Δ-encoder: an effective sample synthesis method for few-shot object recognition

NIPS 2018

Eli Schwartz*[1,2], Leonid Karlinsky*[1],
Joseph Shtok[1], Sivan Harary[1], Mattias Marder[1], Abhishek Kumar[1],
Rogerio Feris[1], Raja Giryes[2] and Alex M. Bronstein[3]

[1]IBM Research AI
[2]School of Electrical Engineering, Tel-Aviv University, Tel-Aviv, Israel
[3]Department of Computer Science, Technion, Haifa, Israel

目录 OUTLINE

01

Few-shot learning

Test on unseen dataset

Table 3: Summary of the datasets used in our experiments

Dataset	Fine grained	Image size	Total # images	Seen classes	Unseen classes
*mini*ImageNet [42]	✗	Medium	$60K$	80	20
CIFAR-100 [22]	✗	Small	$60K$	80	20
Caltech-256 Object Category [15]	✗	Large	$30K$	156	50
Caltech-UCSD Birds 200 (CUB) [47]	✓	Large	$12K$	150	50
Attribute Pascal & Yahoo (aPY) [9]	✗	Large	$14K$	20	12
Scene UNderstanding (SUN) [49]	✓	Large	$14K$	645	72
Animals with Attributes 2 (AWA2) [48]	✗	Large	$37K$	40	10

With many instances

Only k instances for k-shot

Few-shot learning

Few-shot learning by metric learning

Learn an embedding into a metric space where some simple (usually L2) metric is then used to classify instances of new categories

Few-shot meta-learning (learning-to-learn)

The meta-learned classifiers are optimized to be easily fine-tuned on new few-shot tasks using the provided small training data.

Generative and augmentation-based few-shot approaches

Either generative models are trained to synthesize new data based on few examples, or additional examples are obtained by some other form of transfer learning from external data.

The Framework

The framework

Intuition

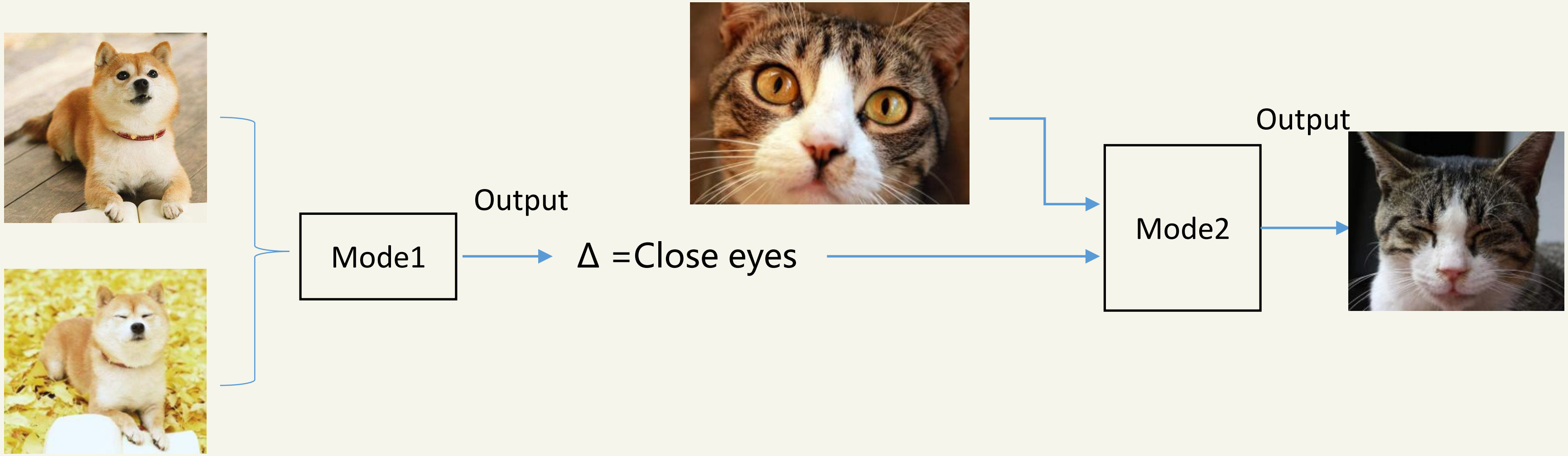
The offset in feature space between a pair of same-class examples conveys information on a valid deformation. The deformation called “Δ”

Δ =Close eyes

Method

A model to capture the deformation between examples in the same class.

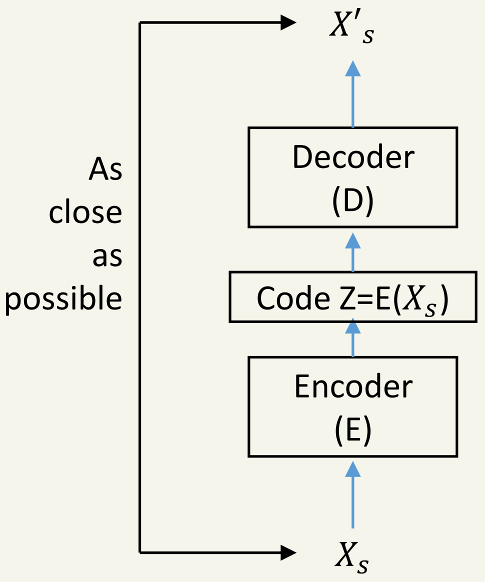
A model use the deformation AND an example X to generate a new example Y.
The expected results is : The generated example Y is a deformation of the given example X.



02

The framework

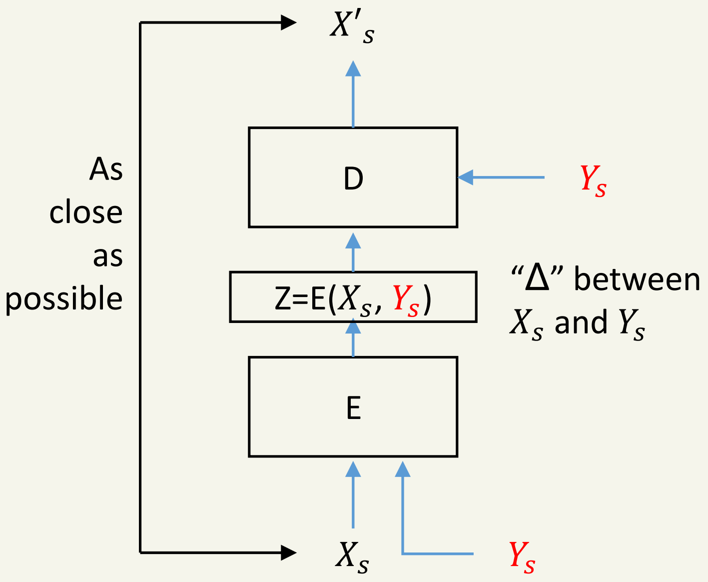
Auto-encoder



Problem:

Z has the information of X. How can you guarantee the reconstructed X'_s is relied on Y?

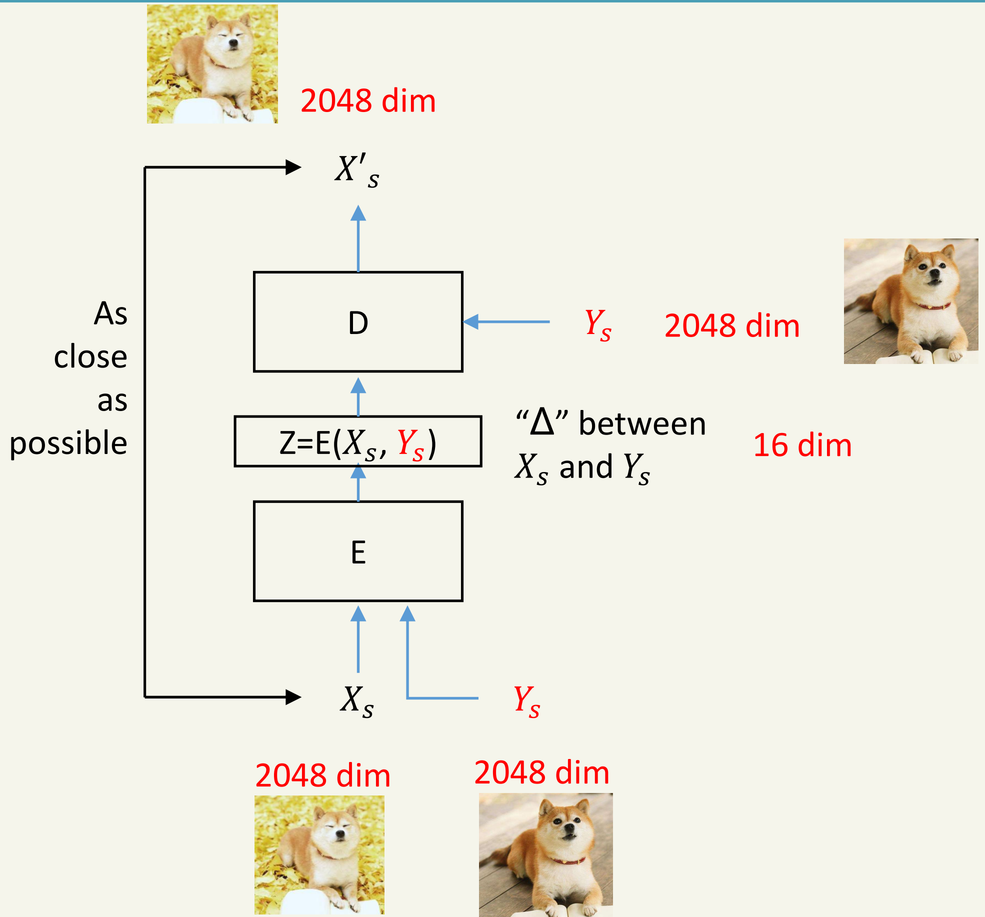
Δ-encoder



Training phase

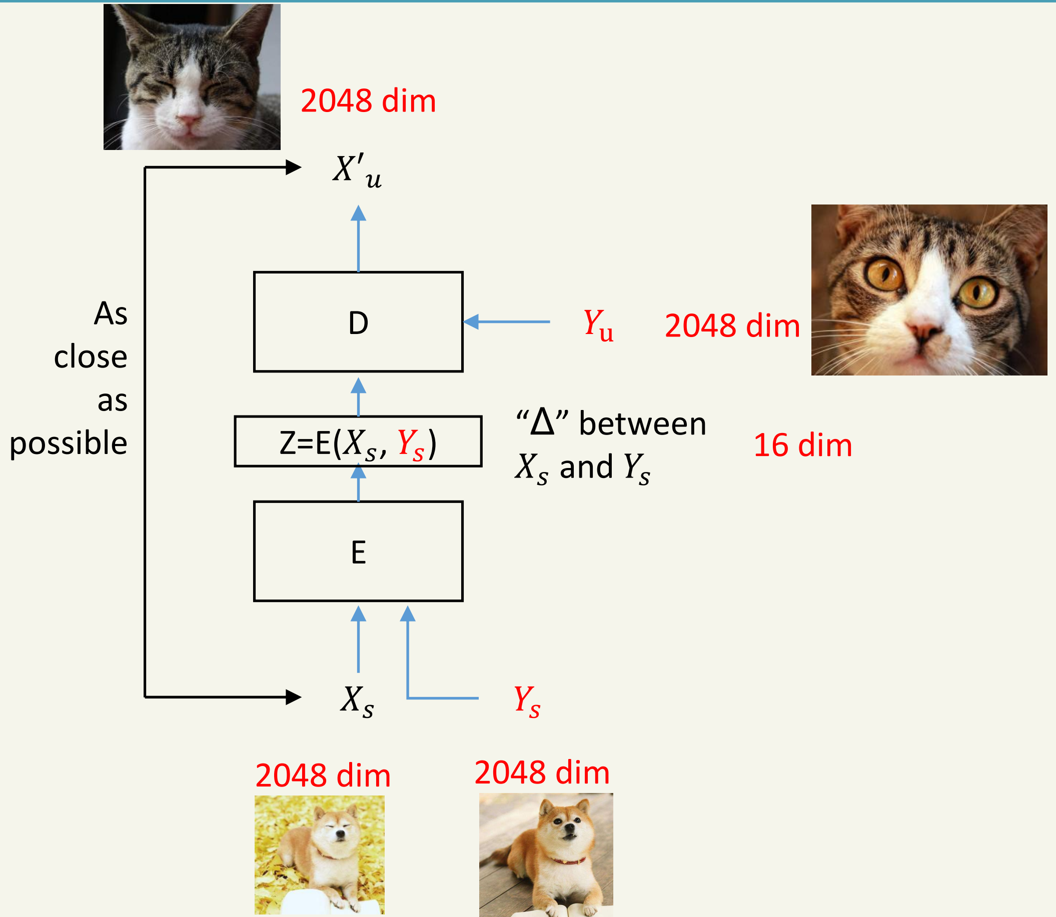
Keeping the dimension of Z small, we ensure that the decoder D cannot use just Z in order to reconstruct X.

This way, we regularize the encoder to strongly rely on the anchor example Y for the reconstruction



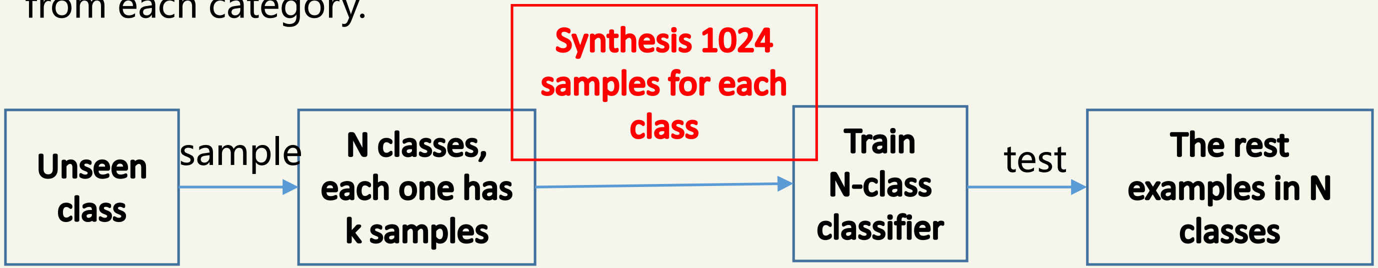
Sample synthesis

Sample W pairs of X_s and Y_s and one Y_u to generate W X'_u



Testing

N-way *k*-shot : draw *N* random unseen categories, and draw *k* random samples from each category.



we use our trained network to synthesize a total of 1024 samples per category based on those k examples. This is followed by training a simple linear N-class classifier over those 1024 · N samples, and finally, the calculation of the few-shot classification accuracy on a set of M real (query) samples from the tested N categories.

Experiment

Experiment

Table 1: 1-shot/5-shot 5-way accuracy results

Method	*mini*ImageNet	CIFAR100	Caltech-256	CUB	Average
Nearest neighbor (baseline)	44.1 / 55.1	56.1 / 68.3	51.3 / 67.5	52.4 / 66.0	51.0 / 64.2
MACO [19]	41.1 / 58.3	-	-	60.8 / 75.0	-
Meta-Learner LSTM [33]	43.4 / 60.6	-	-	40.4 / 49.7	-
Matching Nets [42]	46.6 / 60.0	50.5 / 60.3	48.1 / 57.5	49.3 / 59.3	48.6 / 59.3
MAML [10]	48.7 / 63.1	49.3 / 58.3	45.6 / 54.6	38.4 / 59.1	45.5 / 58.8
Prototypical Networks [38]	49.4 / 68.2	-	-	-	-
SRPN [29]	55.2 / 69.6	-	-	-	-
RELATION NET [40]	57.0 / 71.1	-	-	-	-
DEML+Meta-SGD [50]	58.5 / 71.3 $^{\diamond}$	61.6 / 77.9 $^{\diamond}$	62.2 / 79.5 $^{\diamond}$	66.9 / 77.1 $^{\diamond}$	62.3 / 76.4
Dual TriNet [4]	58.1 / **76.9** †	63.4 / 78.4 †	63.8 / 80.5 †	69.6 / **84.1** *	63.7 / 80.0
Δ-encoder	**58.7** / 73.6	**65.9 / 80.1**	**63.9 / 84.7**	**69.9** / 82.5	**64.6 / 80.2**

$\diamond$ Model also trained on an external large-scale dataset
$\dagger$ Using word embedding trained on large corpus and applied to the label name
$\star$ Using human annotated class attributes

Feature extractor backbone only trained on the subset of training categories of the target dataset

03

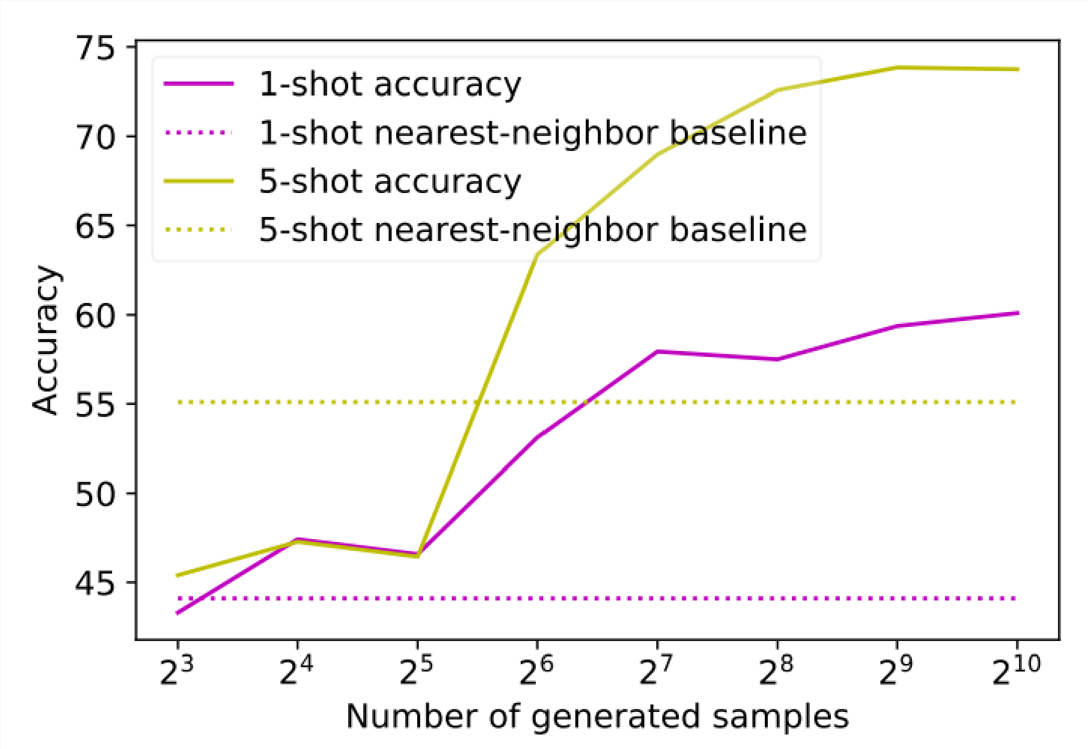
Experiment

Table 2: 1-shot/5-shot 5-way accuracy with ImageNet model features (trained on disjoint categories)

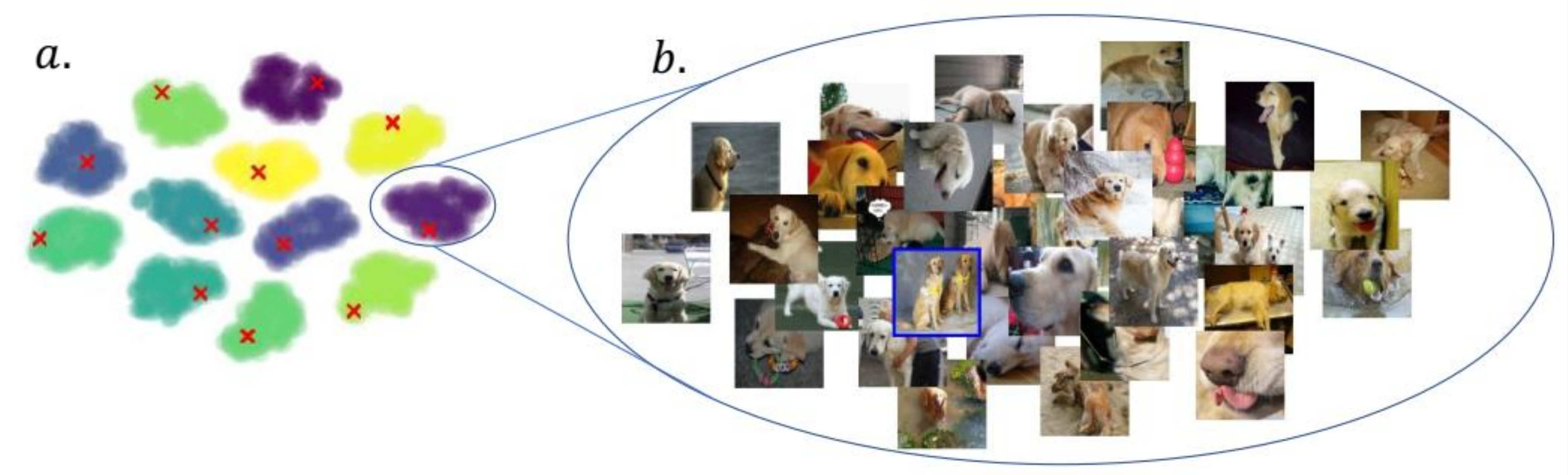
Method	AWA2	APY	SUN	CUB
Nearest neighbor (baseline)	65.9 / 84.2	57.9 / 76.4	72.7 / 86.7	58.7 / 80.2
Prototypical Networks	80.8 / 95.3	69.8 / 90.1	74.7 / **94.8**	71.9 / 92.4
Δ-encoder	**90.5 / 96.4**	**82.5 / 93.4**	**82.0** / 93.0	**82.2 / 92.6**

Feature extractor backbone is a VGG16 backbone pre-trained on ImageNet. The unseen test categories were verified to be disjoint from the ImageNet categories

Are we synthesizing non-trivial samples?

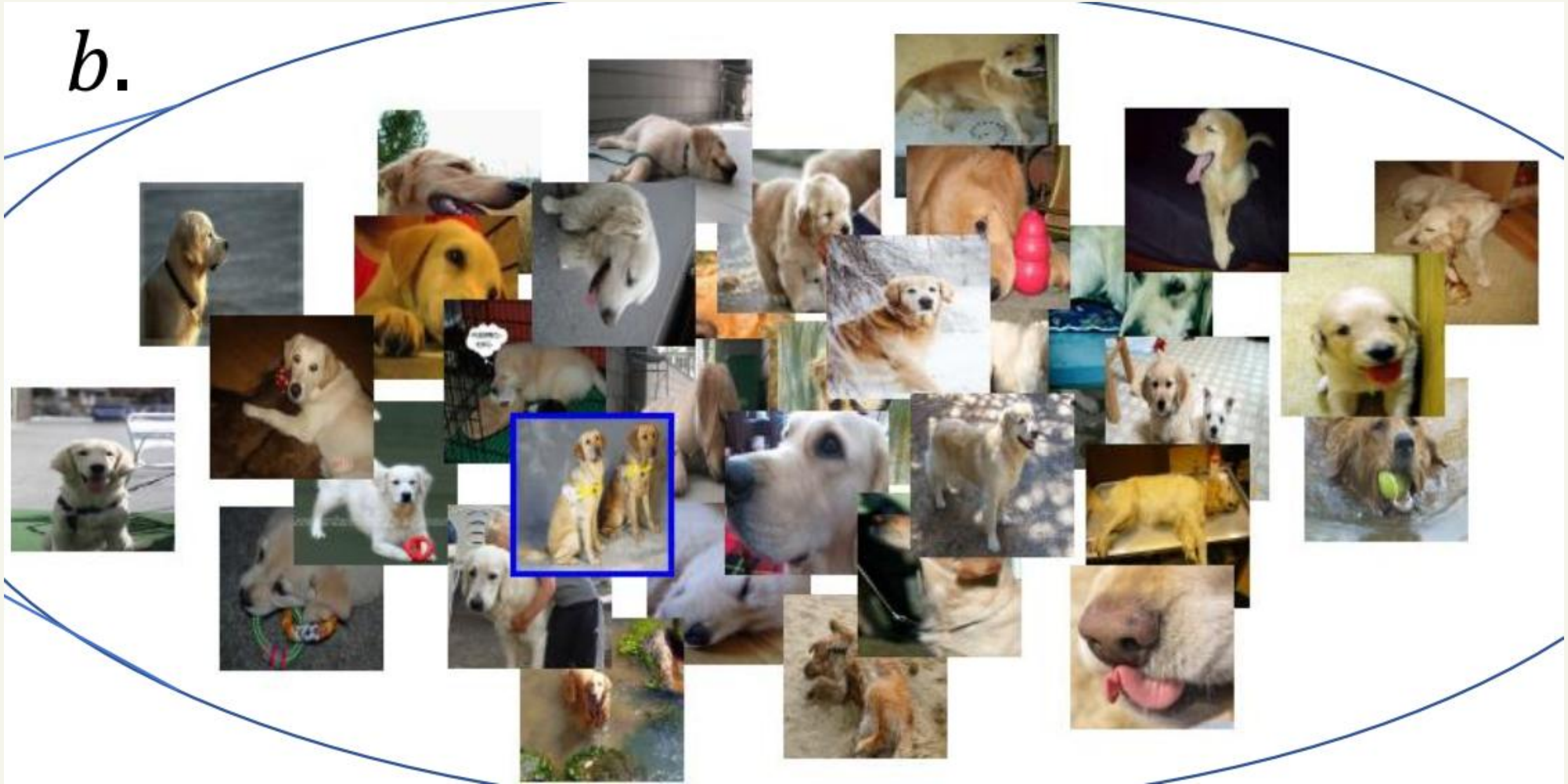


Experiment



Generated samples for 12-way one-shot. The two-dimensional embedding was produced by t-SNE. Best viewed in color.

Are we synthesizing non-trivial samples?



THANKS