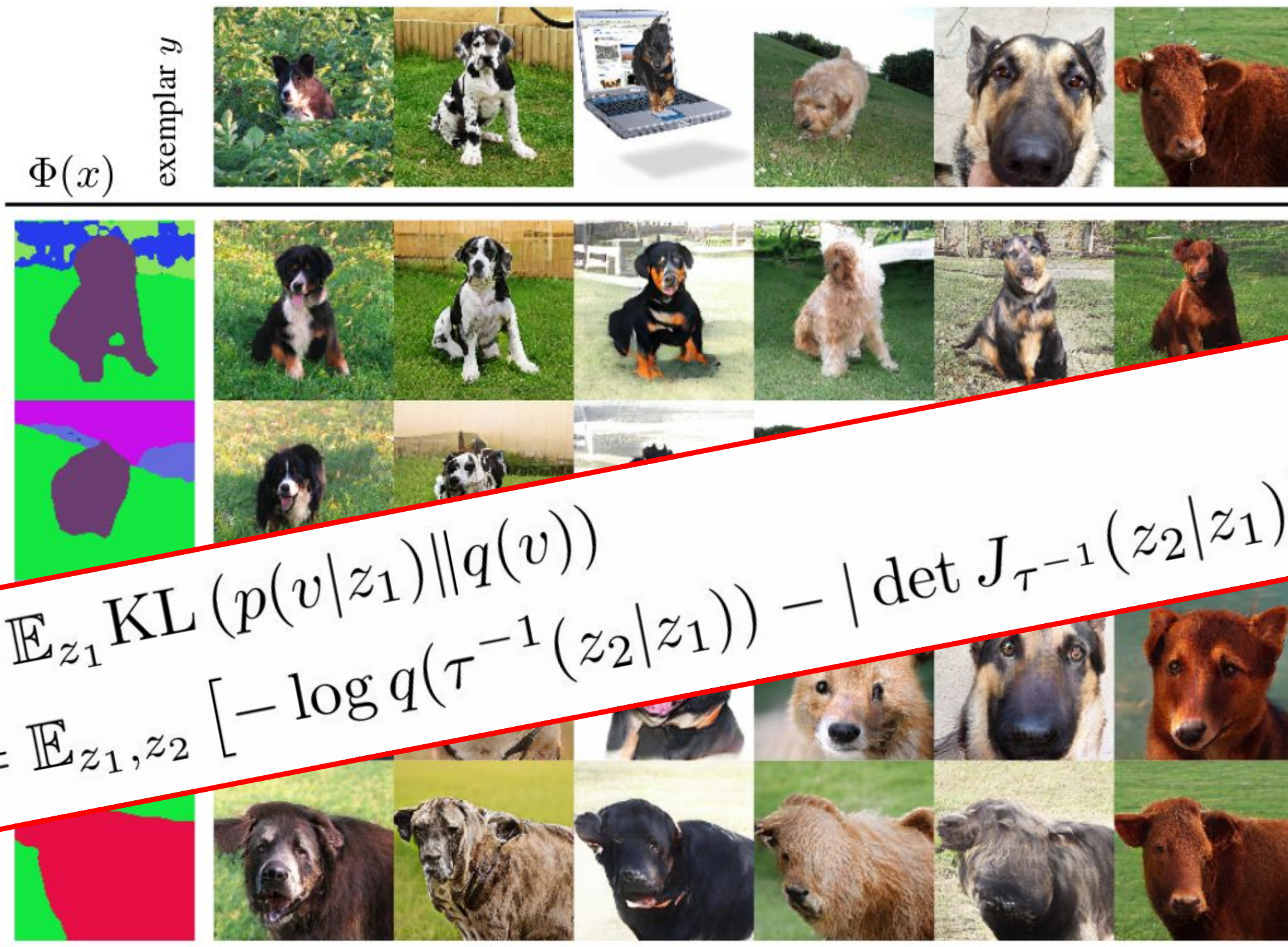
Logits of Segmentation Expert
for low variability



Results: Exemplar-Guided Synthesis with Net2Net



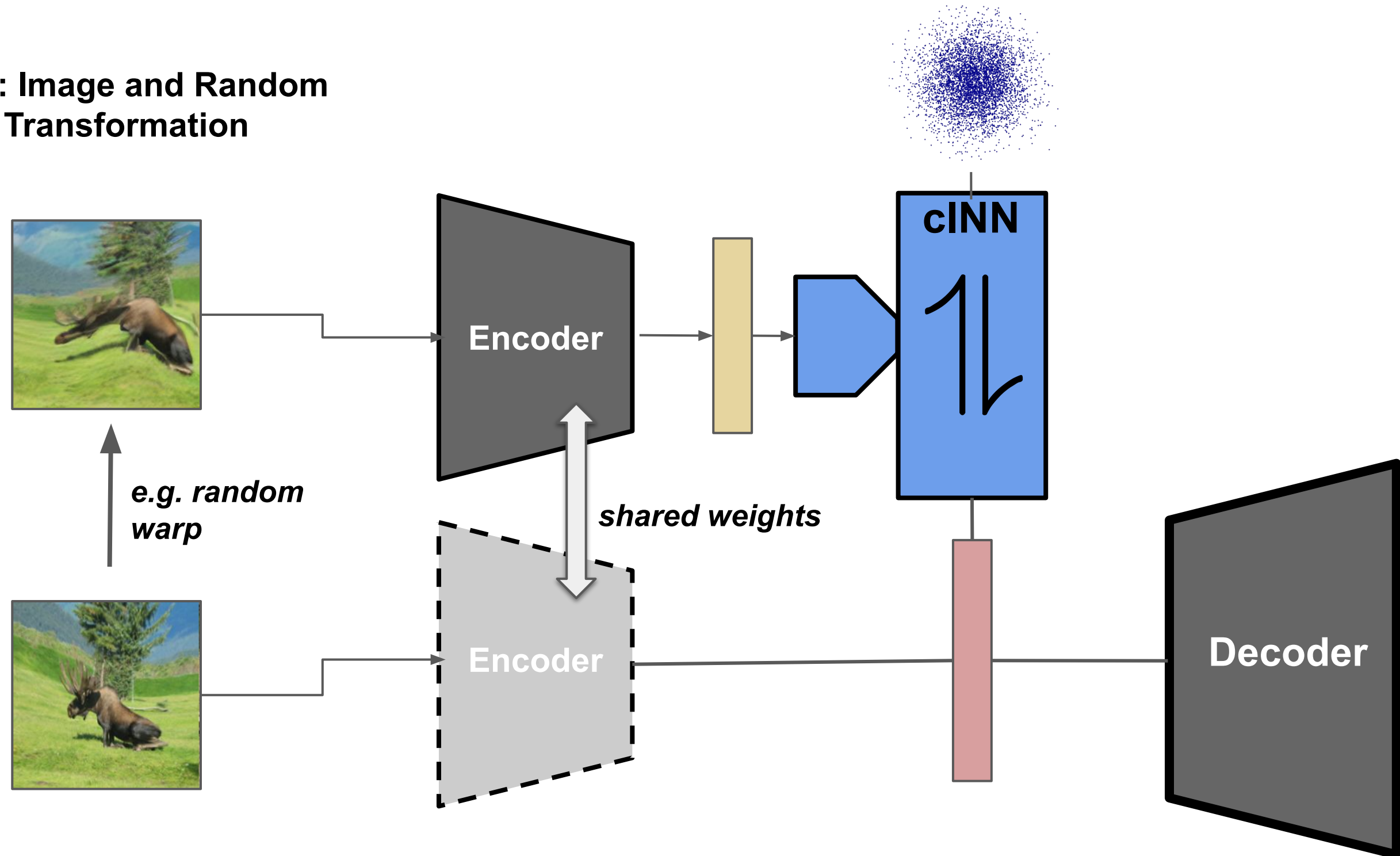
Results: Exemplar-Guided Synthesis with Net2Net



$$I(v, z_1) \leq \mathbb{E}_{z_1} \text{KL} (p(v|z_1) \| q(v)) \\ = \mathbb{E}_{z_1, z_2} \left[-\log q(\tau^{-1}(z_2|z_1)) - |\det J_{\tau^{-1}}(z_2|z_1)| \right] - \mathbb{H}[z_2|z_1]$$

Experiment #5: Unsupervised Disentangling

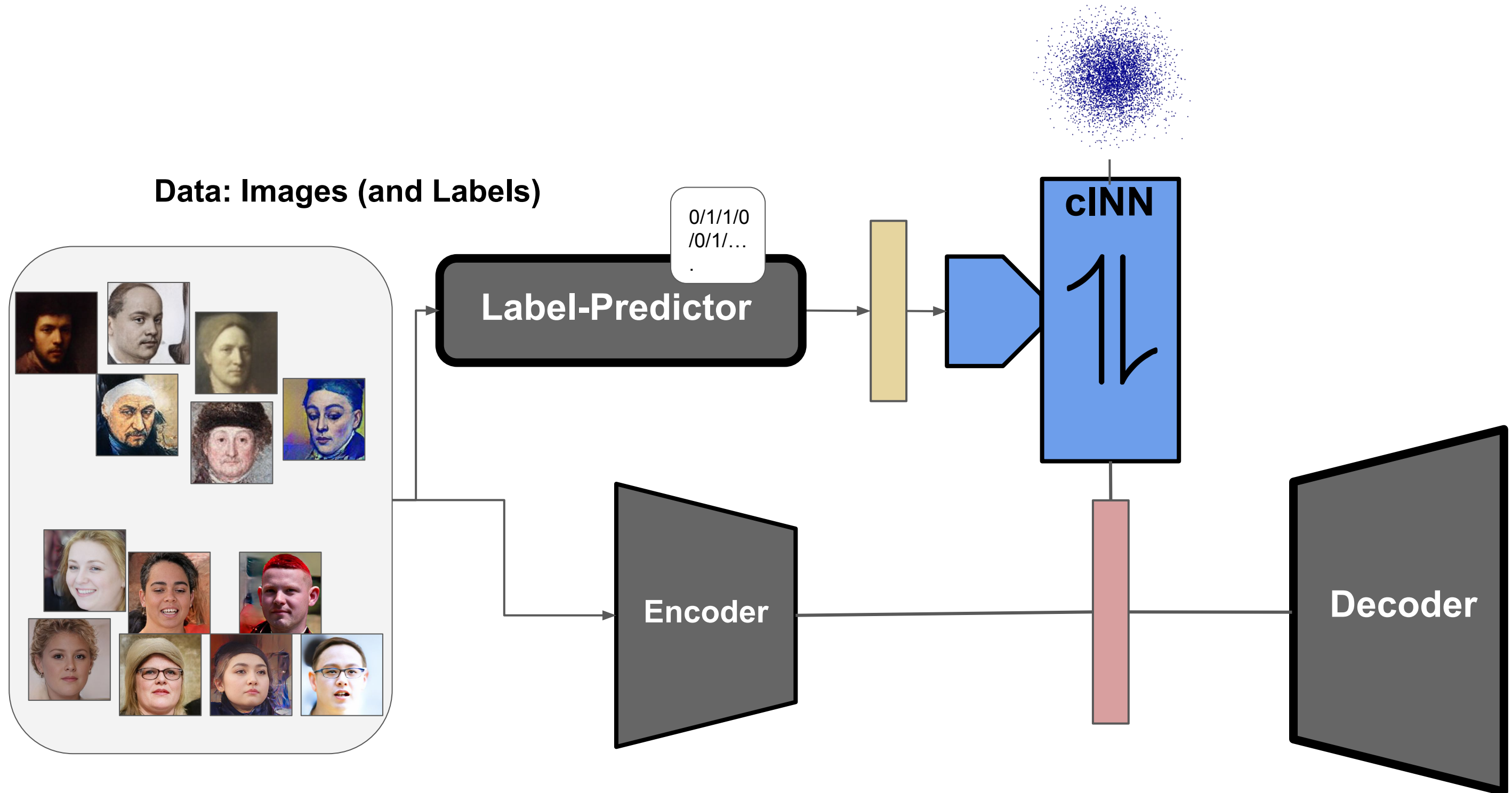
Data: Image and Random Transformation



Results: Unsupervised Disentangling of Shape and Appearance

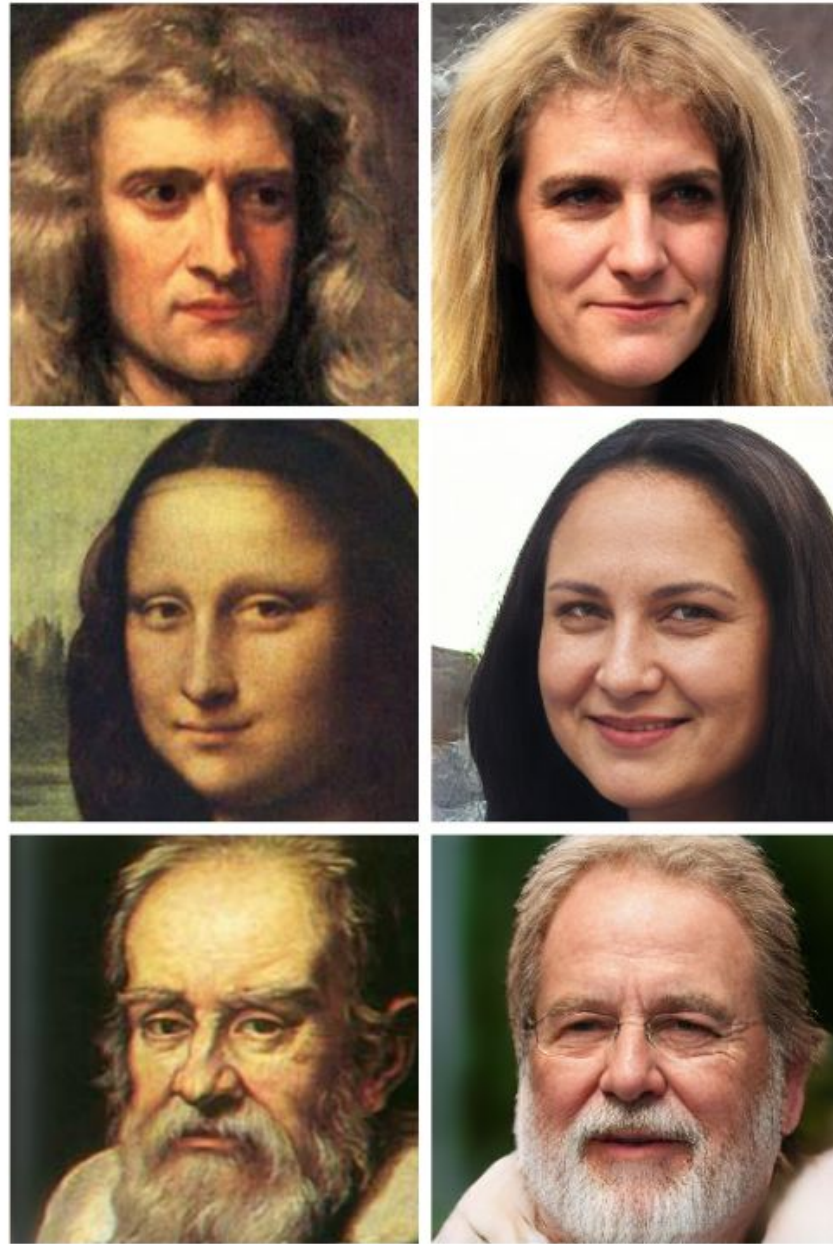


Experiment #6: Unpaired Domain Transfer



Results: Unpaired Domain Transfer with Net2Net

Oil-Portrait to Photography



Anime to Photography




















FFHQ to CelebA-HQ




















Broad Applicability

No gradients of experts required \Rightarrow Labels of human experts can be used

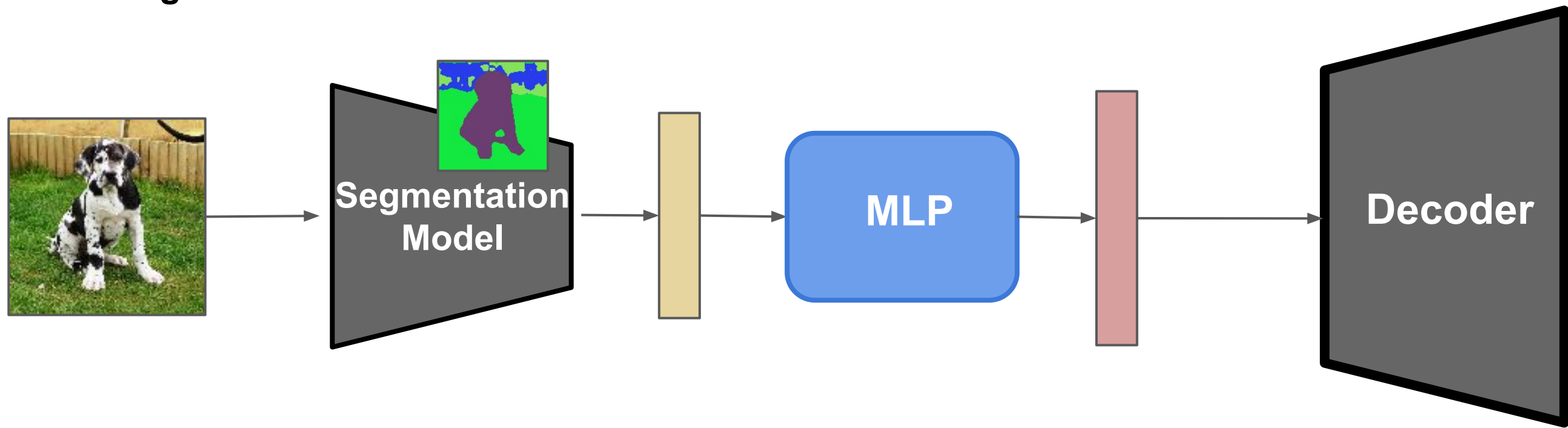
Attribute Modification

input	method	hair	glasses	gender
	our			
				
	our			
				
FID	our	15.18	37.32	16.38
		20.94	41.27	20.04

input	method	beard	age	smiling
	our			
				
	our			
				
FID	our	12.02	10.77	9.57
		19.88	21.77	14.47

Experiment #7: What if we use an MLP?













Data: Image



Experiment #7: What if we use an MLP?

What if we use a standard feedforward network instead?













→ Fails because the translation is not uniquely determined.

x	method	early layer: $\Lambda(z_{\Theta})$	middle layer: $\Lambda(z_{\Theta})$	last layer of f : $\Lambda(z_{\Theta})$
	our MLP			
	our MLP			
	our MLP			

Experiment #7: What if we use an MLP?

What if we use a standard feedforward network instead?













→ Fails because the translation is not uniquely determined.

x	method	early layer: $\Lambda(z_{\Theta})$	middle layer: $\Lambda(z_{\Theta})$	last layer of f : $\Lambda(z_{\Theta})$
	our			
	MLP			
	our			
	MLP			
	our			
	MLP			

Experiment #7: What if we use an MLP?

What if we use a standard feedforward network instead?

→ Fails because the translation is not uniquely determined.

x	method	early layer: $\Lambda(z_{\Theta})$	middle layer: $\Lambda(z_{\Theta})$	last layer of f : $\Lambda(z_{\Theta})$
	our MLP			
	our MLP			
	our MLP			

Try it yourself!

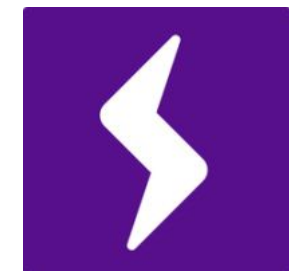


Check out our paper
and visit our github page

<https://github.com/CompVis/net2net>

to train net2net for your own models.

Thanks for your attention!



Model	Time [days]	Hardware	Energy [kWh]	Cost [EUR]	CO_2 [kg]
our cINN	≤ 1	1 NVIDIA Titan X	14.4	3.11	4.26
BigGAN [3]	15	8 NVIDIA V100	1260.0	272.16	372.96
FUNIT [40]	14	8 NVIDIA V100	1176.0	254.02	348.10
BERT [14]	10.3	8 NVIDIA V100	865.2	186.88	256.10