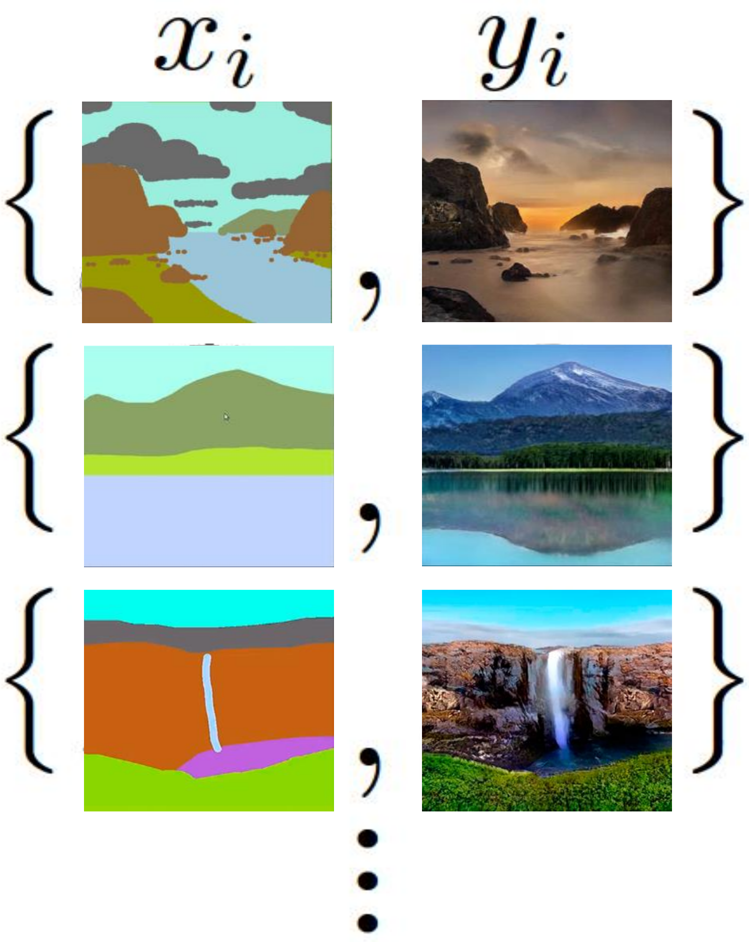
Few-Shot Unsupervised Image-to-Image Translation

Ming-Yu Liu

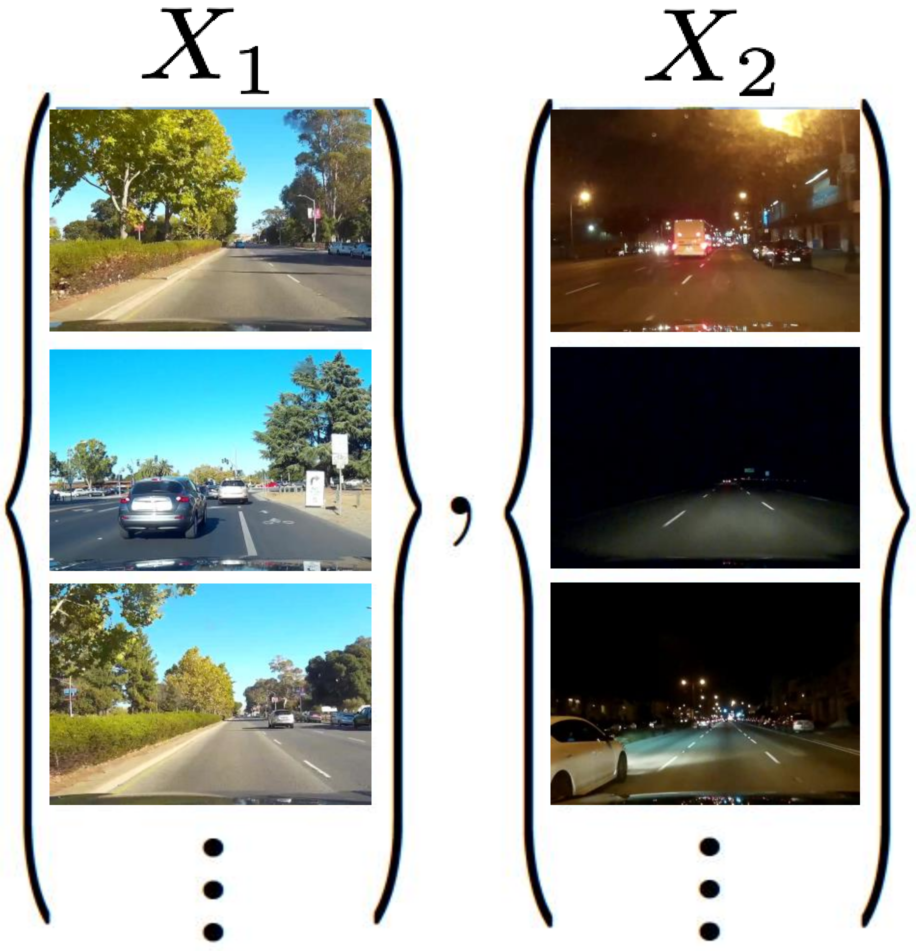
NVIDIA

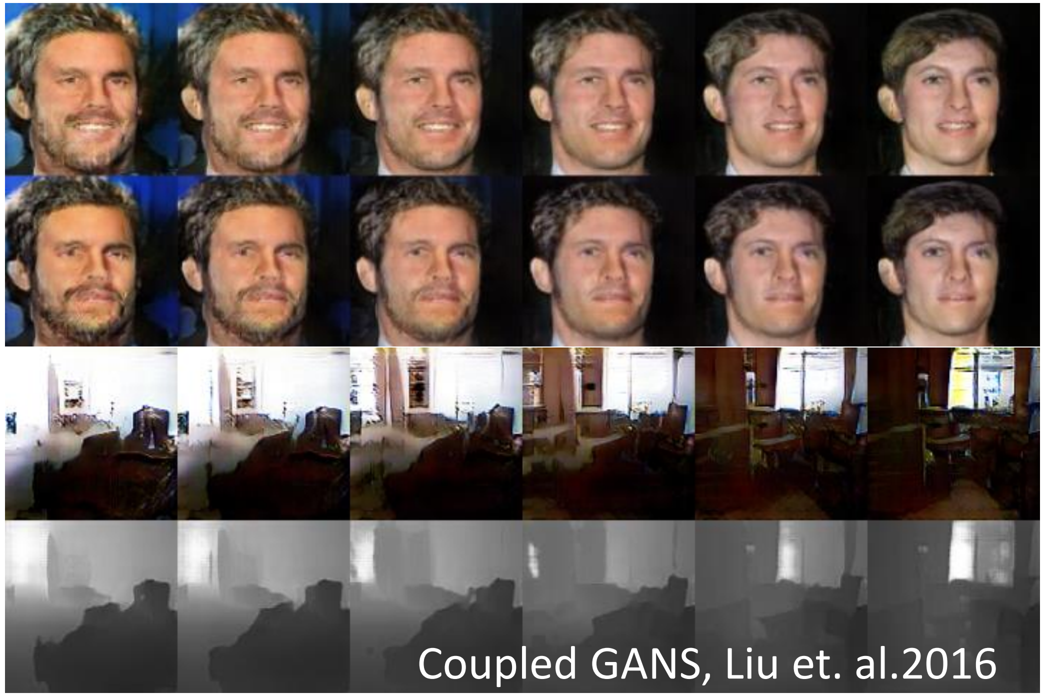
Supervised vs Unsupervised

Supervised/Paired/Aligned/Registered



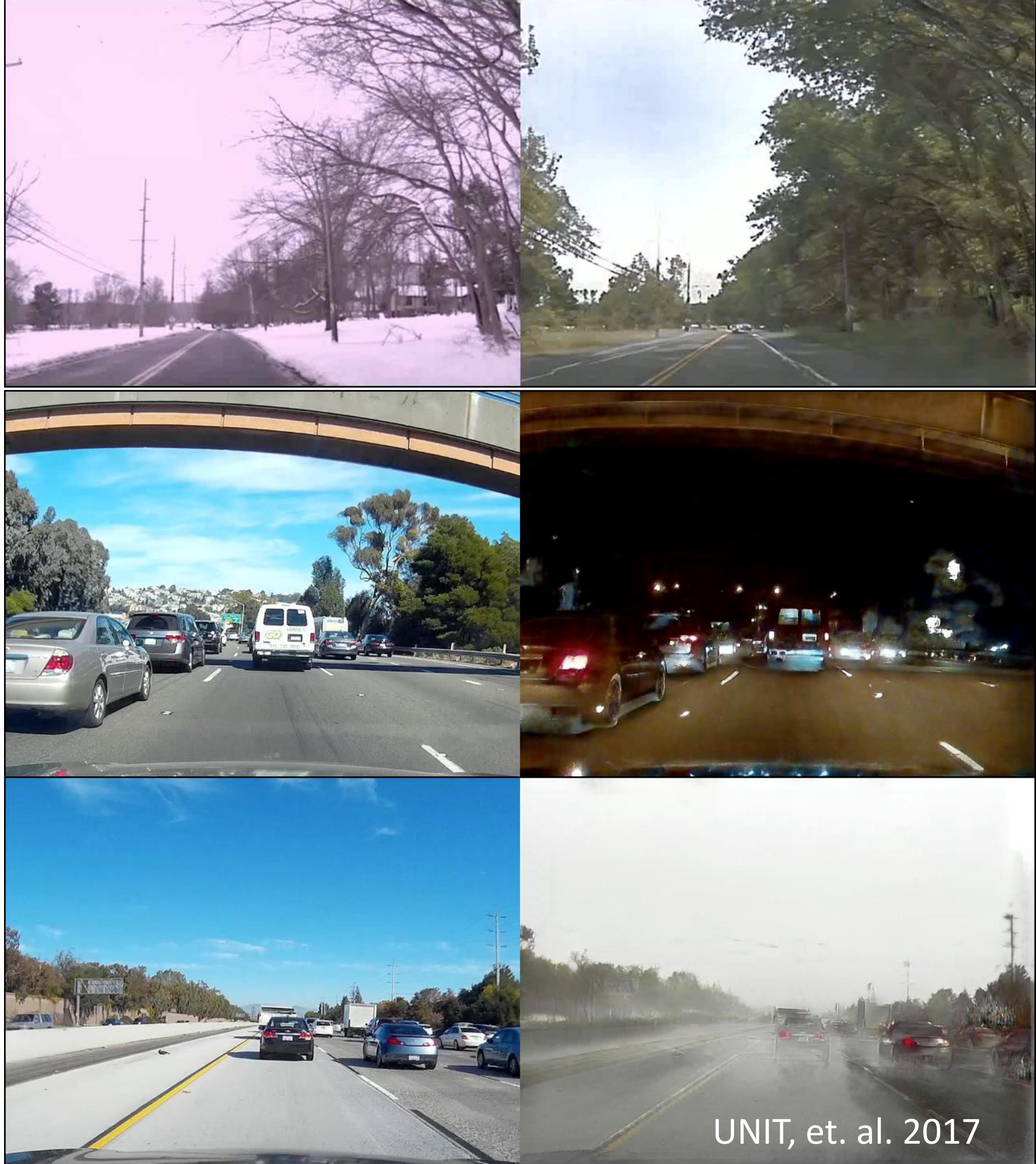
Unsupervised/Unpaired/Unaligned/Unregistered





Coupled GANS, Liu et. al.2016

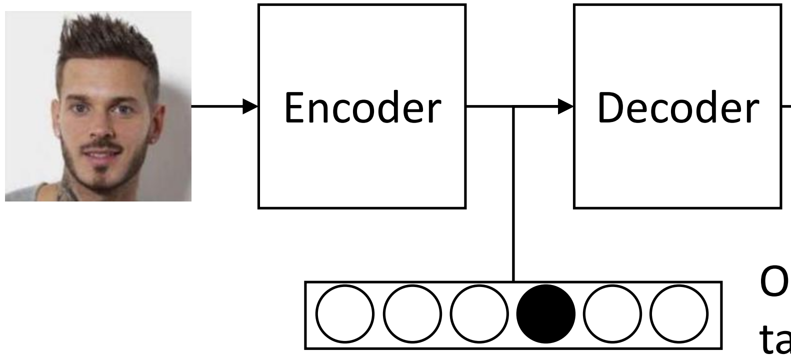
MUNIT, Huang et. al. 2018



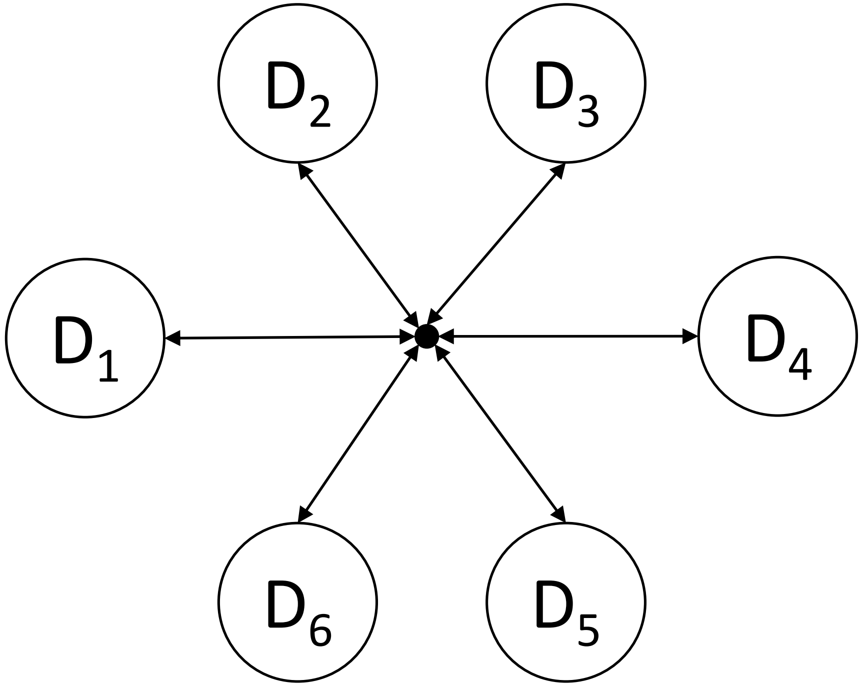
UNIT, et. al. 2017

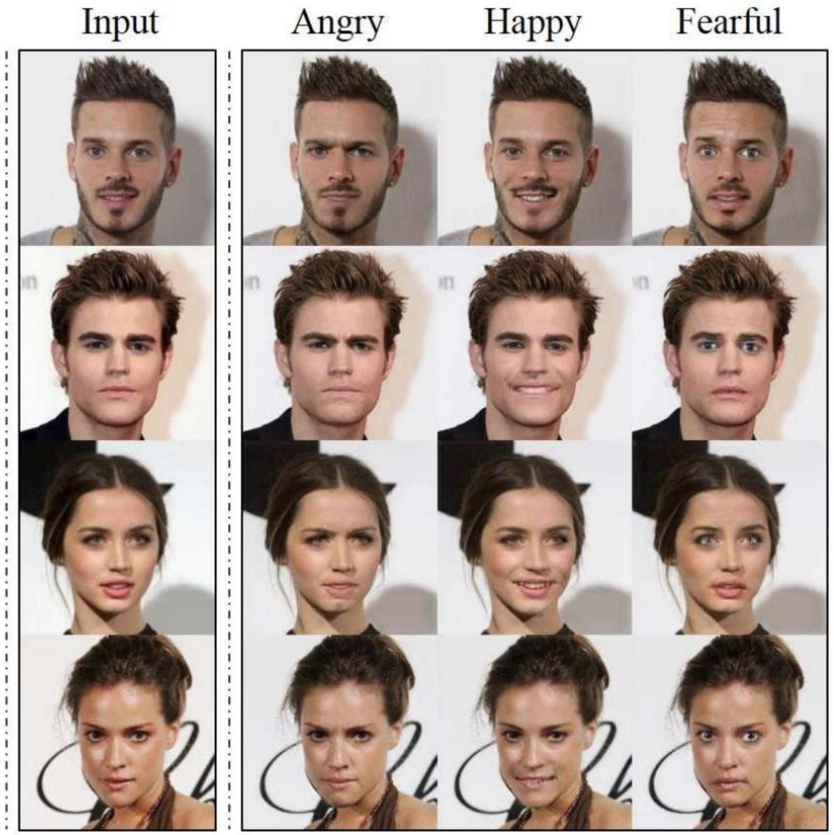
Multiple Domains

- StarGAN (Choi et. al. 2018)
- ComboGAN (Anoosheh et. al. 2017)
- Domain-Bank (Hui et. al. 2017)

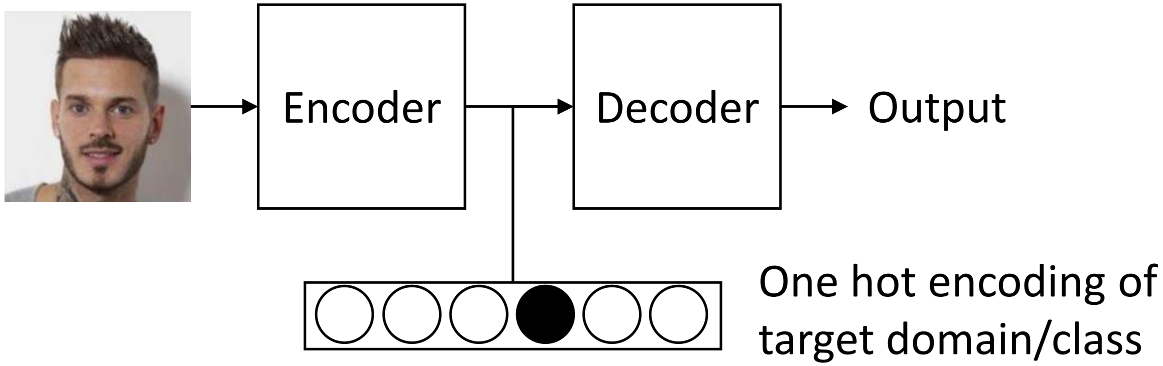


One hot encoding of
target domain/class

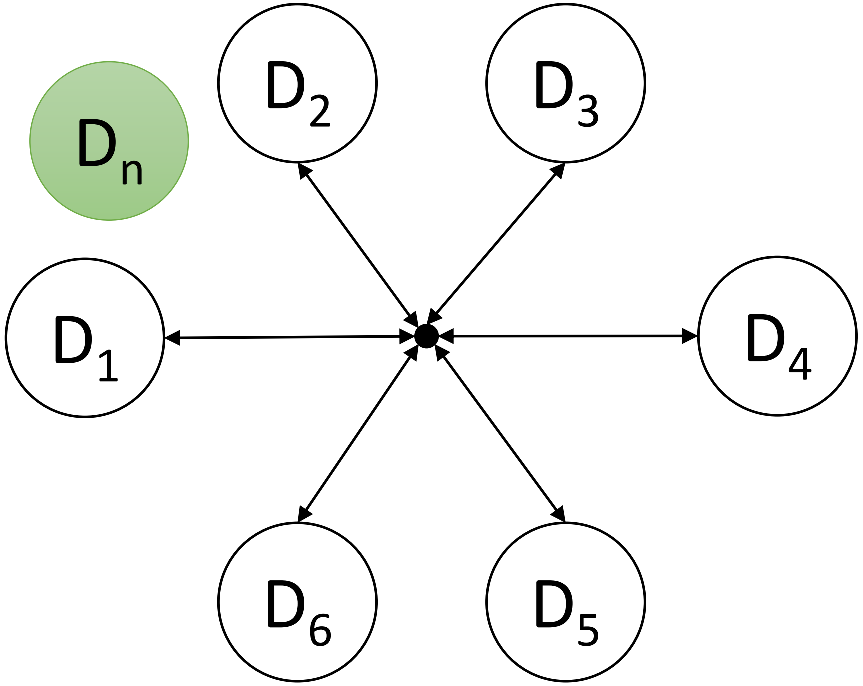


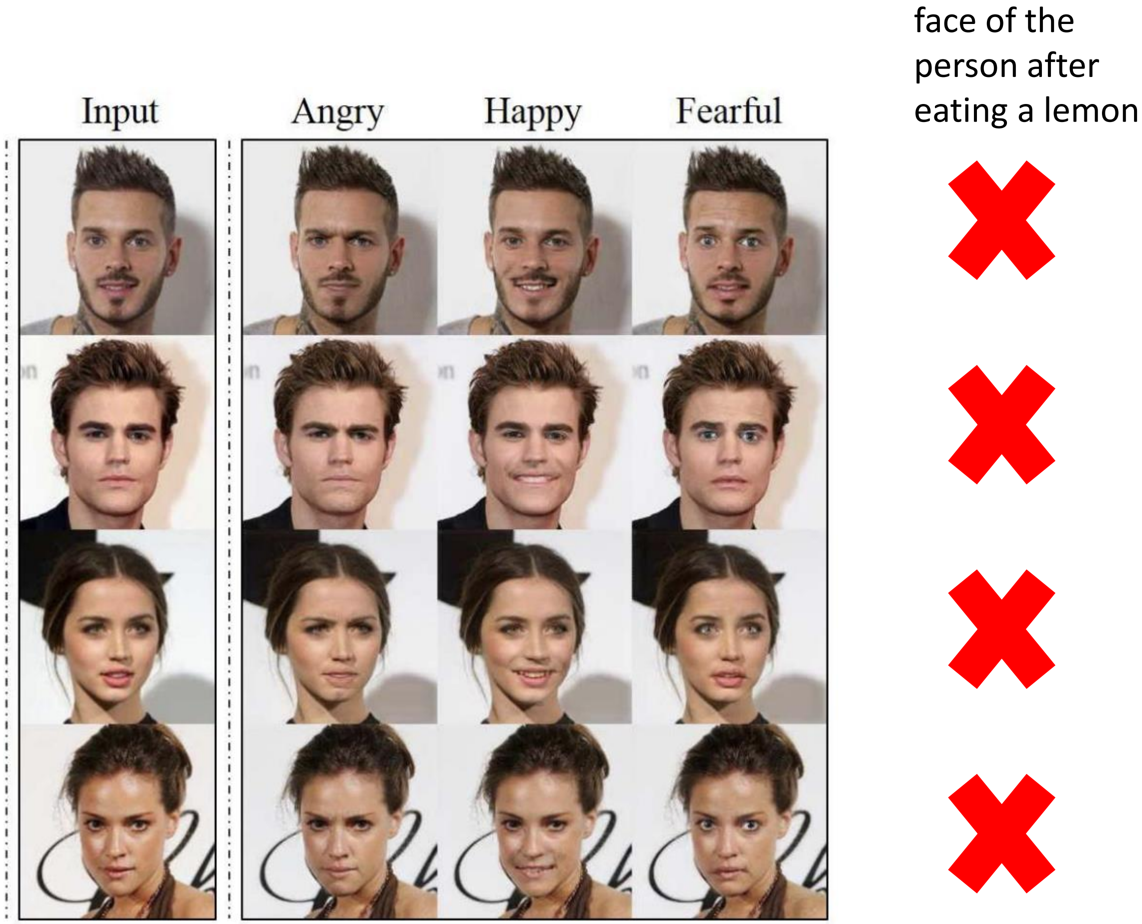


How About Unseen Domain



New class: face of the person after eating a lemon





Few-Shot and **Unseen Domain**

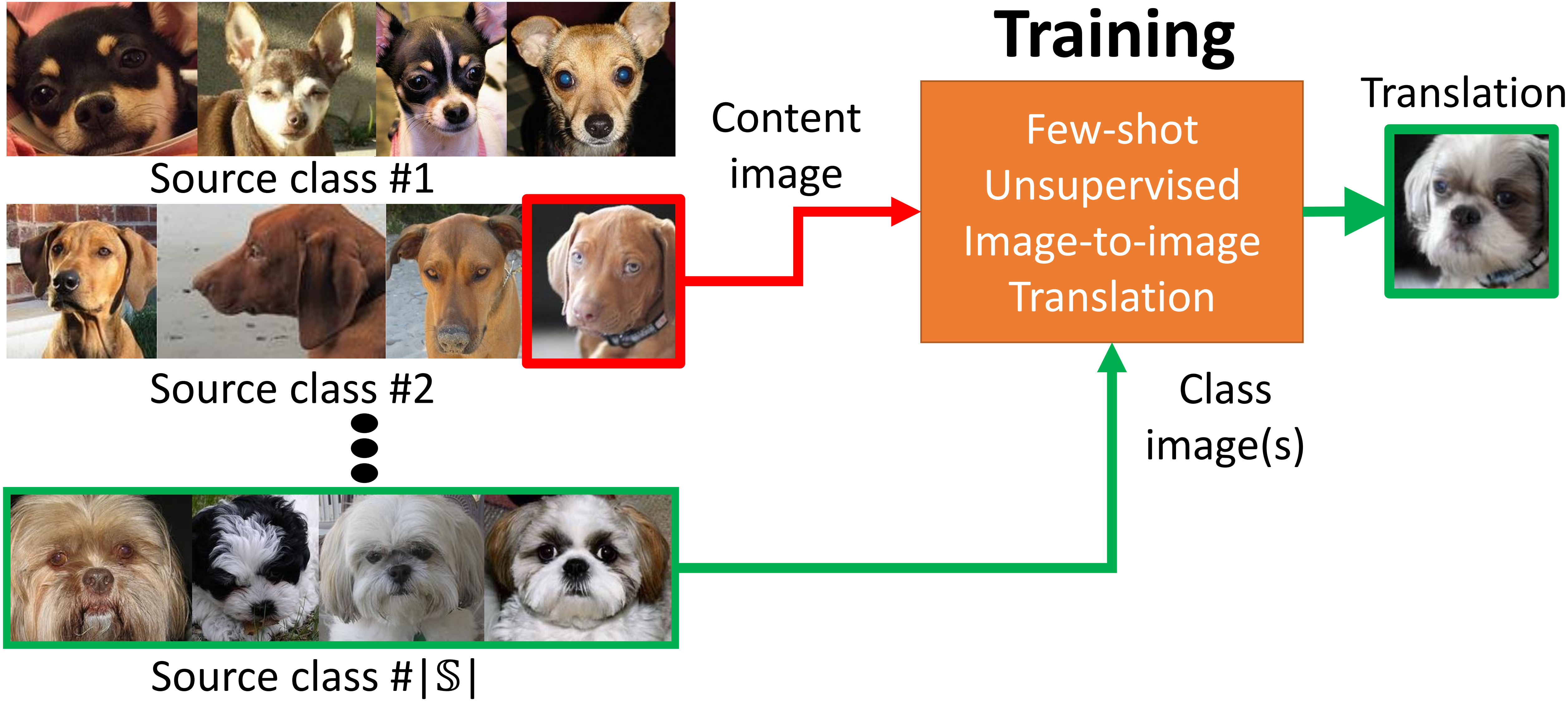
- We want to have an image-to-image translation model that can translate images to an unseen domain by leveraging few images of the unseen domain given in the test time.
  - Need to extract domain characteristics from few images in test time.
  - A model for all the unseen domains.
- This is an interesting combination.
- Challenging but useful.
- Human can do it.

A person seeing a standing tiger **for the first time** will have no trouble imagining what it will look lying down.

given a lifetime experience of other animals

Below is what our model achieves

Training



Source class #1

Source class #2

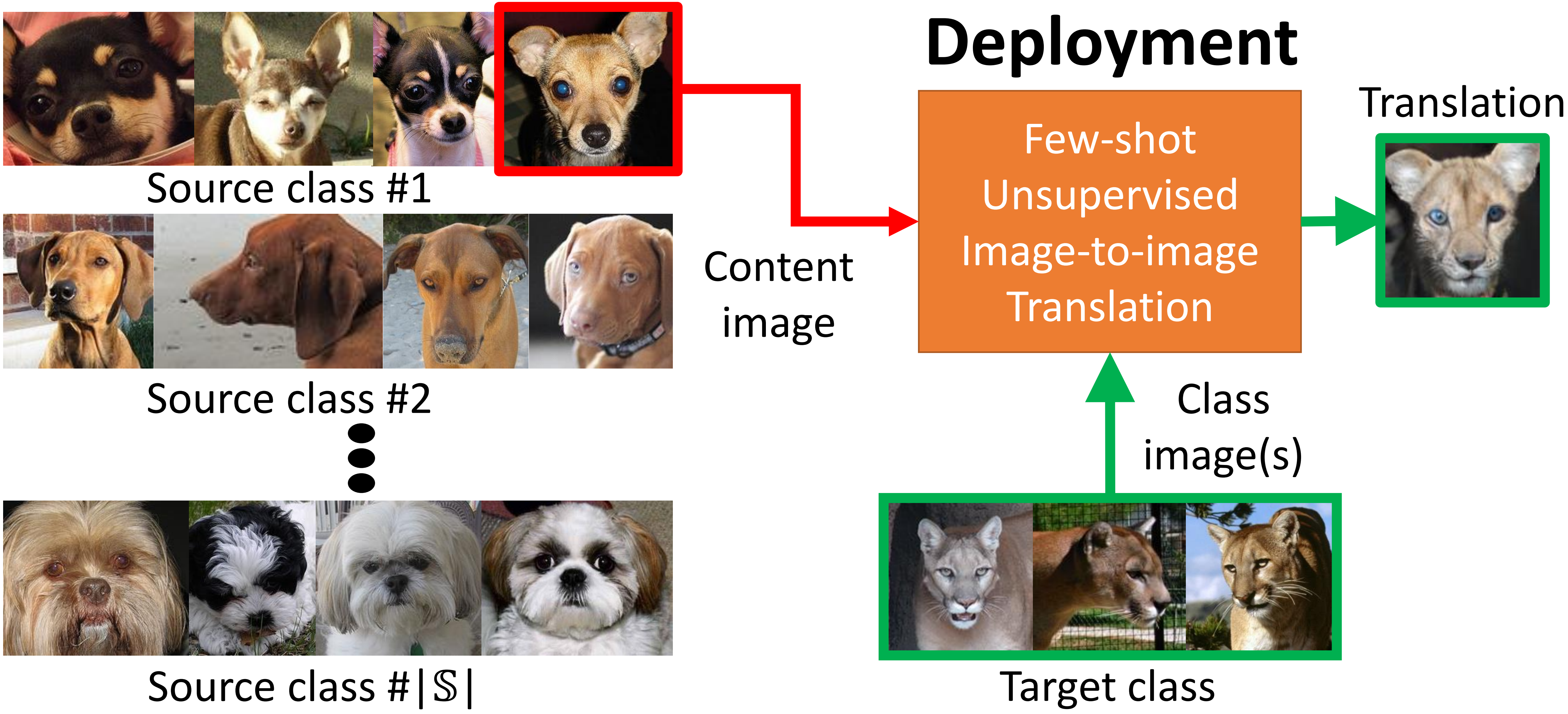
Source class #$|\mathbb{S}|$

Content image

Few-shot Unsupervised Image-to-image Translation

Class image(s)

Translation



Deployment

Source class #1

Source class #2

Source class #|$\mathbb{S}$|

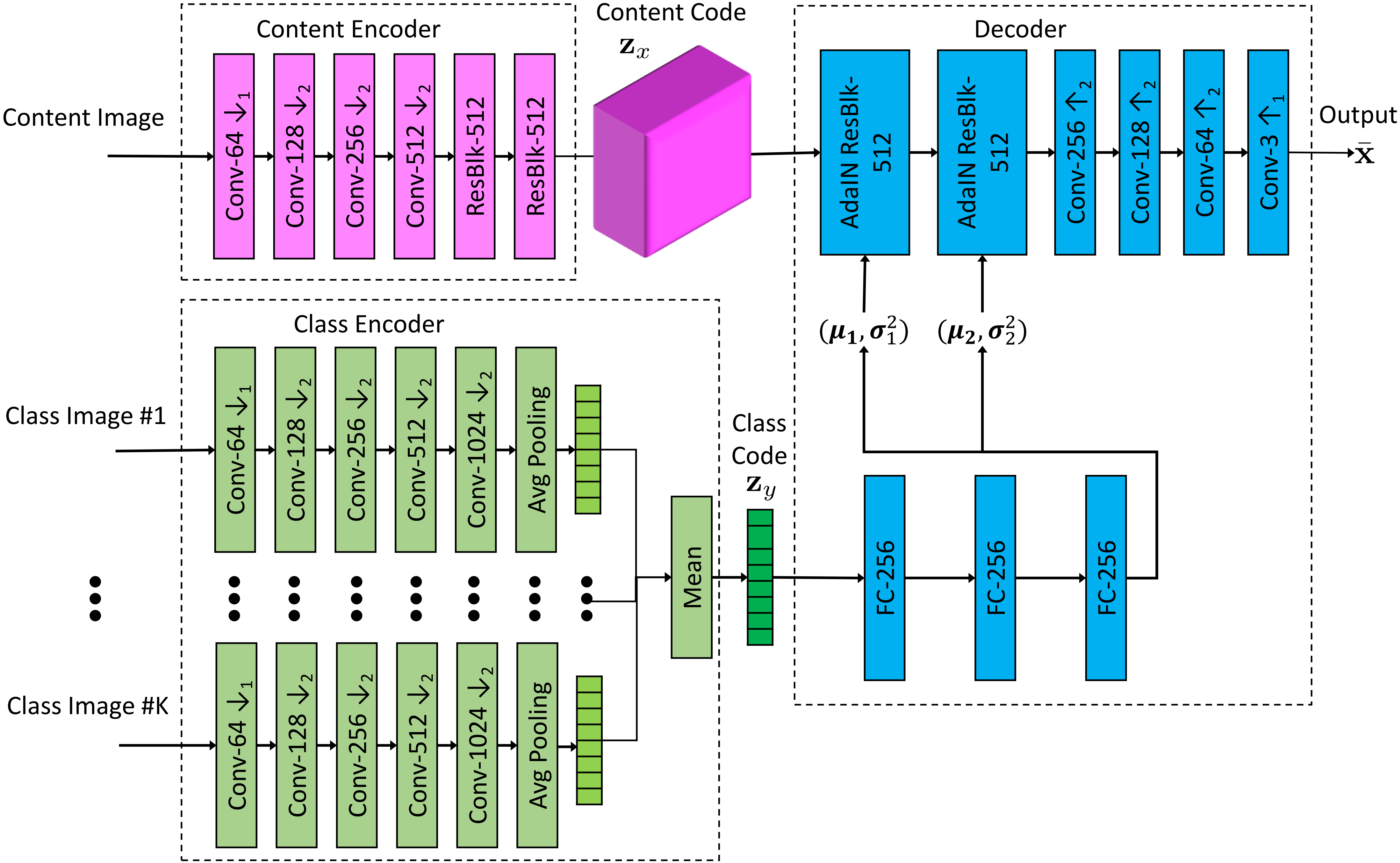
Content image

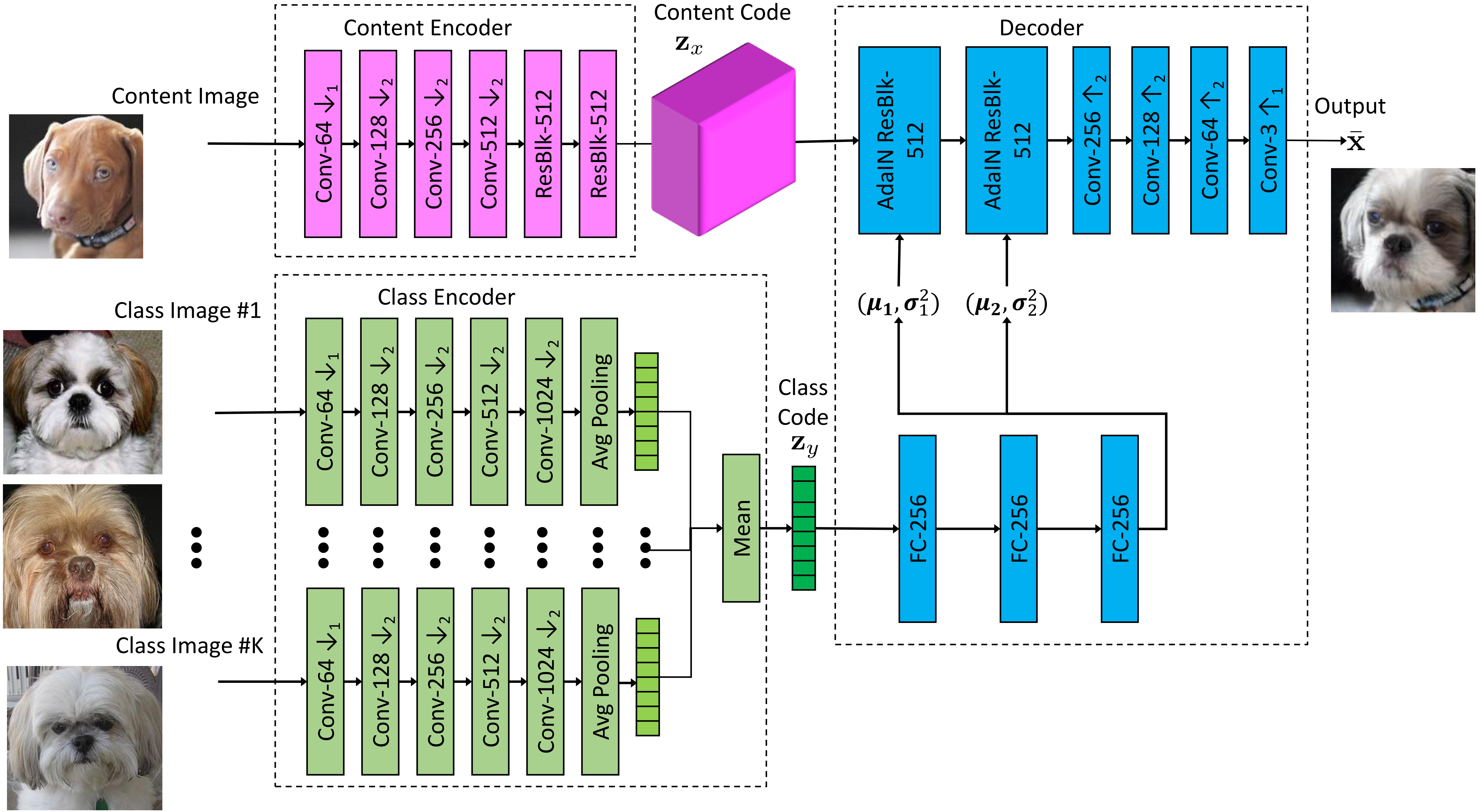
Few-shot Unsupervised Image-to-image Translation

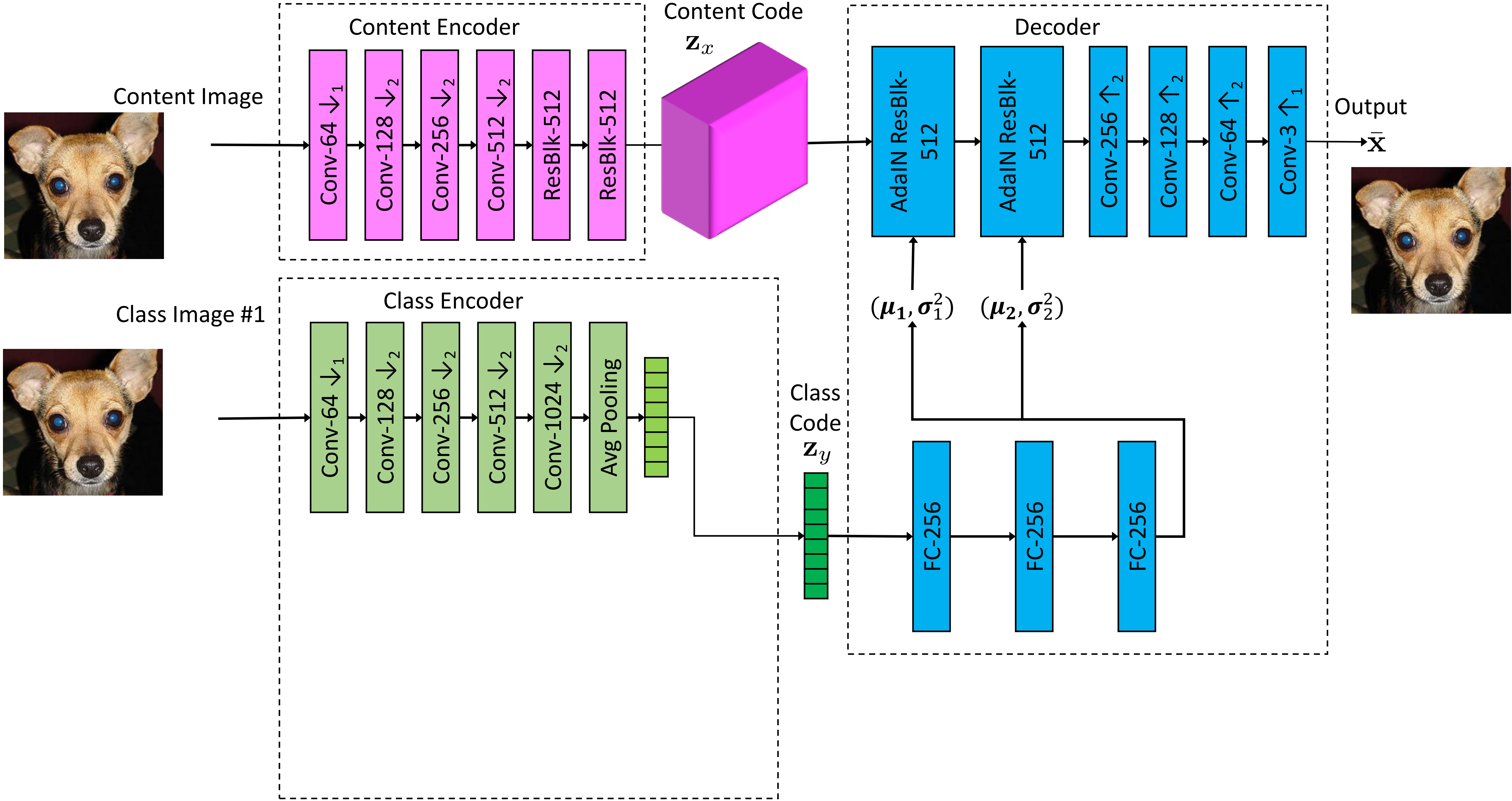
Translation

Class image(s)

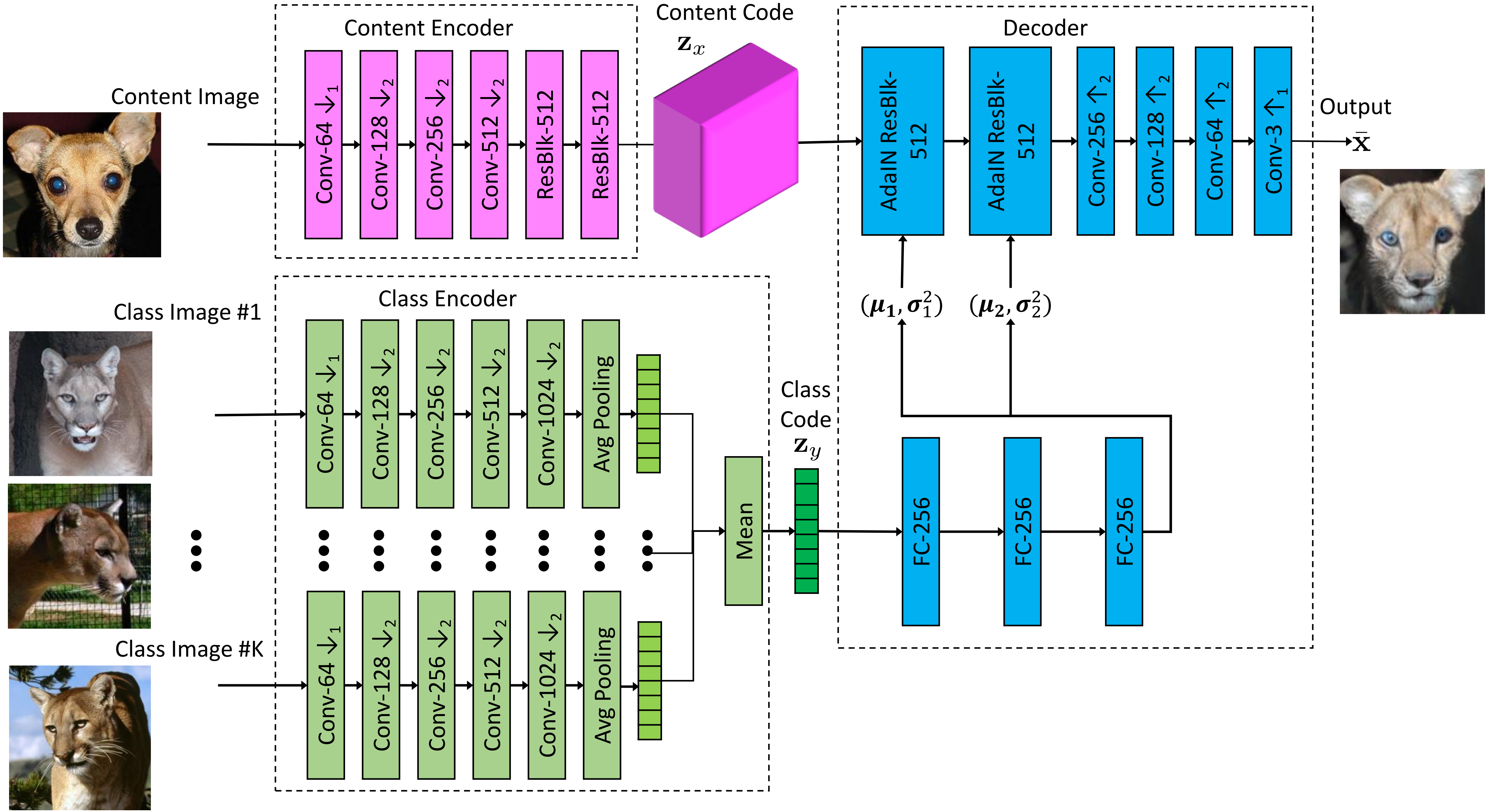
Target class


Content Encoder
Content Image
Conv-64 $\downarrow_1$
Conv-128 $\downarrow_2$
Conv-256 $\downarrow_2$
Conv-512 $\downarrow_2$
ResBlk-512
ResBlk-512
Content Code
$\mathbf{z}_x$
Decoder
AdaIN ResBlk-512
AdaIN ResBlk-512
Conv-256 $\uparrow_2$
Conv-128 $\uparrow_2$
Conv-64 $\uparrow_2$
Conv-3 $\uparrow_1$
Output
$\bar{\mathbf{x}}$
$(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2)$
$(\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2^2)$
Class Encoder
Class Image #1
Conv-64 $\downarrow_1$
Conv-128 $\downarrow_2$
Conv-256 $\downarrow_2$
Conv-512 $\downarrow_2$
Conv-1024 $\downarrow_2$
Avg Pooling
Class Image #K
Conv-64 $\downarrow_1$
Conv-128 $\downarrow_2$
Conv-256 $\downarrow_2$
Conv-512 $\downarrow_2$
Conv-1024 $\downarrow_2$
Avg Pooling
Mean
Class Code
$\mathbf{z}_y$
FC-256
FC-256
FC-256

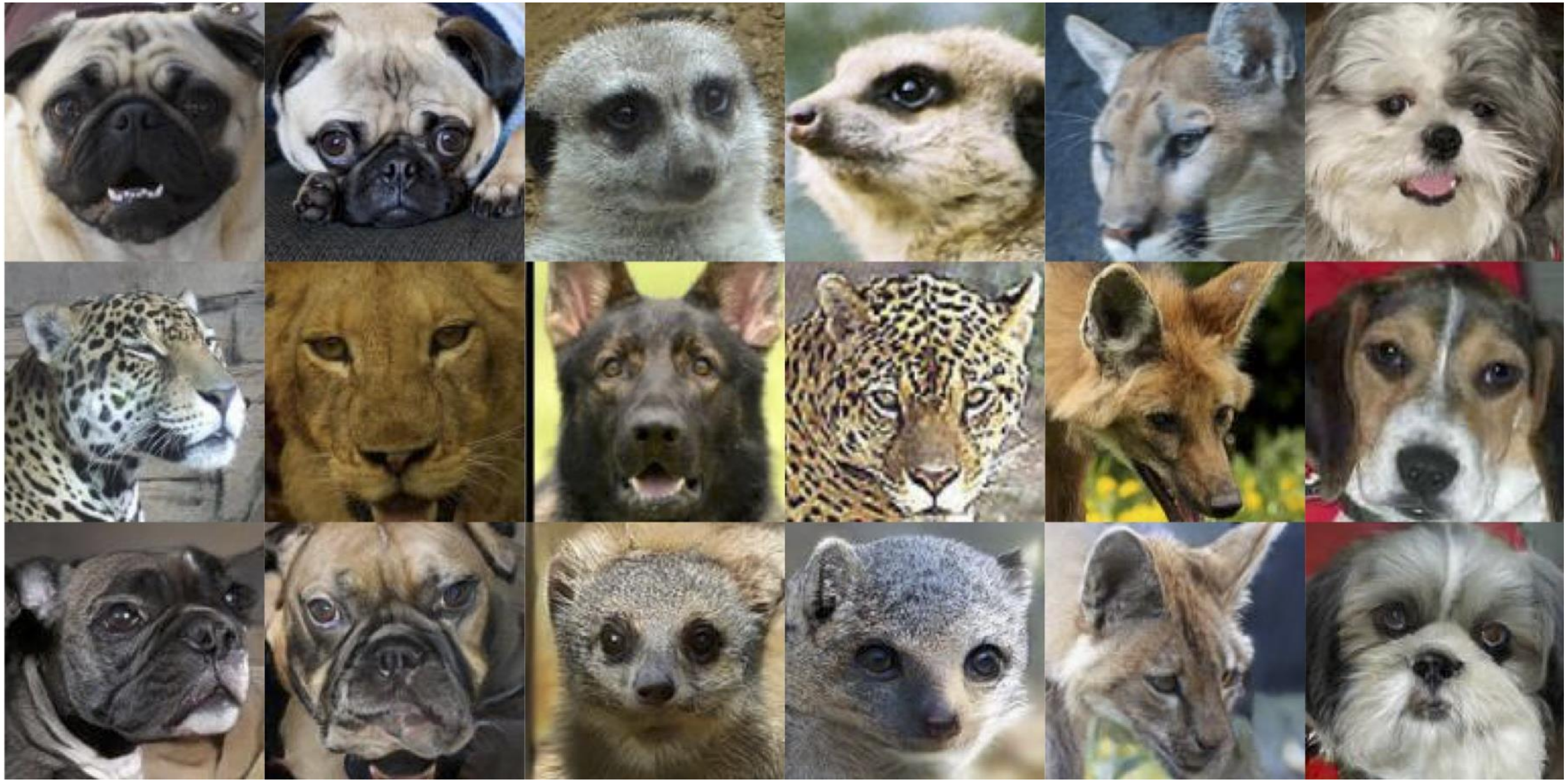

Content Image
Content Encoder
Conv-64 ↓1
Conv-128 ↓2
Conv-256 ↓2
Conv-512 ↓2
ResBlk-512
ResBlk-512
Content Code
$\mathbf{z}_x$
Decoder
AdaIN ResBlk-512
AdaIN ResBlk-512
Conv-256 ↑2
Conv-128 ↑2
Conv-64 ↑2
Conv-3 ↑1
Output
$\bar{\mathbf{x}}$
$(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2)$
$(\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2^2)$
Class Image #1
Class Image #K
Class Encoder
Conv-64 ↓1
Conv-128 ↓2
Conv-256 ↓2
Conv-512 ↓2
Conv-1024 ↓2
Avg Pooling
Mean
Class Code
$\mathbf{z}_y$
FC-256
FC-256
FC-256



Content Image
Content Encoder
Conv-64 ↓1
Conv-128 ↓2
Conv-256 ↓2
Conv-512 ↓2
ResBlk-512
ResBlk-512
Content Code
$\mathbf{z}_x$
Decoder
AdaIN ResBlk-512
AdaIN ResBlk-512
Conv-256 ↑2
Conv-128 ↑2
Conv-64 ↑2
Conv-3 ↑1
Output
$\bar{\mathbf{x}}$
$(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2)$
$(\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2^2)$
Class Image #1
Class Encoder
Conv-64 ↓1
Conv-128 ↓2
Conv-256 ↓2
Conv-512 ↓2
Conv-1024 ↓2
Avg Pooling
Class Code
$\mathbf{z}_y$
FC-256
FC-256
FC-256


Content Image
Content Encoder
Conv-64 $\downarrow_1$
Conv-128 $\downarrow_2$
Conv-256 $\downarrow_2$
Conv-512 $\downarrow_2$
ResBlk-512
ResBlk-512
Content Code
$\mathbf{z}_x$
Decoder
AdaIN ResBlk-512
AdaIN ResBlk-512
Conv-256 $\uparrow_2$
Conv-128 $\uparrow_2$
Conv-64 $\uparrow_2$
Conv-3 $\uparrow_1$
Output
$\bar{\mathbf{x}}$
$(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2)$
$(\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2^2)$
Class Image #1
Class Image #K
Class Encoder
Conv-64 $\downarrow_1$
Conv-128 $\downarrow_2$
Conv-256 $\downarrow_2$
Conv-512 $\downarrow_2$
Conv-1024 $\downarrow_2$
Avg Pooling
Conv-64 $\downarrow_1$
Conv-128 $\downarrow_2$
Conv-256 $\downarrow_2$
Conv-512 $\downarrow_2$
Conv-1024 $\downarrow_2$
Avg Pooling
Mean
Class Code
$\mathbf{z}_y$
FC-256
FC-256
FC-256

Animals

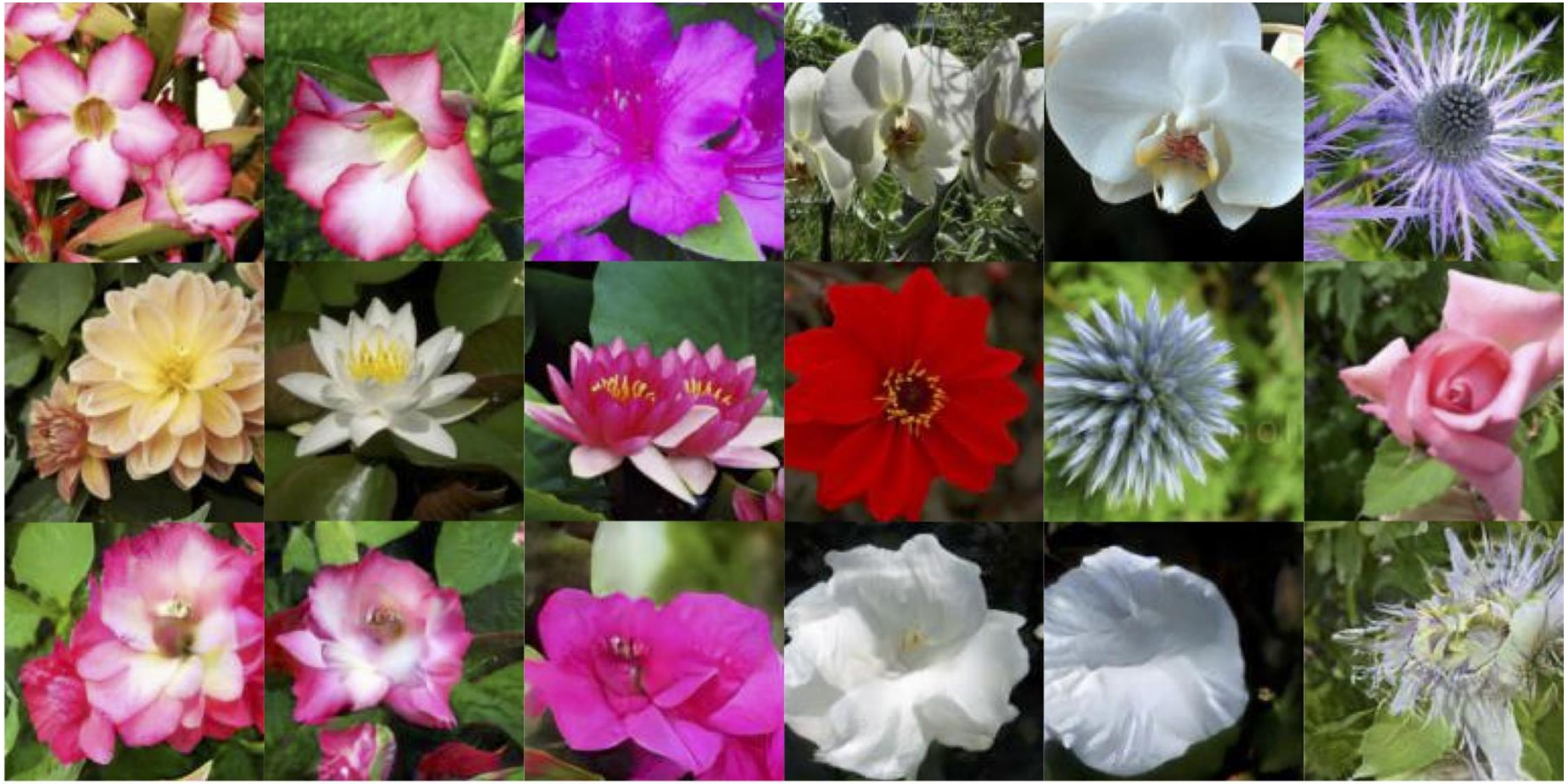


Target

Input

Output

Flowers



Target

Input

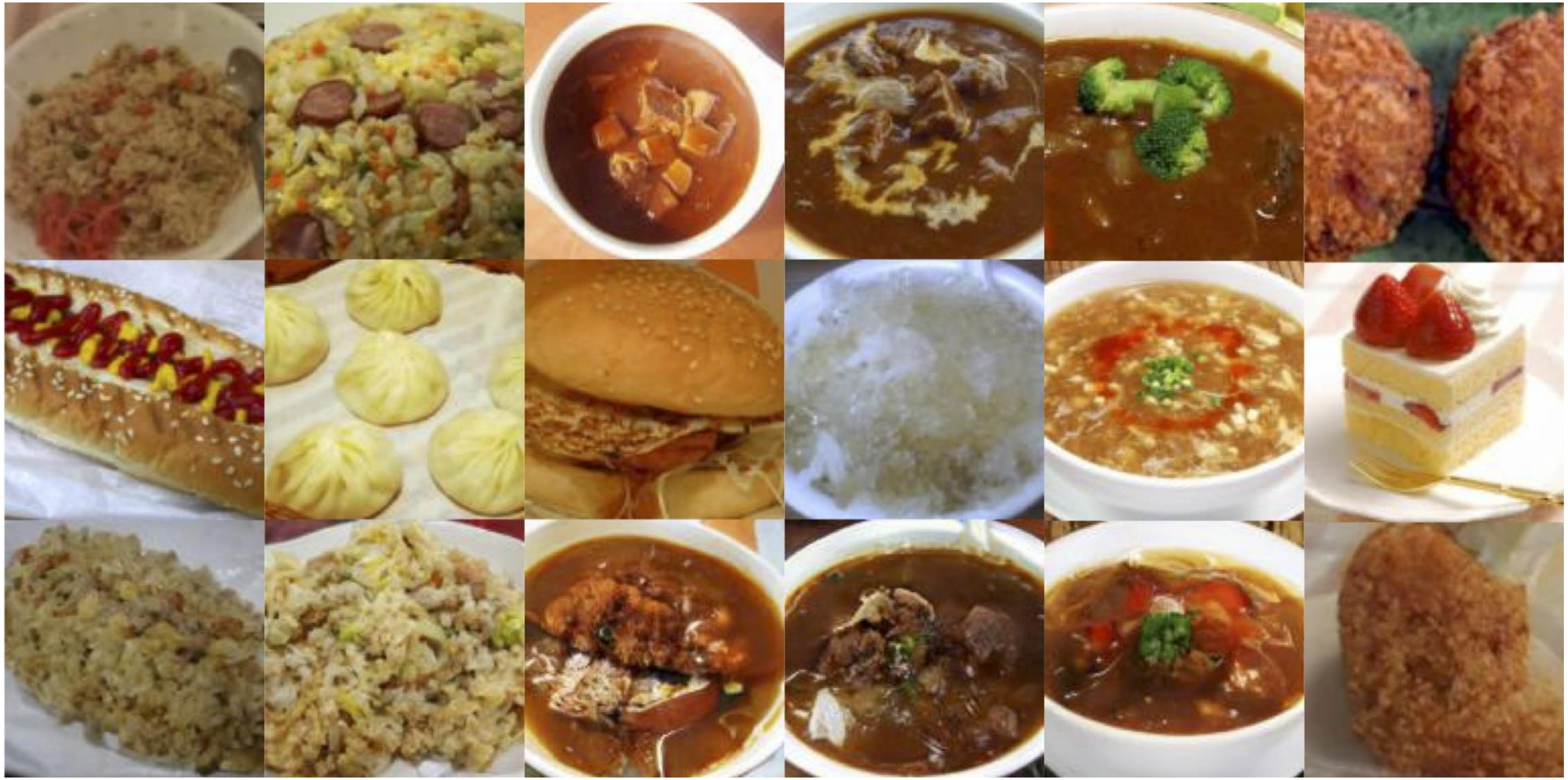
Output

Food

Target

Input

Output



Limitations

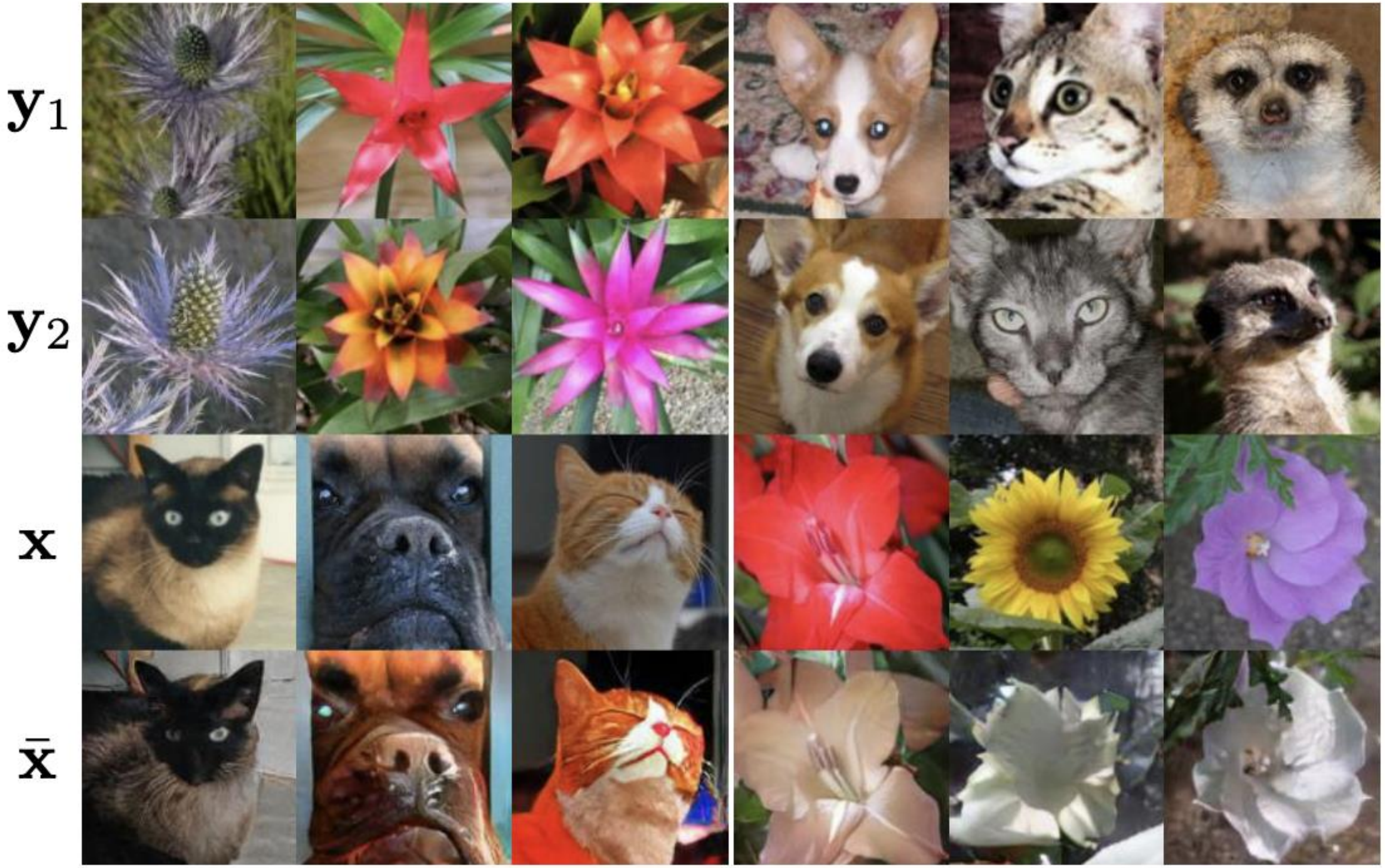


Figure 5. Limitations of the proposed framework. When the appearance of a unseen object class is dramatically different to the appearances of the source classes, (*e.g*. flower and animal face). The proposed `FUNIT` framework fails to generate meaningful translation outputs.

Failure cases

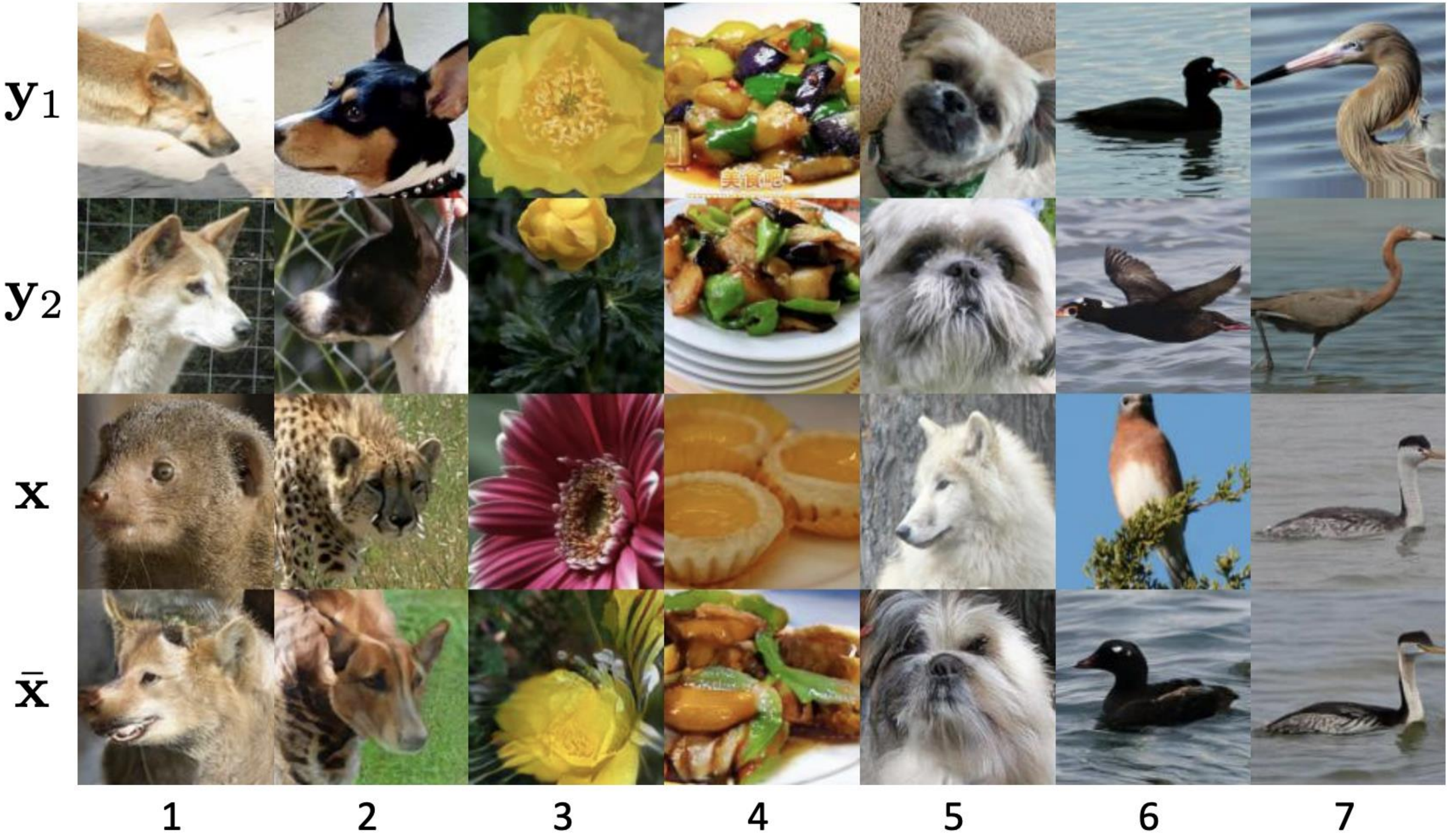
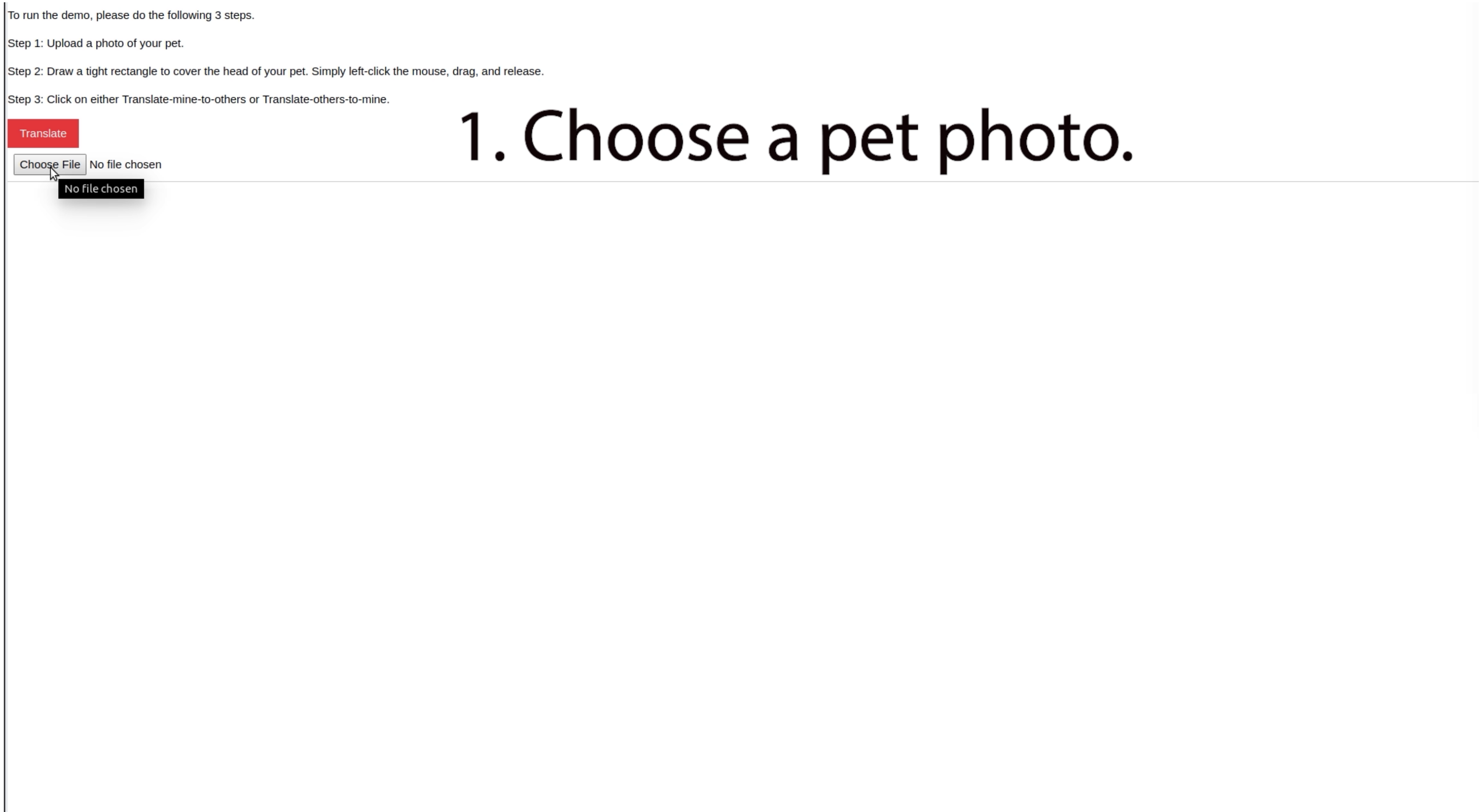


Figure 9. Failure cases. The typical failure cases of the proposed FUNIT model include generating hybrid objects (*e.g*. column 1, 2, 3, and 4), ignoring input content images (*e.g*. column 5 and 6), and ignoring input class images (*e.g*. column 7).

Petswap Demo
https://nvlabs.github.io/FUNIT/petswap.html



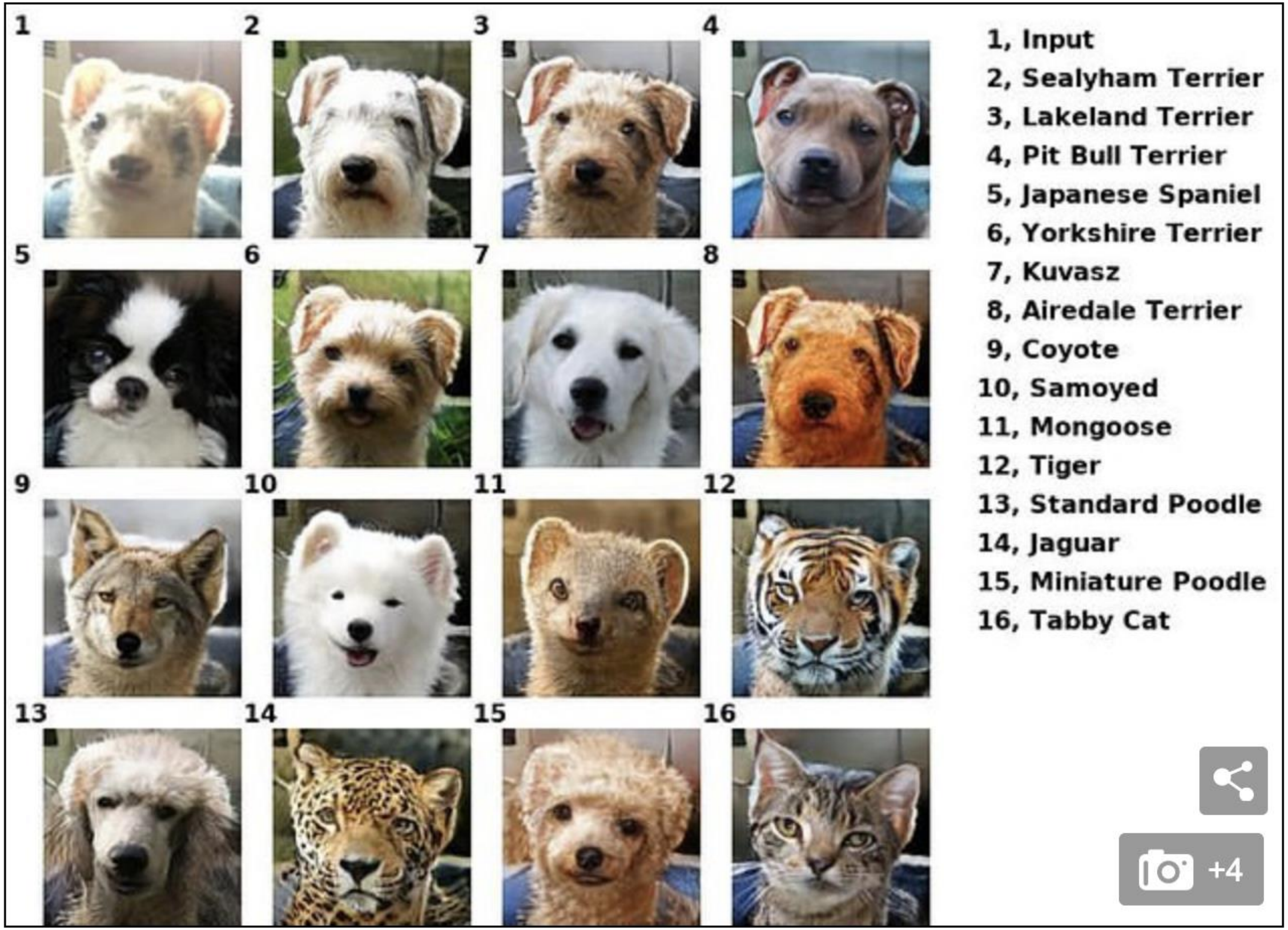
To run the demo, please do the following 3 steps.

Step 1: Upload a photo of your pet.

Step 2: Draw a tight rectangle to cover the head of your pet. Simply left-click the mouse, drag, and release.

Step 3: Click on either Translate-mine-to-others or Translate-others-to-mine.

1. Choose a pet photo.


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
1, Input
2, Sealyham Terrier
3, Lakeland Terrier
4, Pit Bull Terrier
5, Japanese Spaniel
6, Yorkshire Terrier
7, Kuvasz
8, Airedale Terrier
9, Coyote
10, Samoyed
11, Mongoose
12, Tiger
13, Standard Poodle
14, Jaguar
15, Miniature Poodle
16, Tabby Cat
+4

1

2

5

6

9

10

13

14

1

2

5

6

9

10

13

14

1

2

5

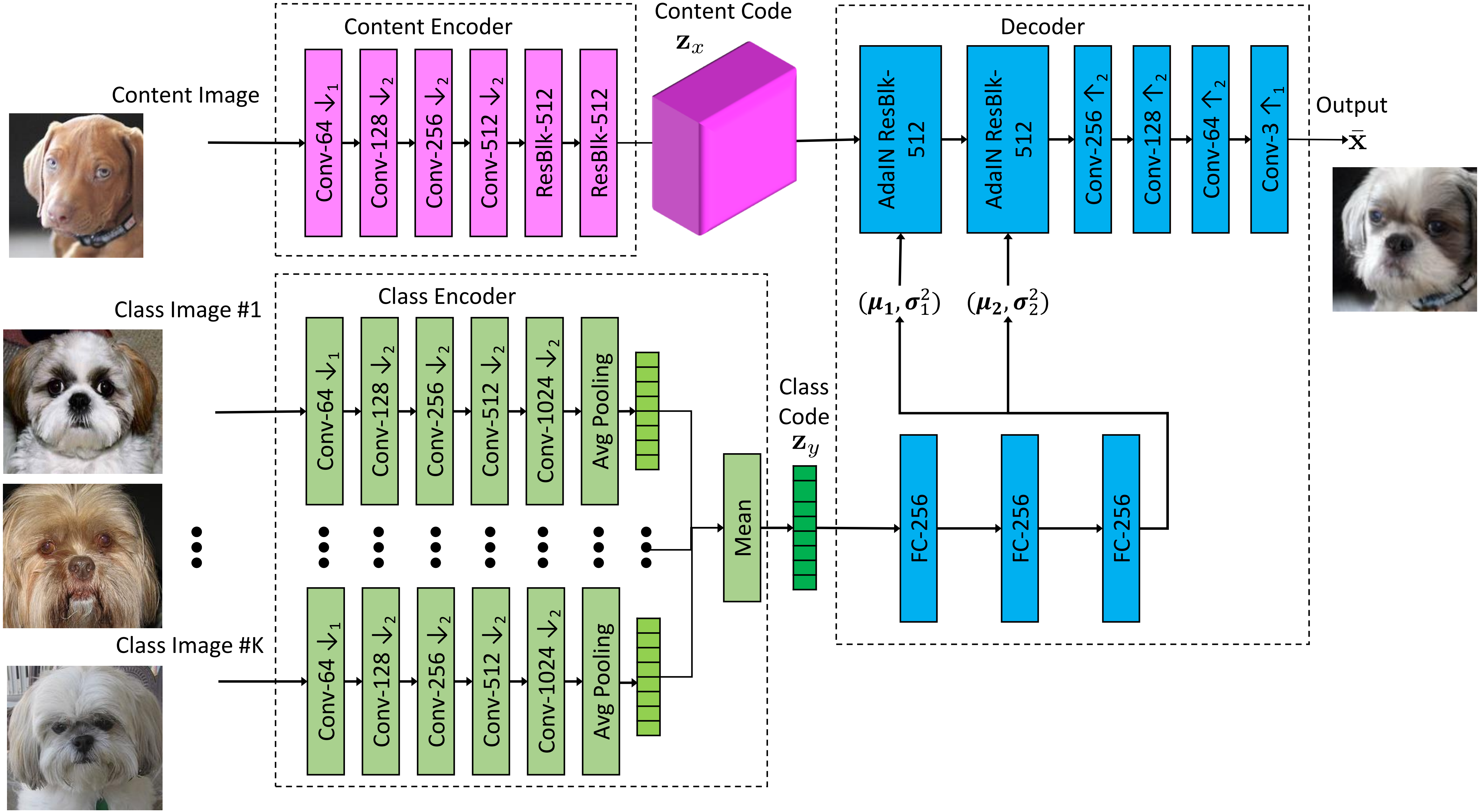
6

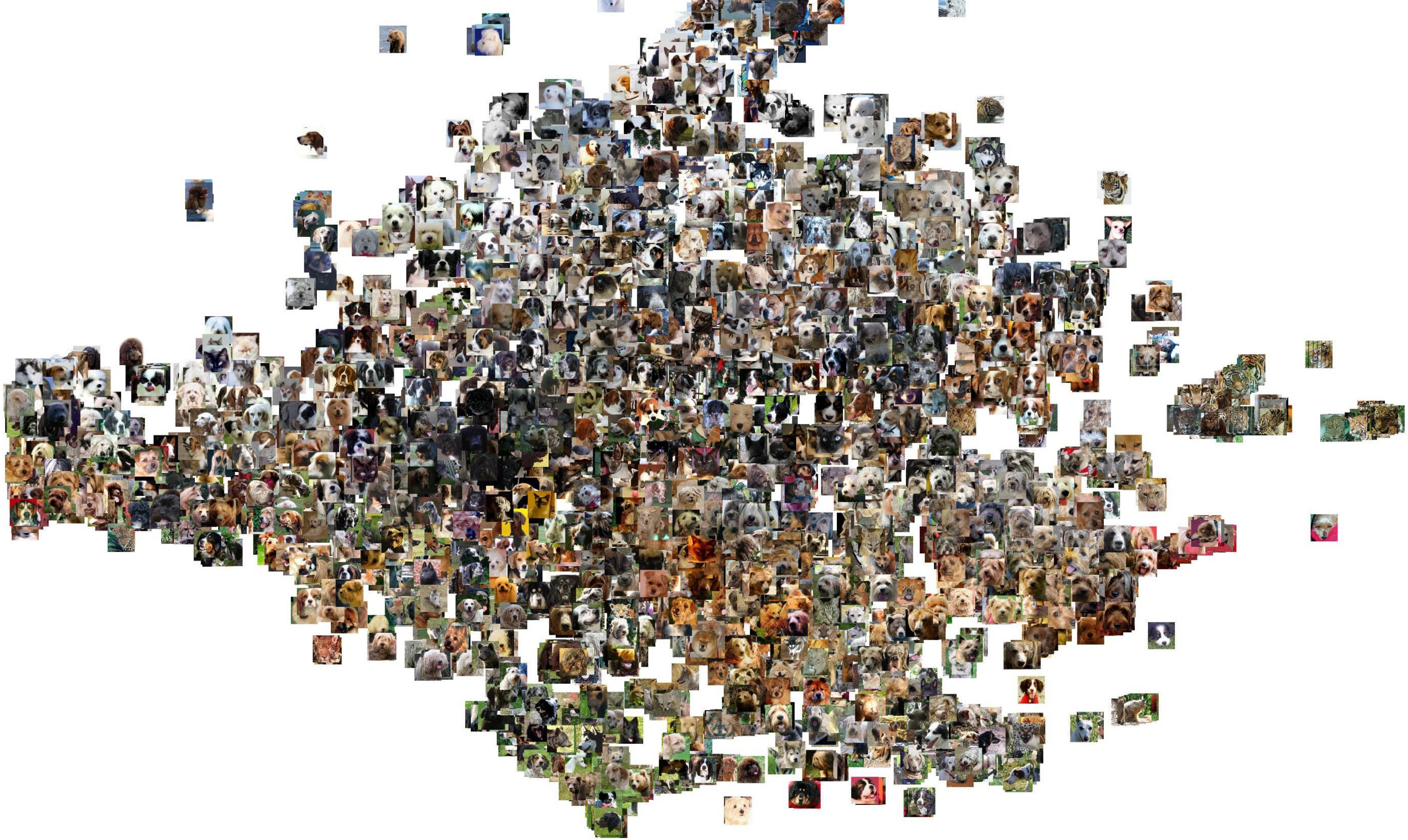
9

10

13

14


Content Encoder
Content Code
$\mathbf{z}_x$
Decoder
Content Image
Conv-64 ↓1
Conv-128 ↓2
Conv-256 ↓2
Conv-512 ↓2
ResBlk-512
ResBlk-512
AdaIN ResBlk-512
AdaIN ResBlk-512
Conv-256 ↑2
Conv-128 ↑2
Conv-64 ↑2
Conv-3 ↑1
Output
$\bar{\mathbf{x}}$
$(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2)$
$(\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2^2)$
Class Image #1
Class Encoder
Conv-64 ↓1
Conv-128 ↓2
Conv-256 ↓2
Conv-512 ↓2
Conv-1024 ↓2
Avg Pooling
Class Code
$\mathbf{z}_y$
Mean
FC-256
FC-256
FC-256
Class Image #K



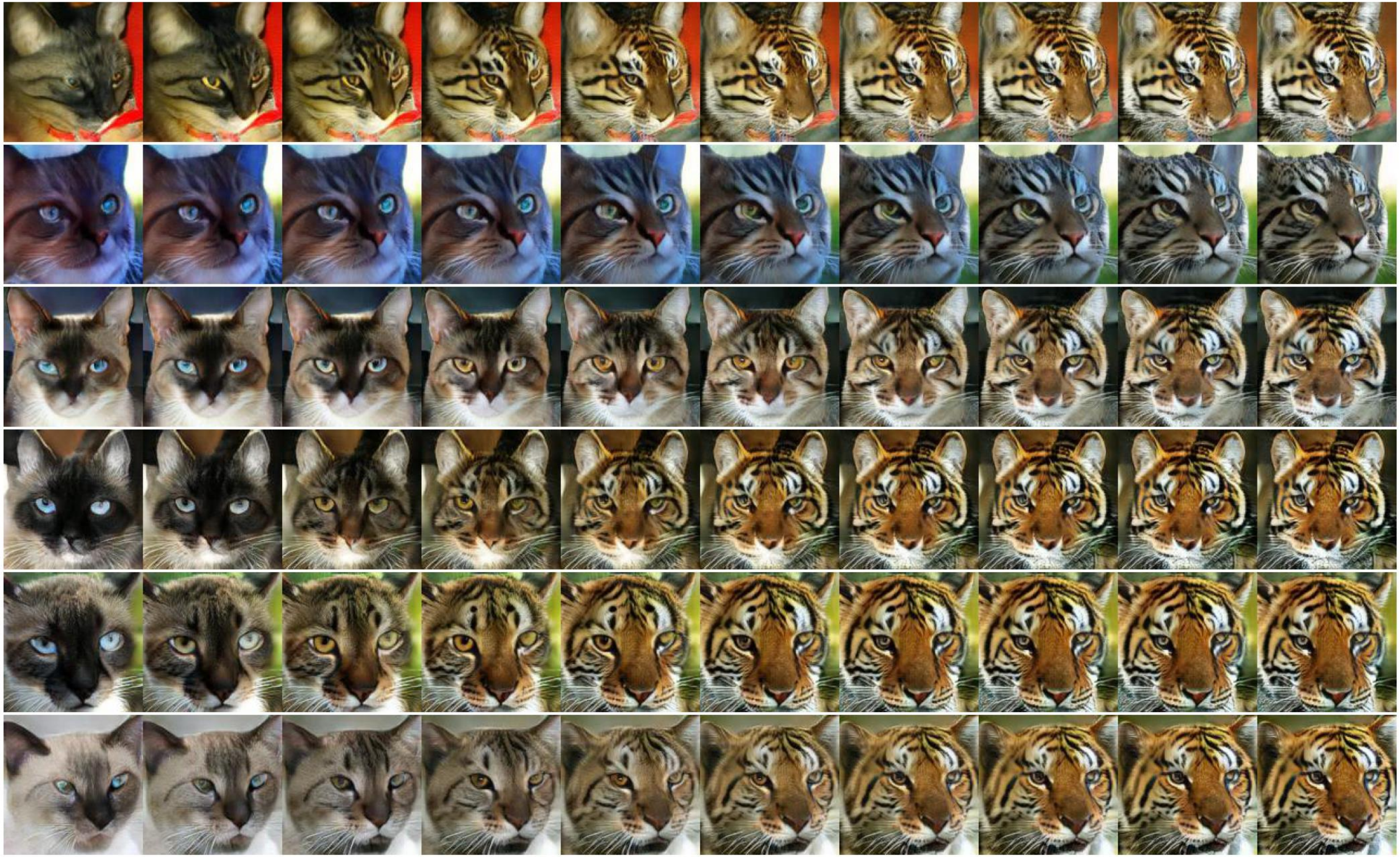
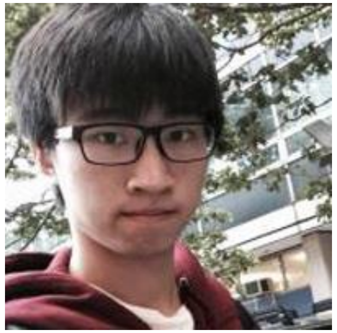


Figure 12. Interpolation by keeping the content code fixed while interpolating between two class codes of source classes.

Conclusion

- Generating images of unseen domain in the few shot setting
- Extract target domain appearance information via a class encoder with AdaIN
- ICCV 2019
- Online demo link: http://nvidia-research-mingyuliu.com/funit
- FUNIT code: https://github.com/nvlabs/funit/
- Paper: https://arxiv.org/abs/1905.01723

Ming-Yu Liu
NVIDIA



Xun Huang
NVIDIA, Cornell

Arun Mallya
NVIDIA

Tero Karras
NVIDIA

Timo Aila
NVIDIA

Jaakko Lehtinen
NVIDIA

Jan Kautz
NVIDIA