ECCV'20 ONLINE
23-28 AUGUST 2020

國立臺灣大學
National Taiwan University

Domain-Specific Mapping for Generative Adversarial Style Transfers

Hsin-Yu Chang Zhixiang Wang Yung-Yu Chuang

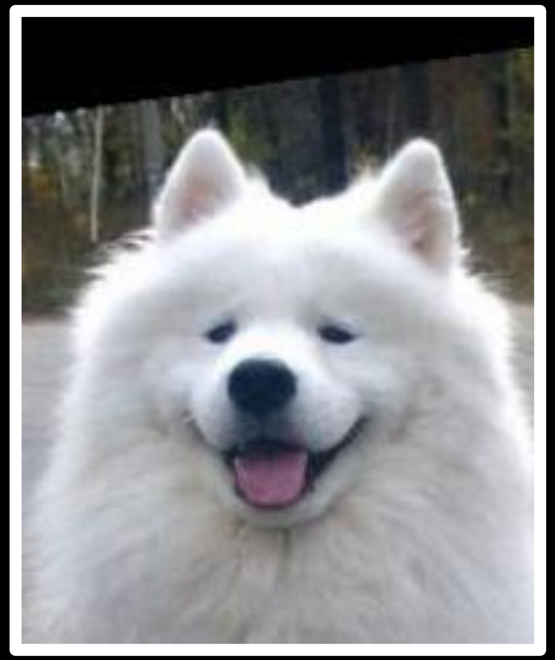


National Taiwan University

Style Transfer

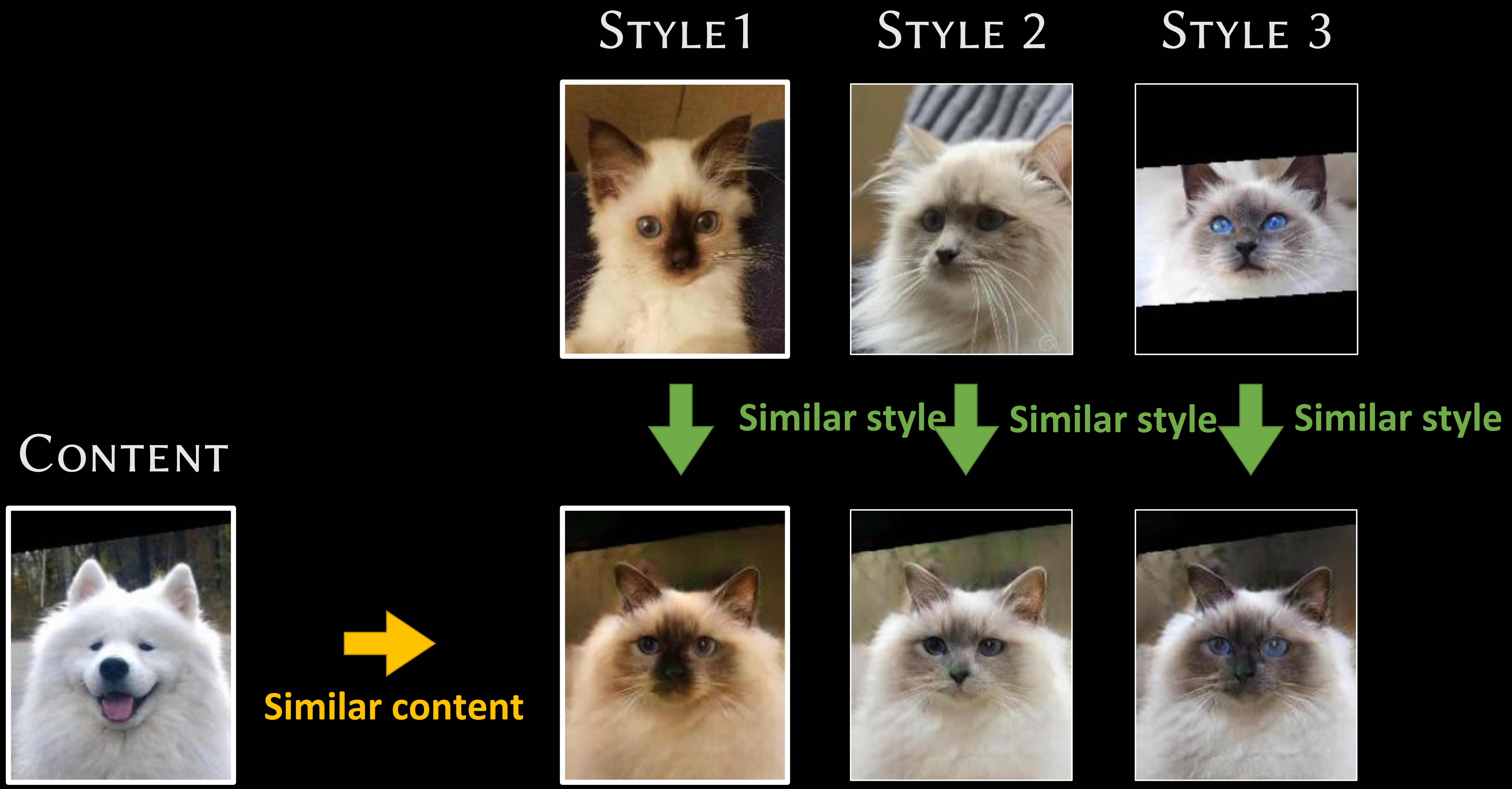
Style 1

Content



Our goal is to learn style transfer.

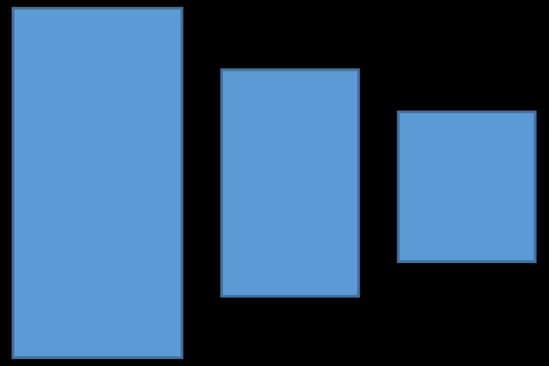
Style Transfer



Given the content and style images, we want to generate results that preserve the content information while performing style translation.

Solution- Unsupervised image-to-image translation

Examplar-guided

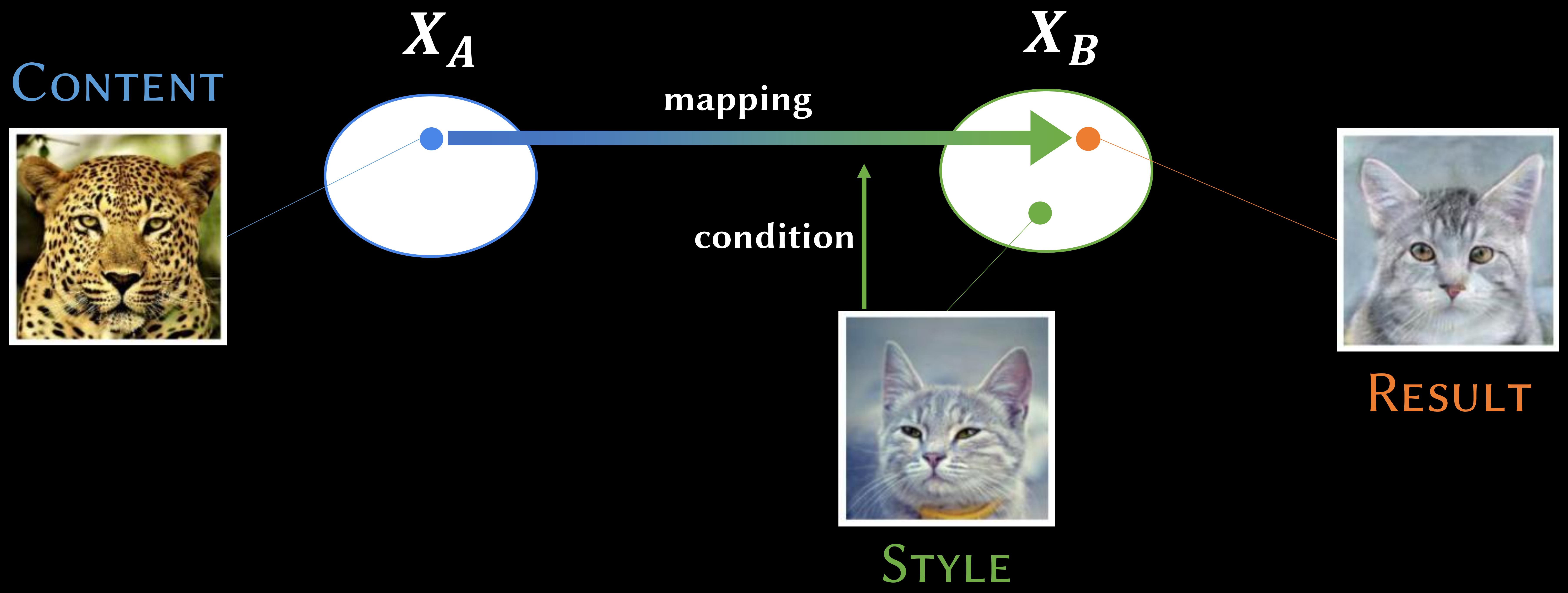


Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to- image translation. In: ECCV(2018)

Examplar-guided I2I translation approaches have been shown effective for style transfer.

Solution- Unsupervised image-to-image translation

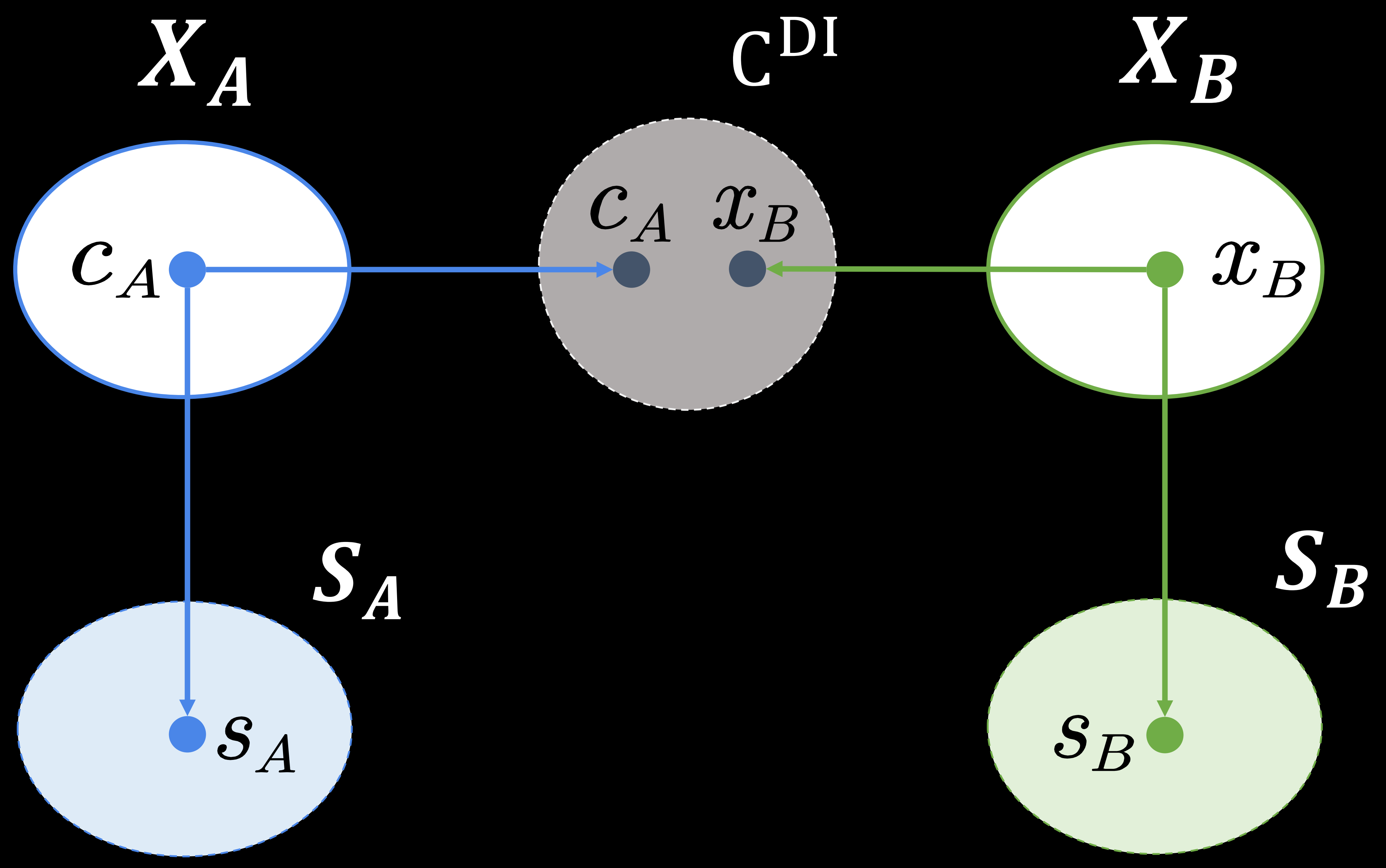
- Learn the mapping between two image domains



I2I translation aims at learning the mapping between images of two domains.
It can be employed for style transfer when
given the content image in domain A and the style image in domain B.

Related Work- Unsupervised image-to-image translation

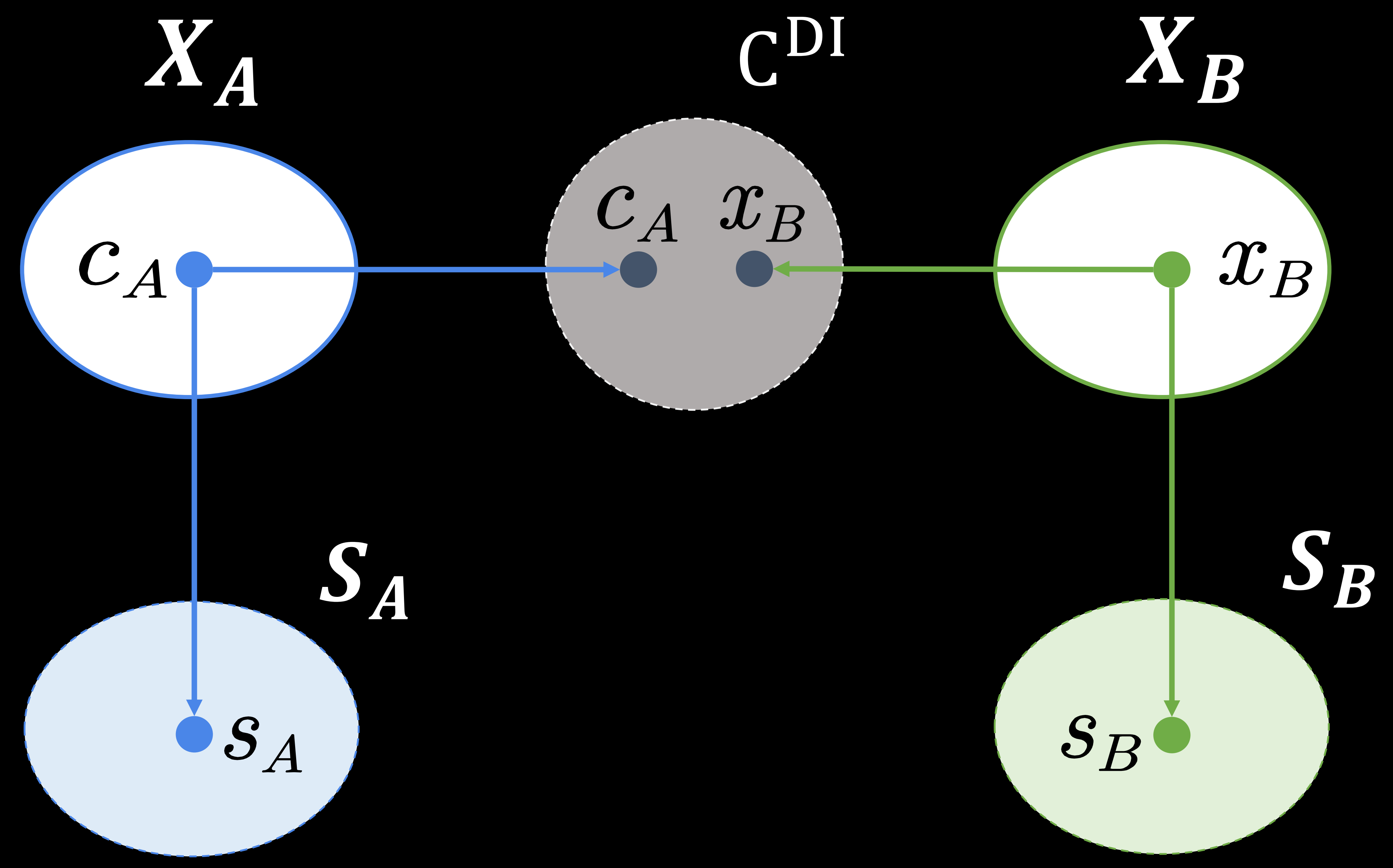
- MUNIT [Huang et al. 2018], DRIT [Lee et al. 2018]



MUNIT and DRIT have shown great success through disentangled representations.

Related Work- Unsupervised image-to-image translation

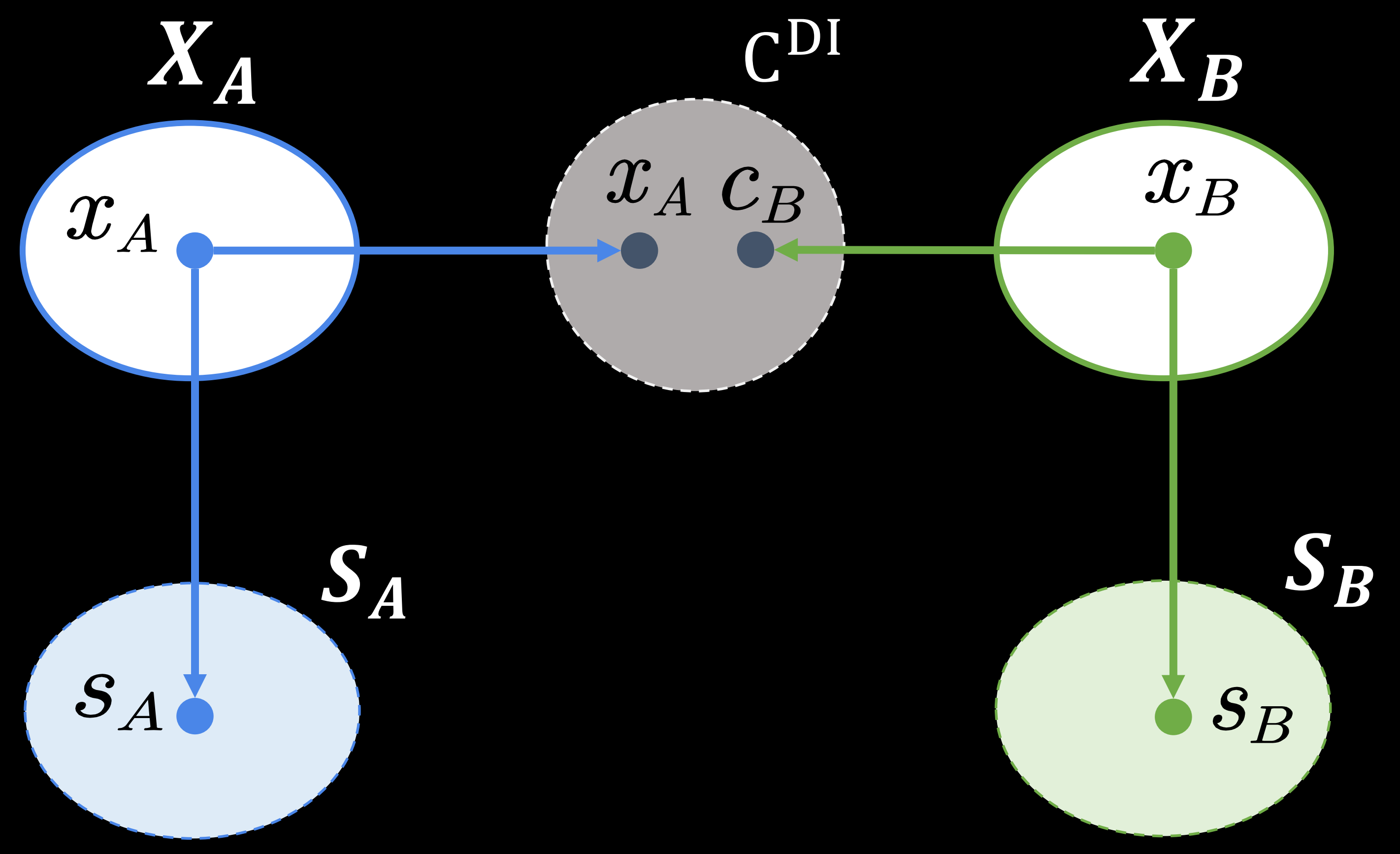
- MUNIT [Huang et al. 2018], DRIT [Lee et al. 2018]



They decompose an image into a content feature in the shared domain-invariant content space and a style feature in the domain-specific style space.

Related Work- Unsupervised image-to-image translation

- MSGAN [Mao et al. 2019]
  - Add mode seeking loss to improve the diversity of generated images
- GDWCT [Choi et al. 2019]
  - Apply WCT to unsupervised I2I translation



The advanced work MSGAN and GDWCT also assume the share content space.
The shared space could limit representation power.

Style Transfer- Global style & local style

Photo → Monet

more global ⟵ ⟶ more local

color tune, texture

structural semantic

For the Photo → Monet task, the style transfer is "more global"
as its success mainly counts on adjusting global attributes such as color tones and textures.

Style Transfer- Global style & local style

Dog ⟶ Cat

more global

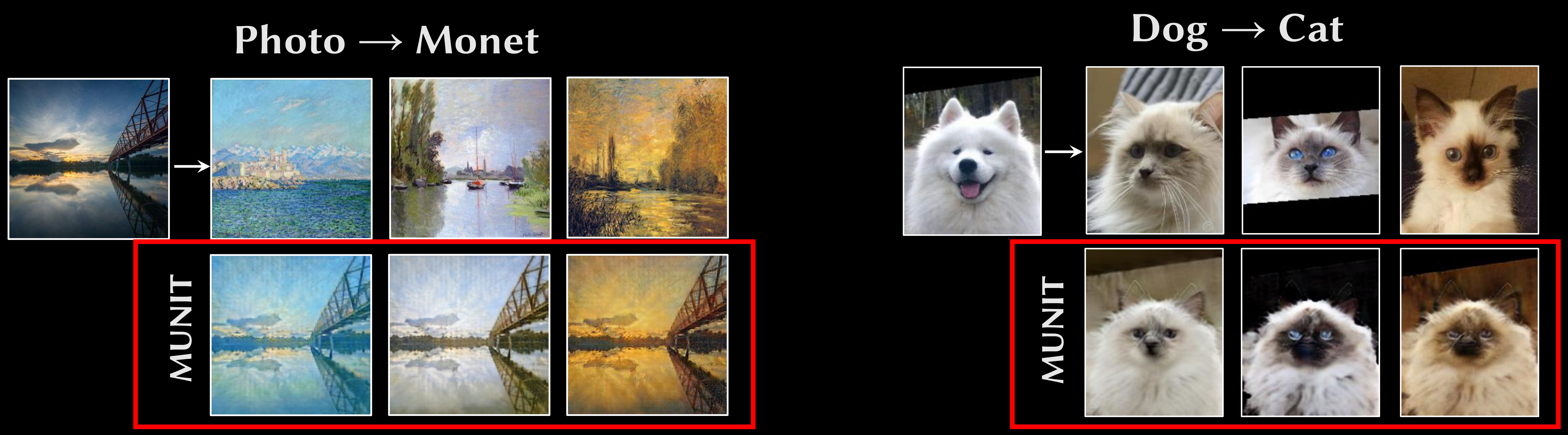
color tune, texture

structural semantic

more local

For the Dog ⟶ Cat task, the style transfer is "more local"
as its success requires more attention on local and structural semantic correspondences.

Style Transfer- Global style & local style



content representation problem

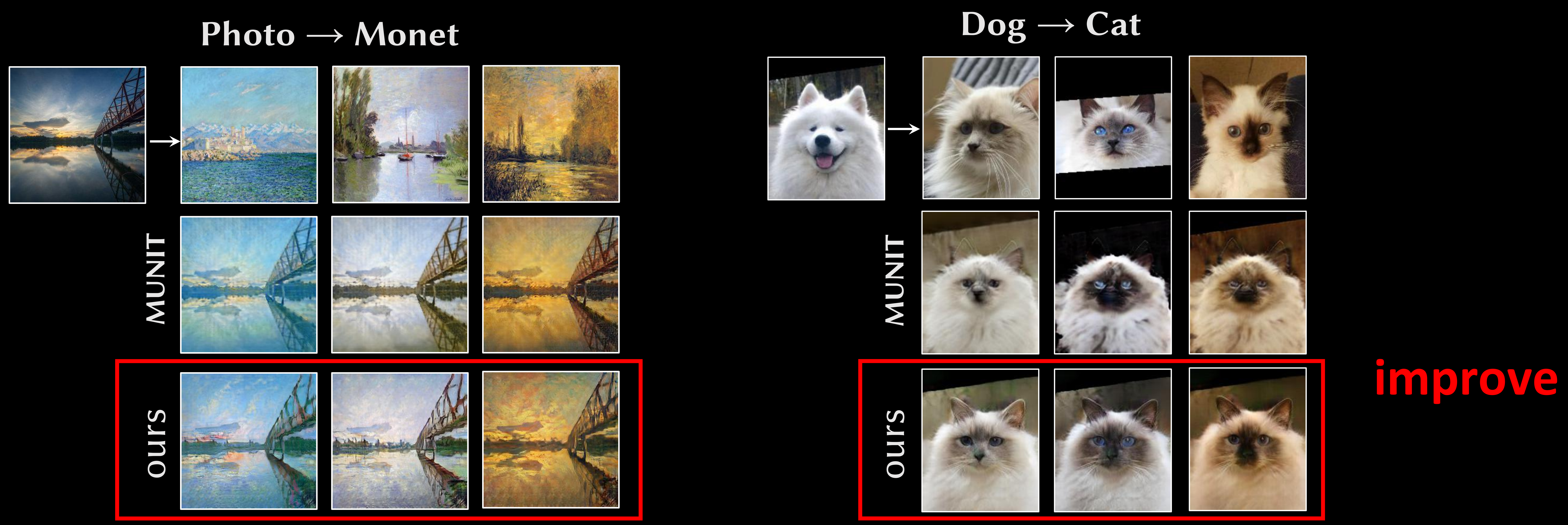
more global ← → more local

color tune, texture

structural semantic

However, previous I2I methods with disentangled representations often run into problems in "more local" style transfer scenarios.

Style Transfer- Global style & local style



more global

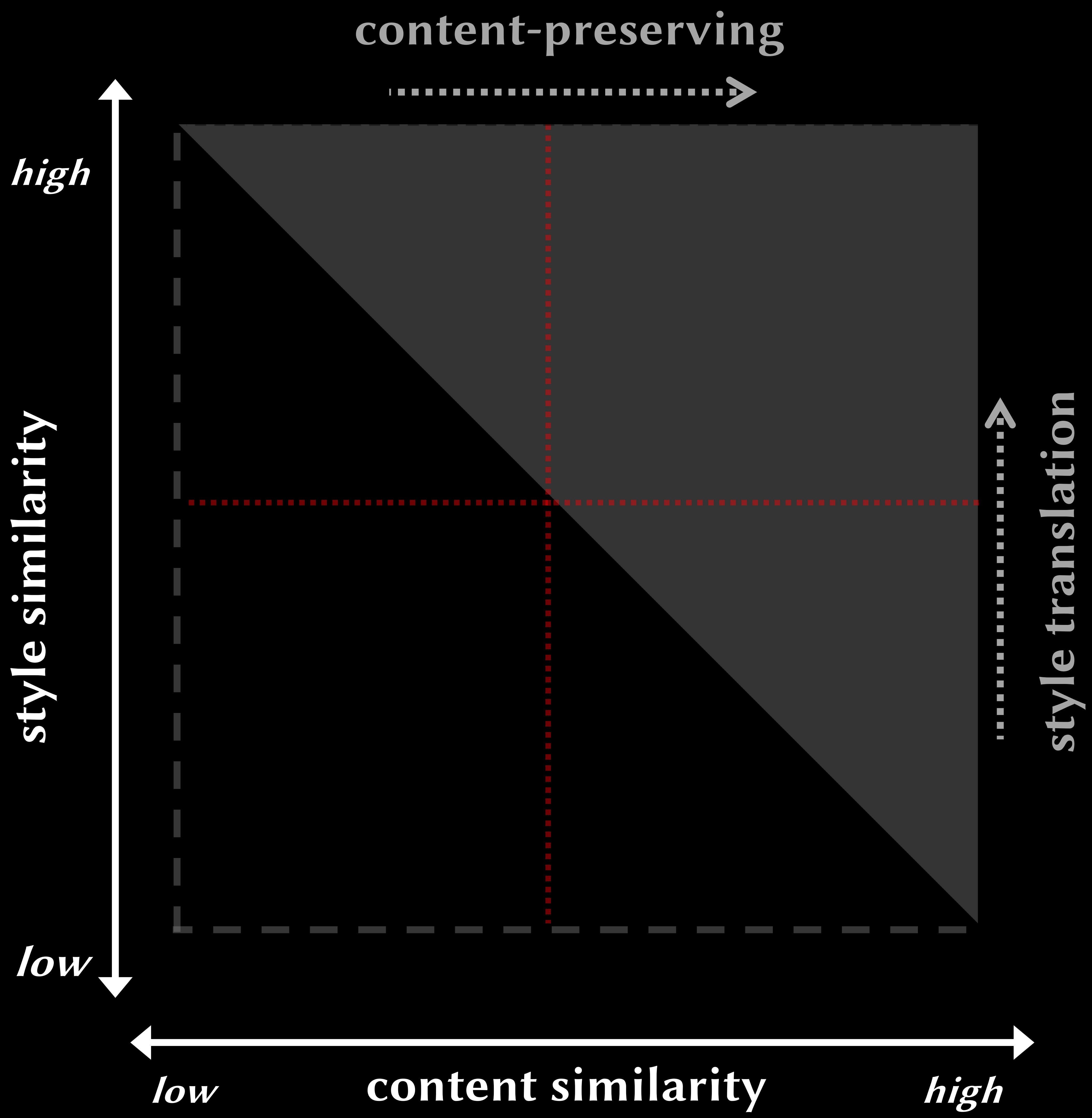
color tune, texture

more local

structural semantic

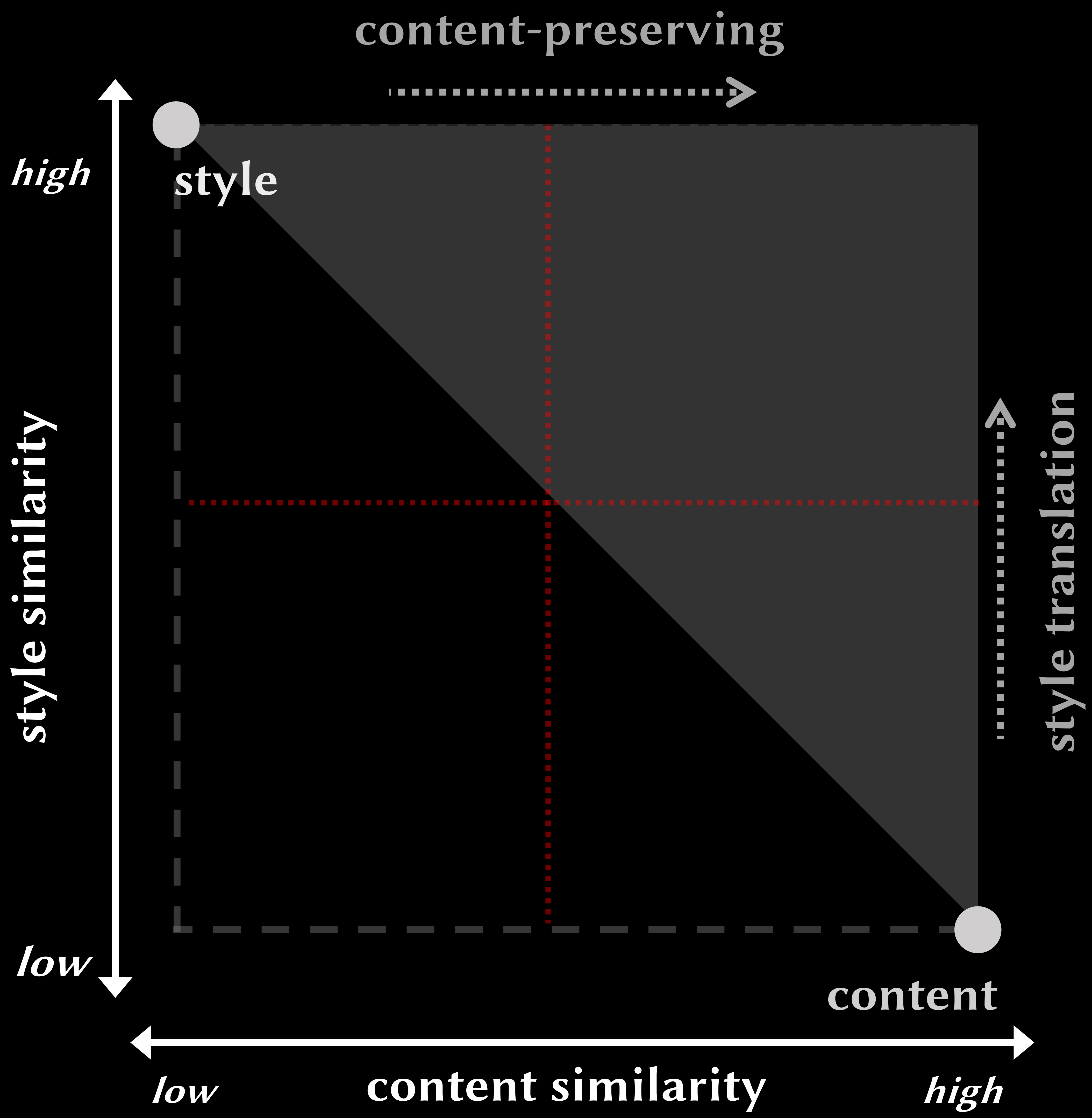
Our method improves the quality of translation and handles both local and global style transfer scenarios well.

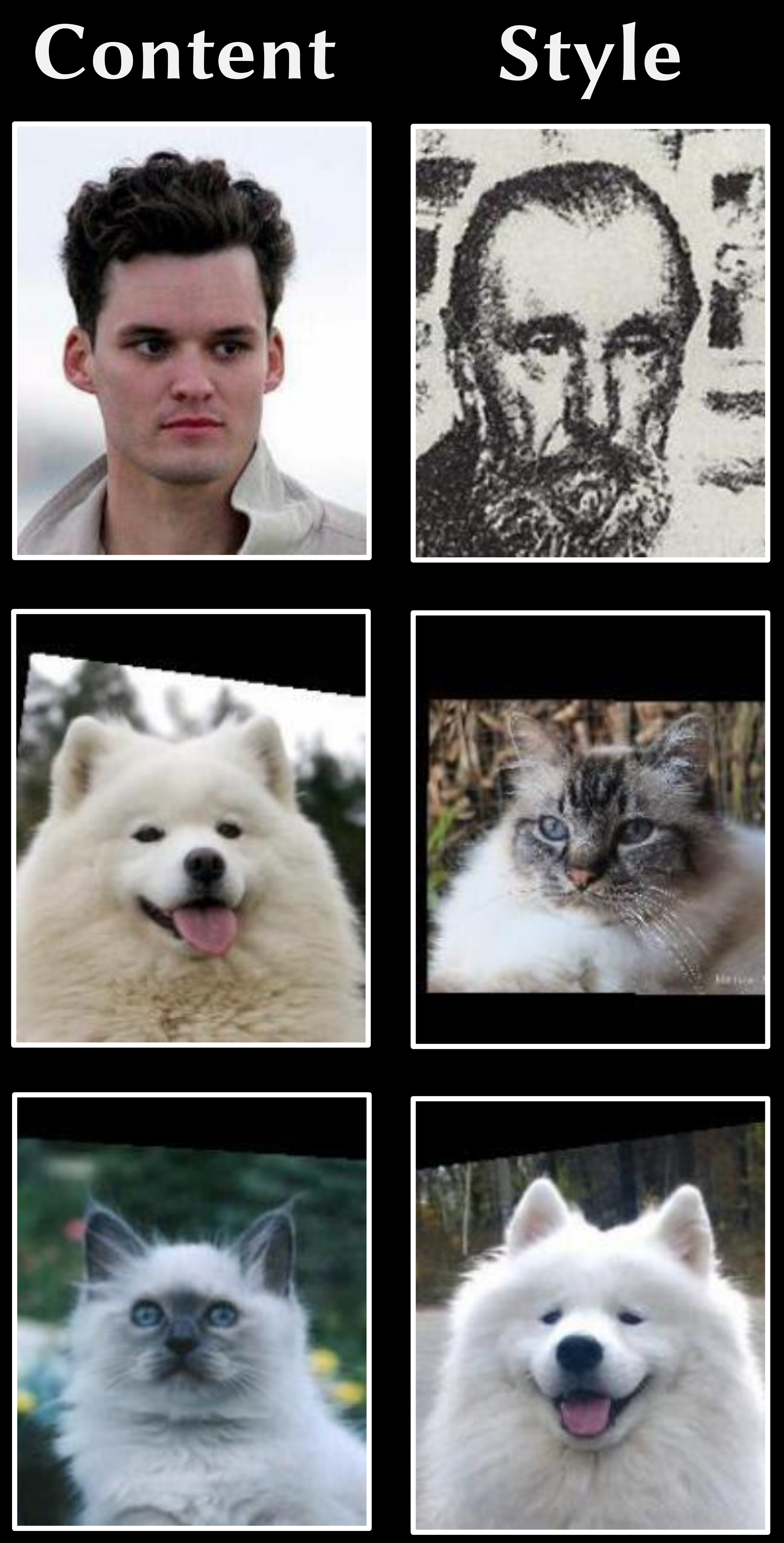
Observation



Most I2I methods make trade-offs between content preservation and style translation.

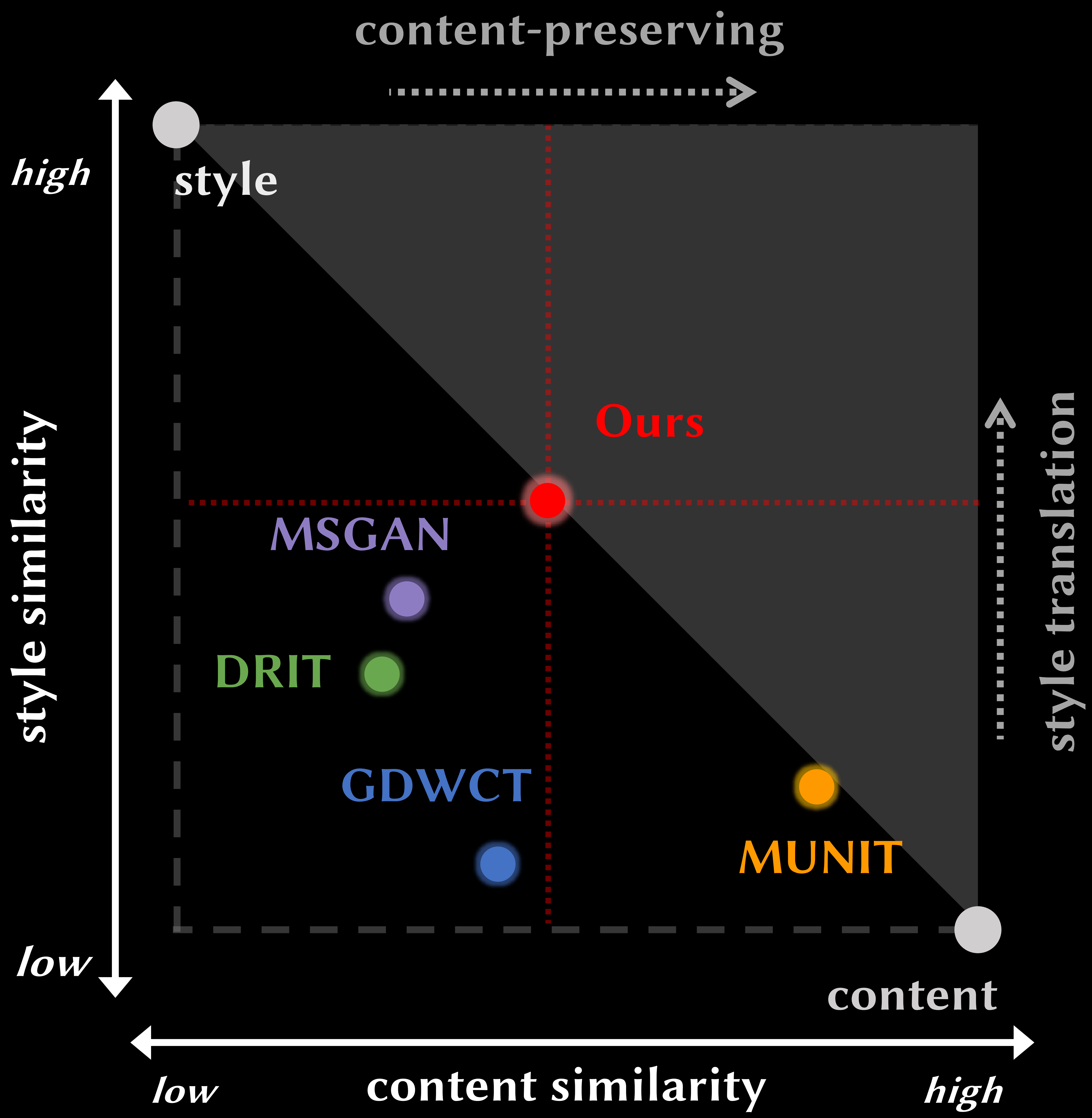
Observation





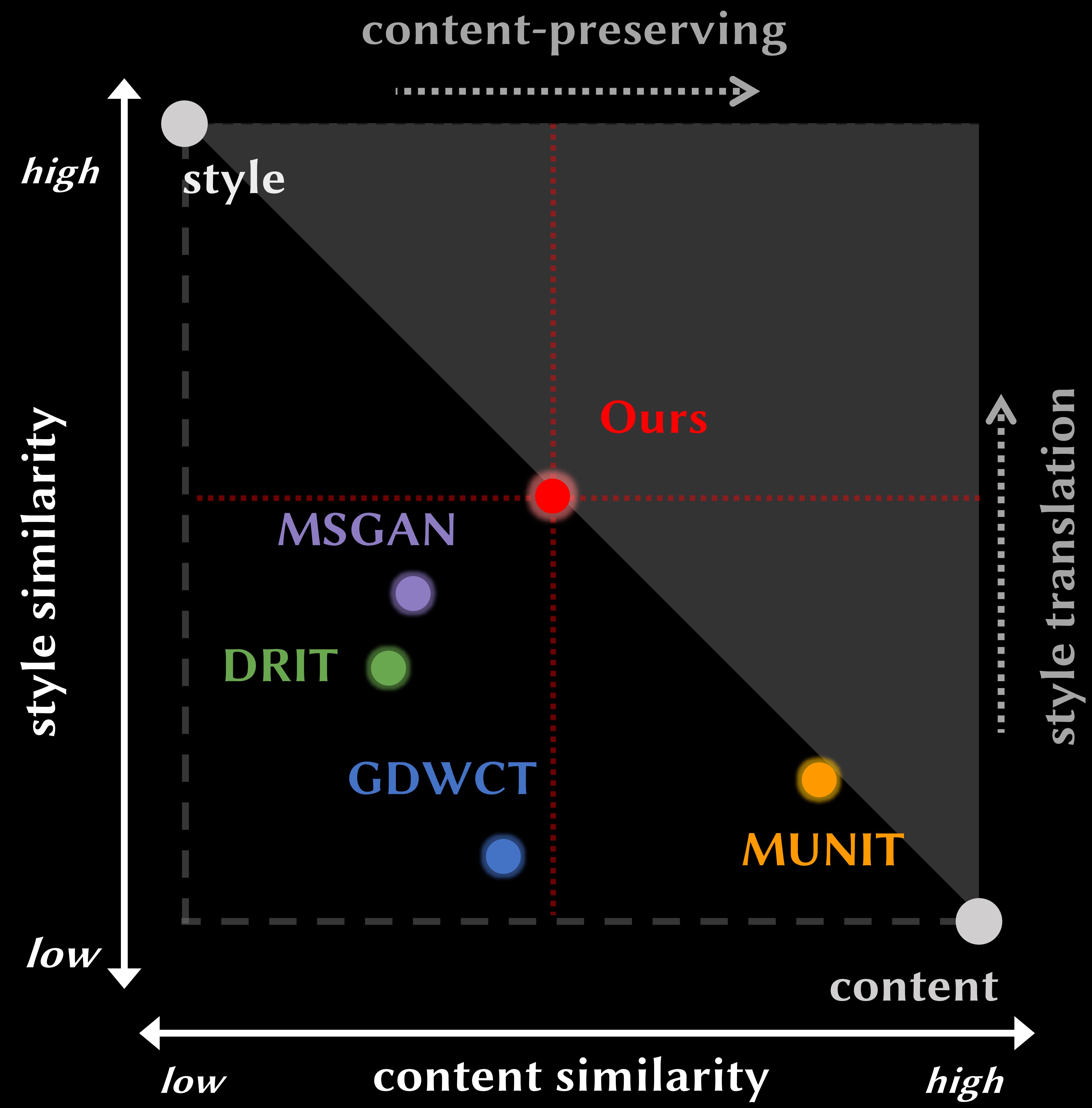
We use the figure to classify I2I methods.
The x-axis shows the ability of content-preserving,
and the y-axis shows the ability of style translation.

Observation



1.Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to- image translation. In: ECCV(2018)
2.Cho, W., Choi, S., Keetae Park, D., Shin, I., Choo, J.: Image-to-image translation via group-wise deep whitening-and-coloring transformation. In: CVPR(2019)
3.Mao, Q., Lee, H.Y., Tseng, H.Y., Ma, S., Yang, M.H.: Mode seeking generative adversarial networks for diverse image synthesis. In: CVPR(2019)
4.Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to- image translation via disentangled representations. In: ECCV(2018)

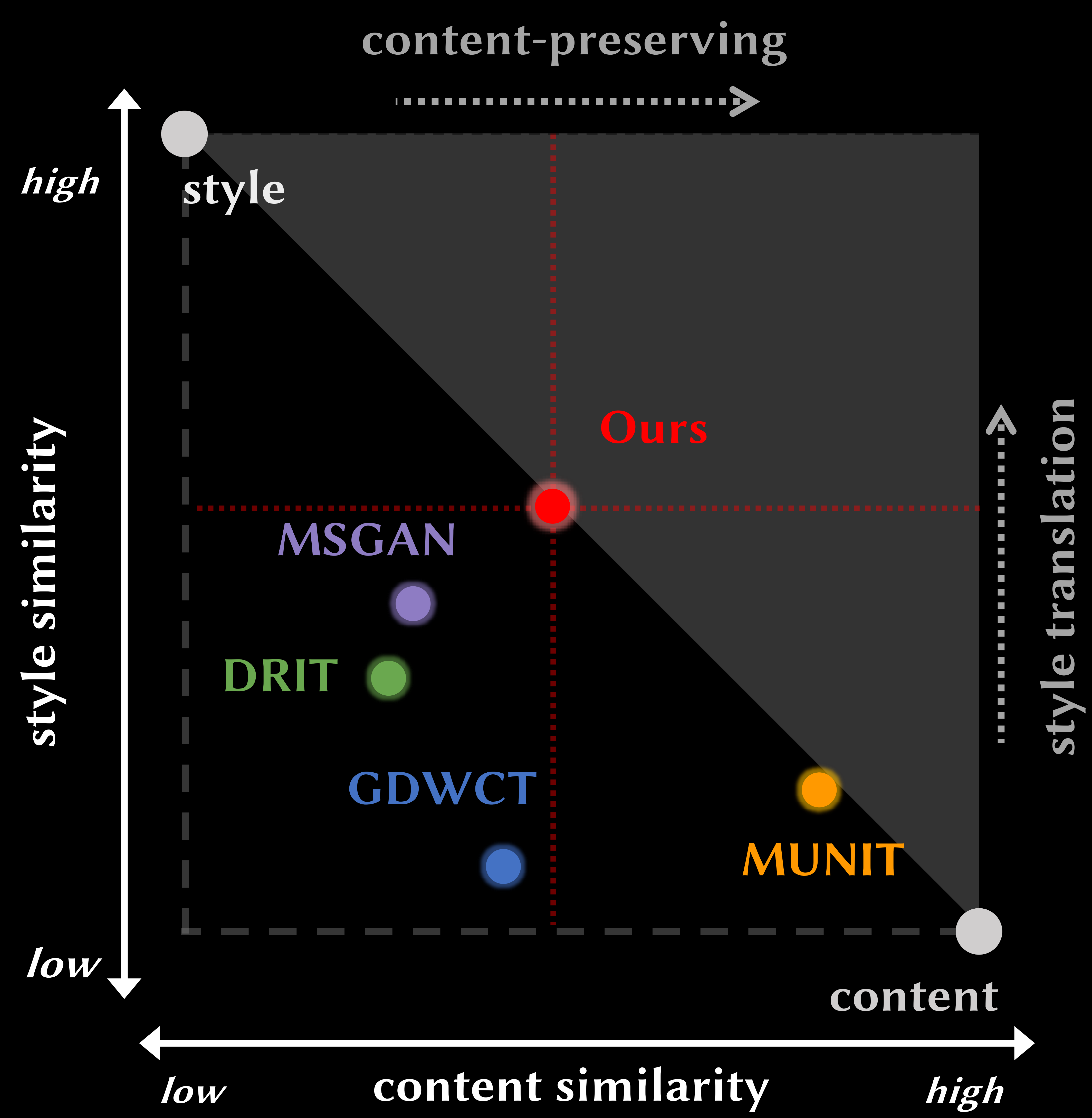
Observation



We can observe that the other I2I methods can only preserve the content information or do style translation.

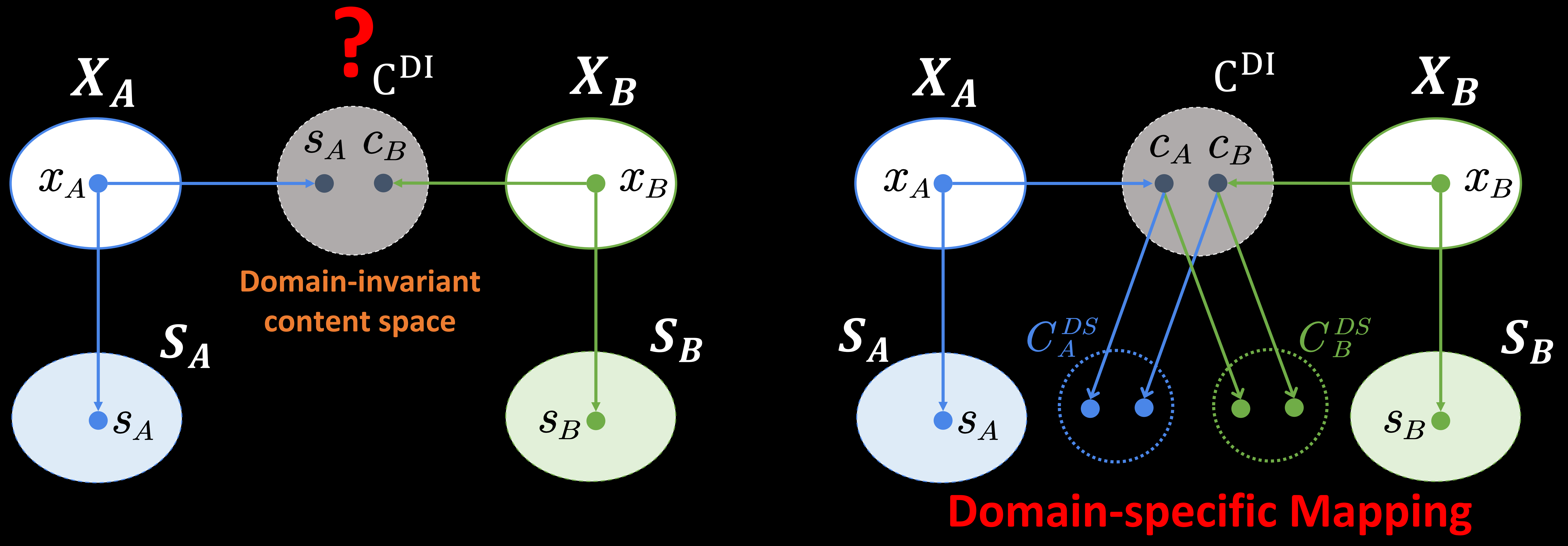
Observation

1.Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to- image translation. In: ECCV(2018)
2.Cho, W., Choi, S., Keetae Park, D., Shin, I., Choo, J.: Image-to-image translation via group-wise deep whitening-and-coloring transformation. In: CVPR(2019)
3.Mao, Q., Lee, H.Y., Tseng, H.Y., Ma, S., Yang, M.H.: Mode seeking generative adversarial networks for diverse image synthesis. In: CVPR(2019)
4.Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to- image translation via disentangled representations. In: ECCV(2018)



The shared domain-invariant content space could compromise the ability to represent content.
Such that their (MUNIT/GDWCT/MSGAN) results are unsatisfactory.

Main Idea

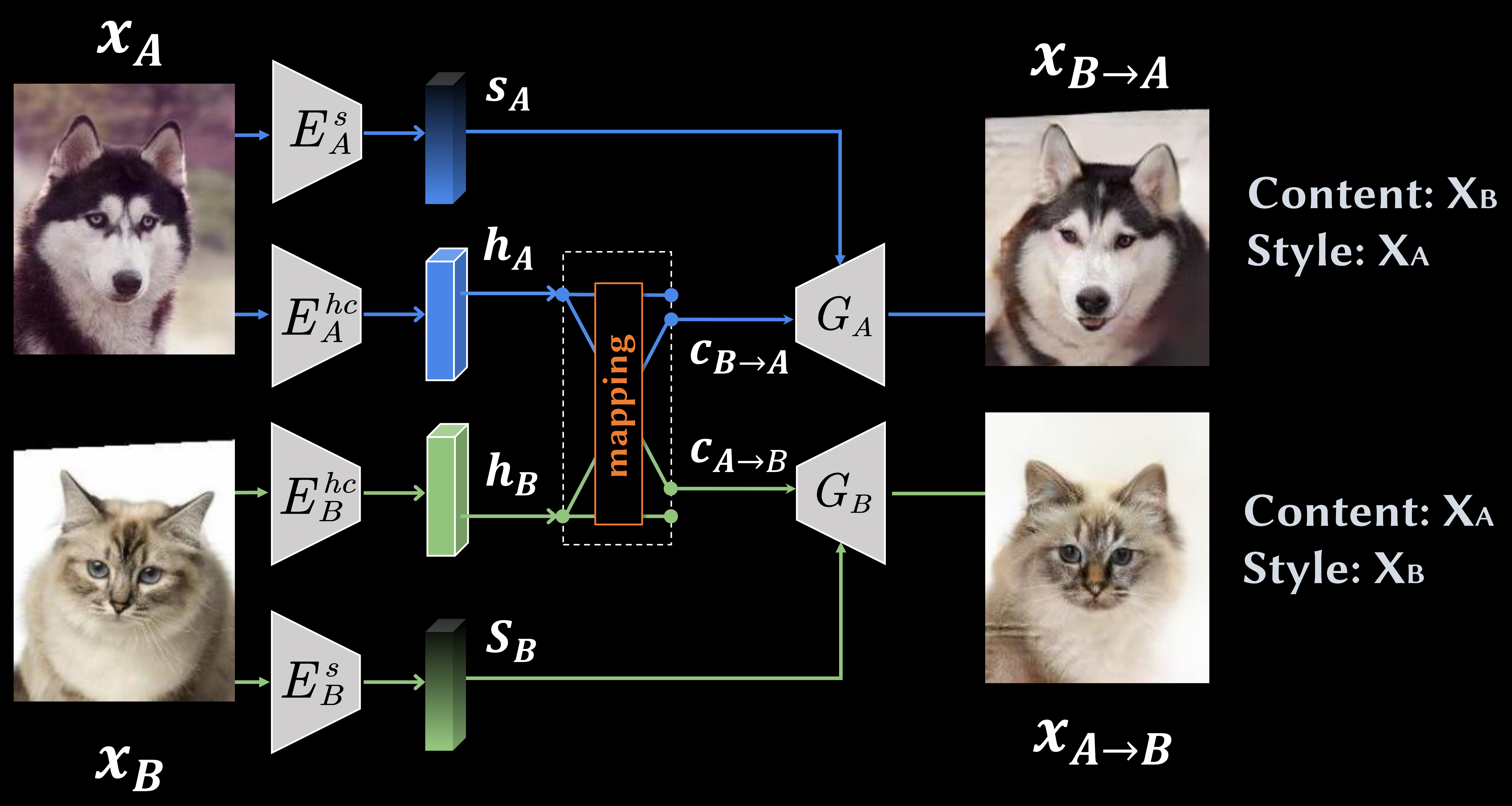


(a) Previous I2I methods (MUNIT/DRIT)

(b) Ours

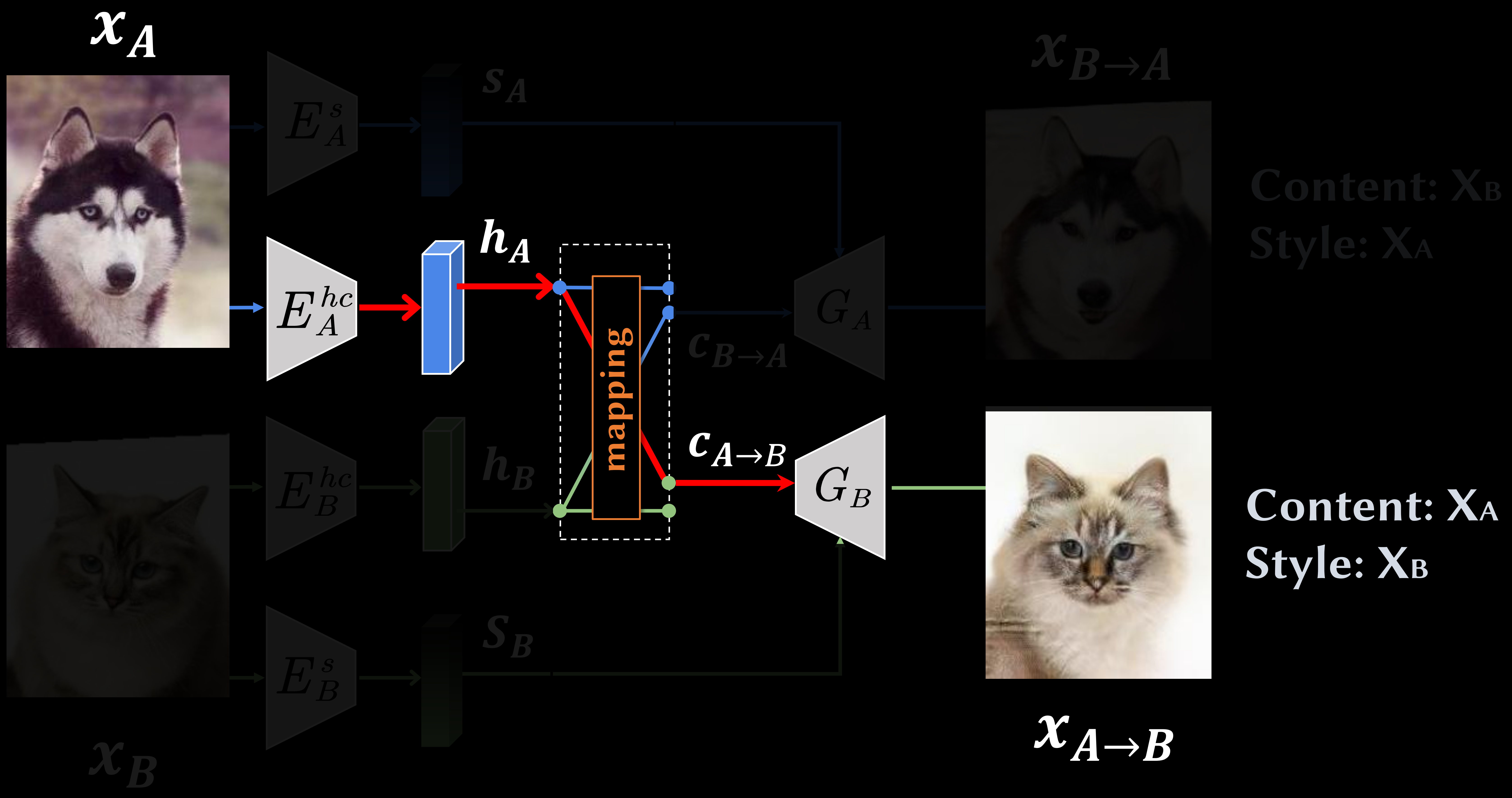
To address the issue, we propose domain-specific mapping functions
to remap the content features in the shared latent space to content spaces for different domains.

Overview



In the training stage, we need to learn the mapping between domains.

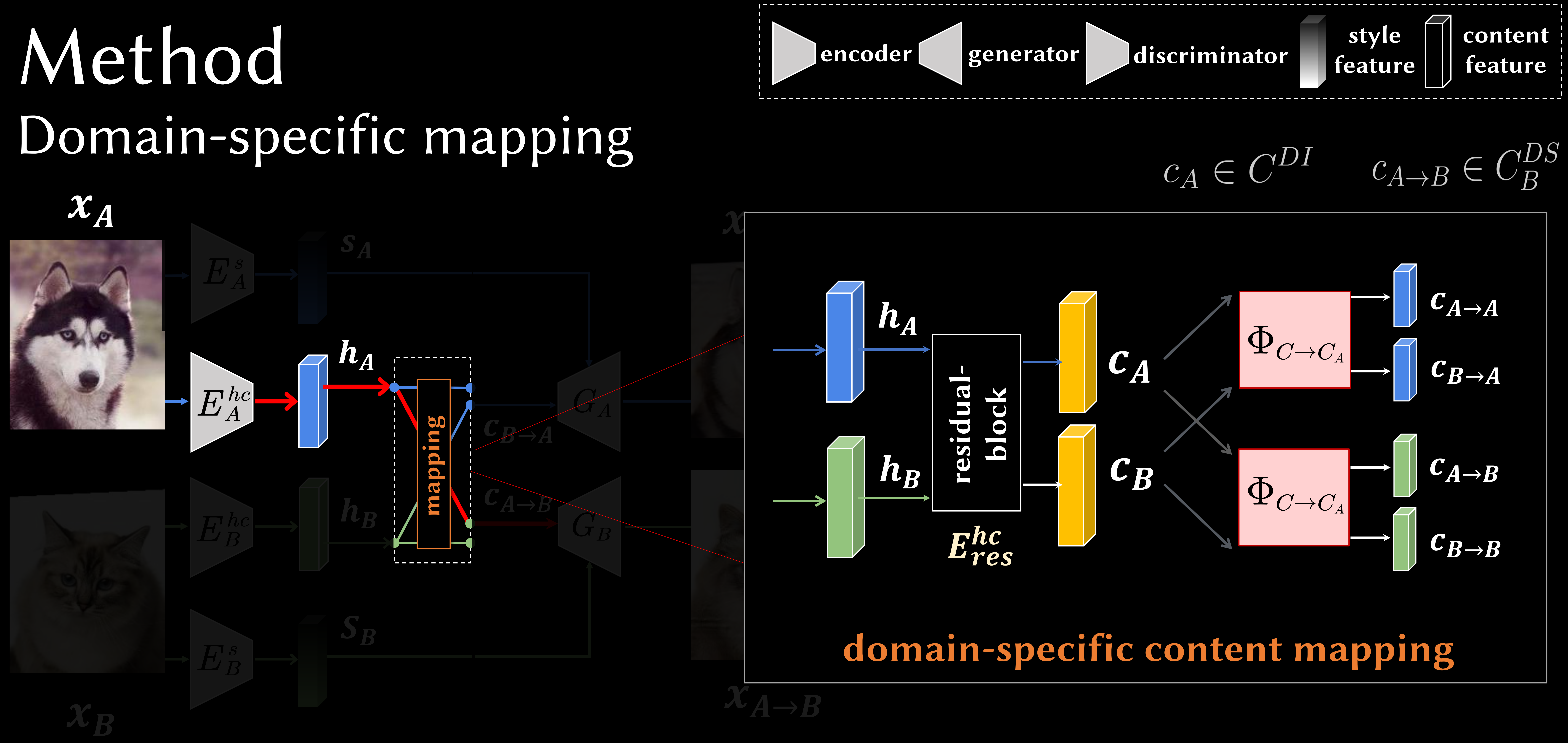
Method



Take the mapping A to B for example. We first encode x_A into a latent content space.

Method

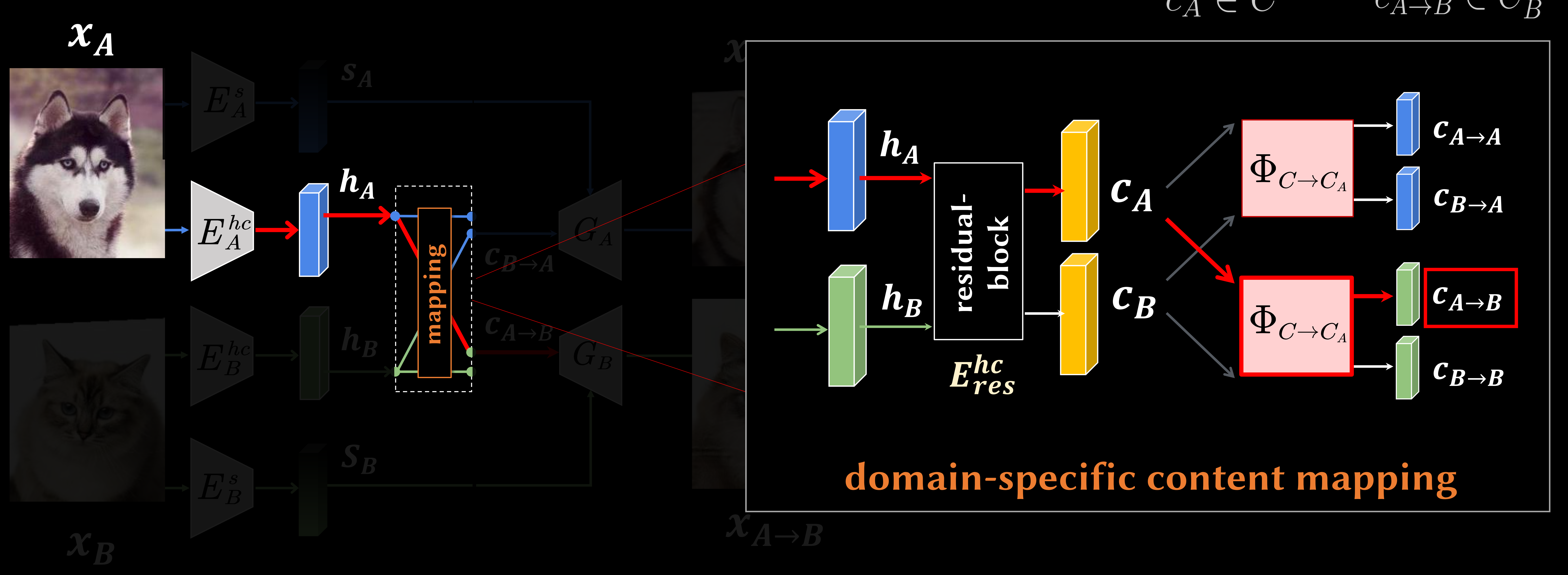
Domain-specific mapping



For the part of domain mapping.

Method

Domain-specific mapping



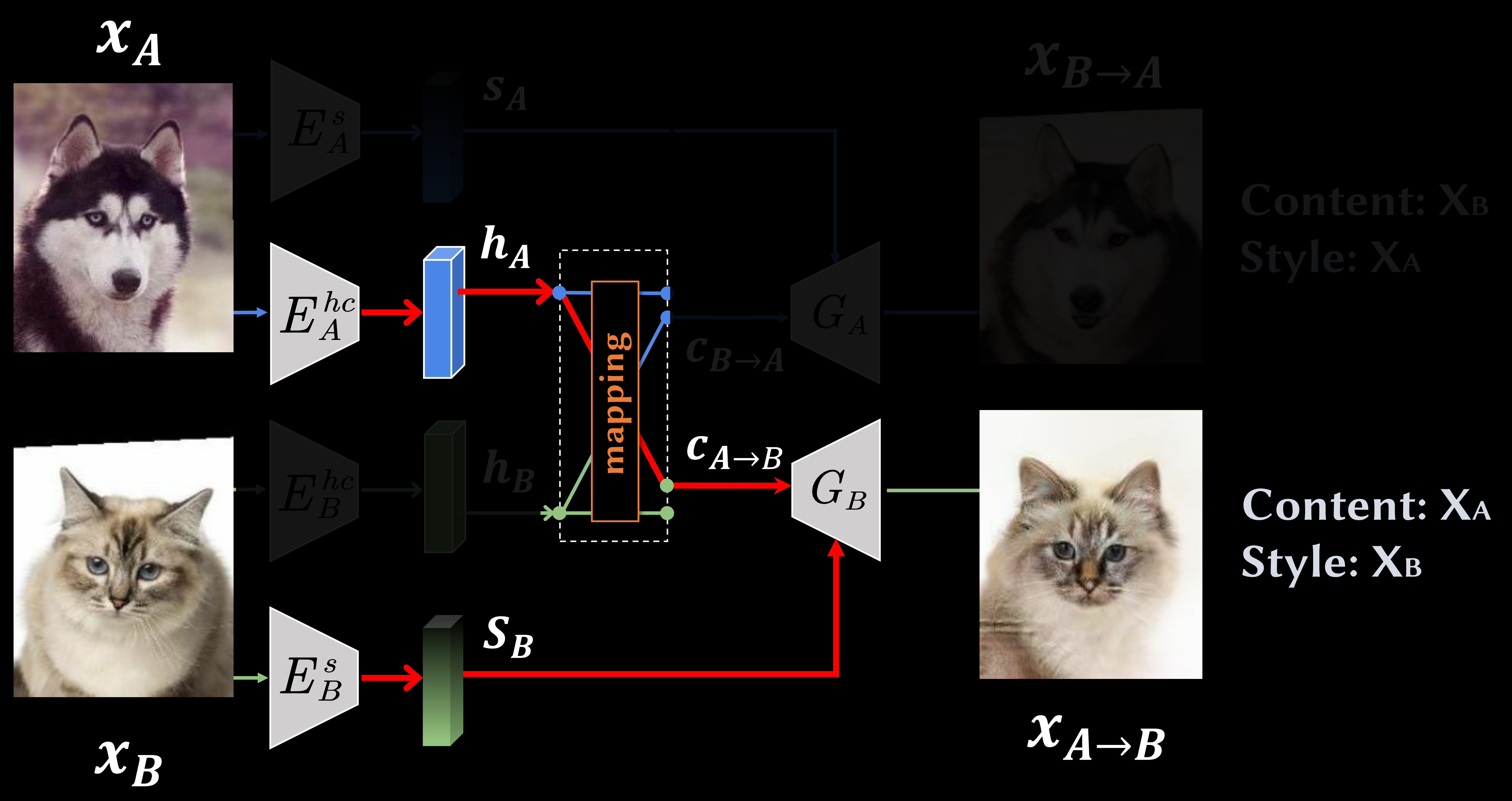
We encode h_A into domain-invariant content space and get c_A,
then use the proposed mapping $\Phi_{C\to C_B}$ to get the content feature in domain B.
In the training stage, we use $c_{A\to B}$ instead of c_A.

Method

Domain-specific mapping

In order to learn the mapping function,
we require that its output resembles the domain-specific content feature h_A and h_B.
Thus we have the domain-specific content loss.

Method



Then, combine with the style feature encoded from x_B, we can get the generated result that preserves content information in x_A while performing style translation of x_B.

Results

Results

Style 1

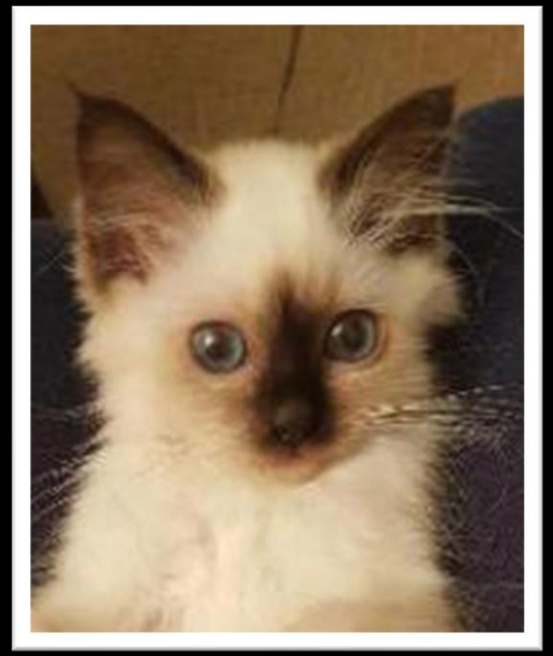
Style 2

Content

DOG → CAT

Style 4

Style 3



Here, we show the translated results of the task dog → cat.

Results

Content

Dog → Cat

Style 1

Result

Result

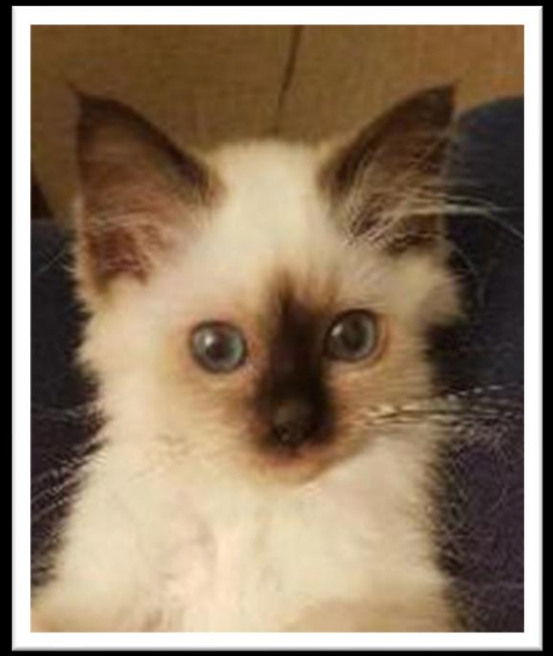
Style 2

Style 4

Result

Result

Style 3



Results

Style 1

Result

Result

Style 2

Content

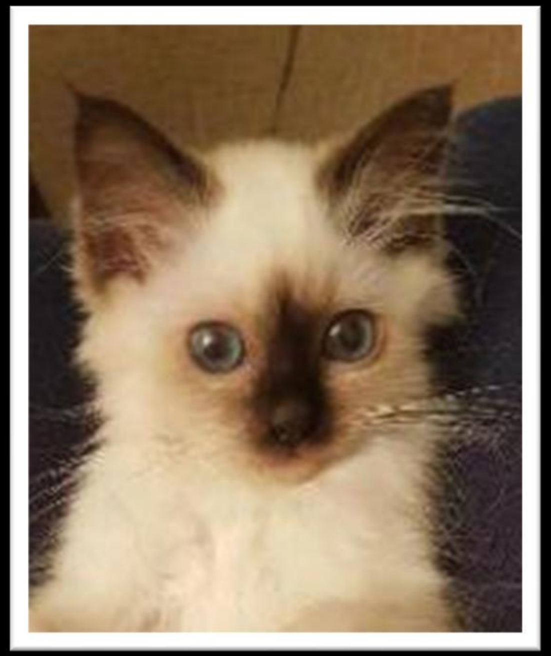
DOG ⟶ CAT

Style 4

Result

Result

Style 3



Results

Content Result

Cat $\longrightarrow$ Dog

Content Result

Dog $\longrightarrow$ Cat

Content Result

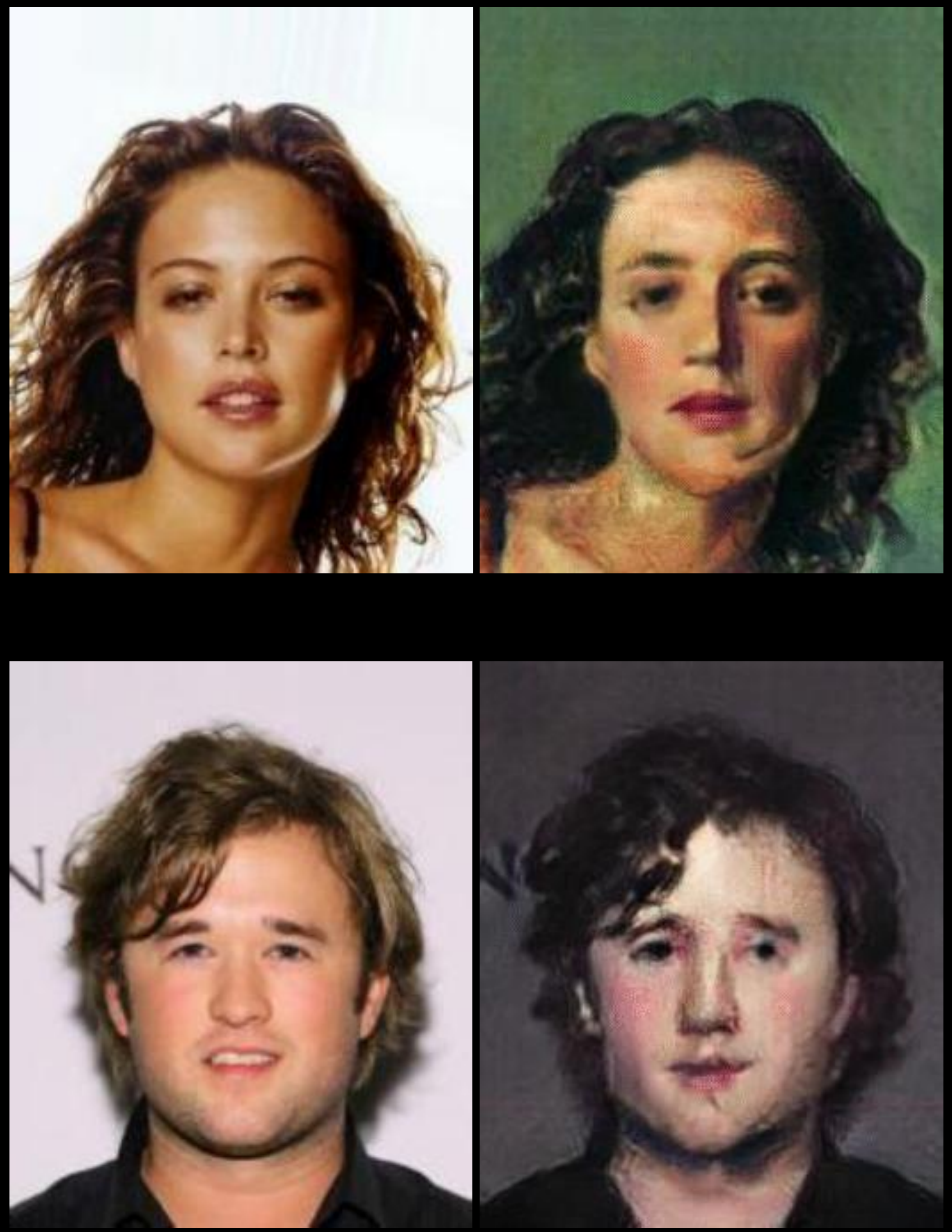


Photo $\longrightarrow$ Portrait

Content Result

Photo $\longrightarrow$ Monet

There are more latent interpolated results on different tasks when given different style images.

Comparisons

We compare our method with three I2I translation methods and three style transfer methods.

Monet ↔ Photo

Content | Style | MUNIT | GDWCT | MSGAN | AdaIN | Liao et al | Luan et al. | Ours

Cat ↔ Dog

Content
Style
MUNIT
GDWCT
MSGAN
AdaIN
Liao et al.
Luan et al.
Ours

Photograph ↔ Portrait

The results of MUNIT and GDWCT have the same problem
that the characteristics of the species are not clear.
MSGAN generates images with more obvious characteristics of the target species.
However, it does not preserve the content information.

Our method generates much clearer results that
better exhibit the characteristics of target species and preserve layouts of the content images.

The style transfer methods have poor performance
due to the different assumption of styles and the use of less information.

The compared style transfer methods cannot perform cross-domain style transfer well.
Thus, we only include image-to-image translation methods in the quantitative comparison.

Quantitative Comparison- FID & LPIPS

	FID ↓				LPIPS ↑			
	MUNIT	GDWCT	MSGAN	Ours	MUNIT	GDWCT	MSGAN	Ours
Cat ⟶ Dog	38.09	91.40	<u>20.80</u>	**13.60**	0.3501	0.1804	**0.5051**	<u>0.4149</u>
Dog ⟶ Cat	39.71	59.72	<u>28.30</u>	**19.69**	0.3167	0.1573	**0.4334**	<u>0.3174</u>
Monet ⟶ Photo	<u>85.06</u>	113.16	86.72	**81.61**	<u>0.4282</u>	0.2478	0.4229	**0.5379**
Photo ⟶ Monet	77.85	<u>71.68</u>	80.37	**63.94**	0.4128	0.2097	<u>0.4306</u>	**0.4340**
Portrait ⟶ Photo	93.45	83.69	**57.07**	<u>62.44</u>	0.1819	0.1563	<u>0.3061</u>	**0.3160**
Photo ⟶ Portrait	89.97	75.86	<u>57.84</u>	**45.81**	0.1929	0.1785	<u>0.2917</u>	**0.3699**
Avg.	70.69	82.59	<u>55.18</u>	**47.85**	0.3131	0.1881	<u>0.3978</u>	**0.3980**

Red texts indicate the best and <u>blue texts</u> indicate the second best method.

Quantitative Comparison- FID & LPIPS

	FID ↓				LPIPS ↑			
	MUNIT	GDWCT	MSGAN	Ours	MUNIT	GDWCT	MSGAN	Ours
Cat $\longrightarrow$ Dog	38.09	91.40	20.80	**13.60**	0.3501	0.1804	**0.5051**	0.4149
Dog $\longrightarrow$ Cat	39.71	59.72	28.30	**19.69**	0.3167	0.1573	**0.4334**	0.3174
Monet $\longrightarrow$ Photo	85.06	113.16	86.72	**81.61**	0.4282	0.2478	0.4229	**0.5379**
Photo $\longrightarrow$ Monet	77.85	71.68	80.37	**63.94**	0.4128	0.2097	0.4306	**0.4340**
Portrait $\longrightarrow$ Photo	93.45	83.69	**57.07**	62.44	0.1819	0.1563	0.3061	**0.3160**
Photo $\longrightarrow$ Portrait	89.97	75.86	57.84	**45.81**	0.1929	0.1785	0.2917	**0.3699**
Avg.	70.69	82.59	55.18	**47.85**	0.3131	0.1881	0.3978	**0.3980**

Our method often has a significantly lower score than other methods in FID score.
For LPIPS, even if our mapping function is not designed to increase diversity,
our method achieves good diversity and performs very well.

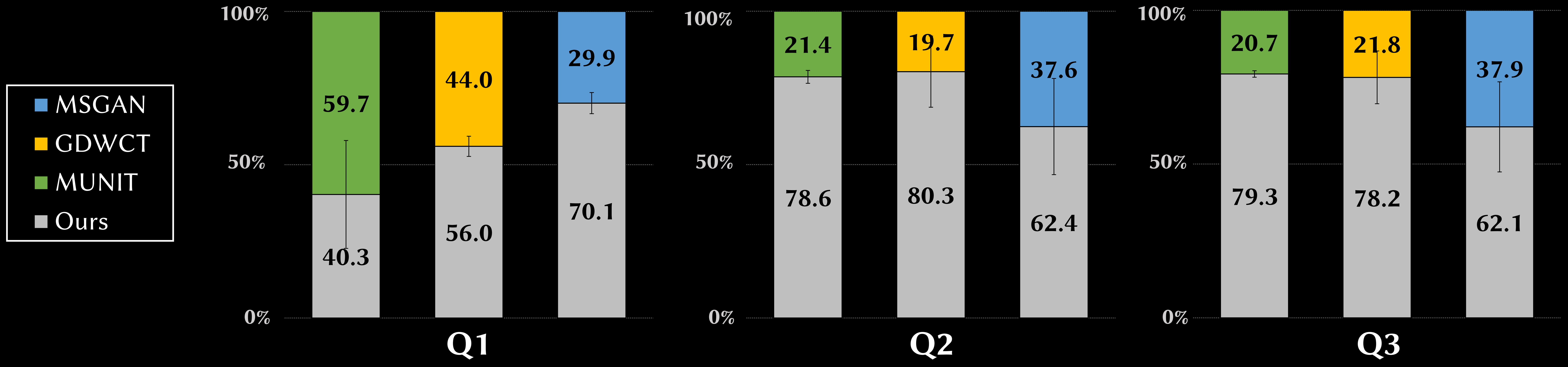
Quantitative Comparison- User study

- User should answer the following questions:
    1. Which one preserves content information (identity, shape, semantic) better?
    2. Which one performs better style translation (in terms of color, pattern)?
    3. Which one is more likely to be a member of the domain B?

For each test set, users are presented with
the content image (domain A), the style image (domain B),
and two result images generated from us and another approach.

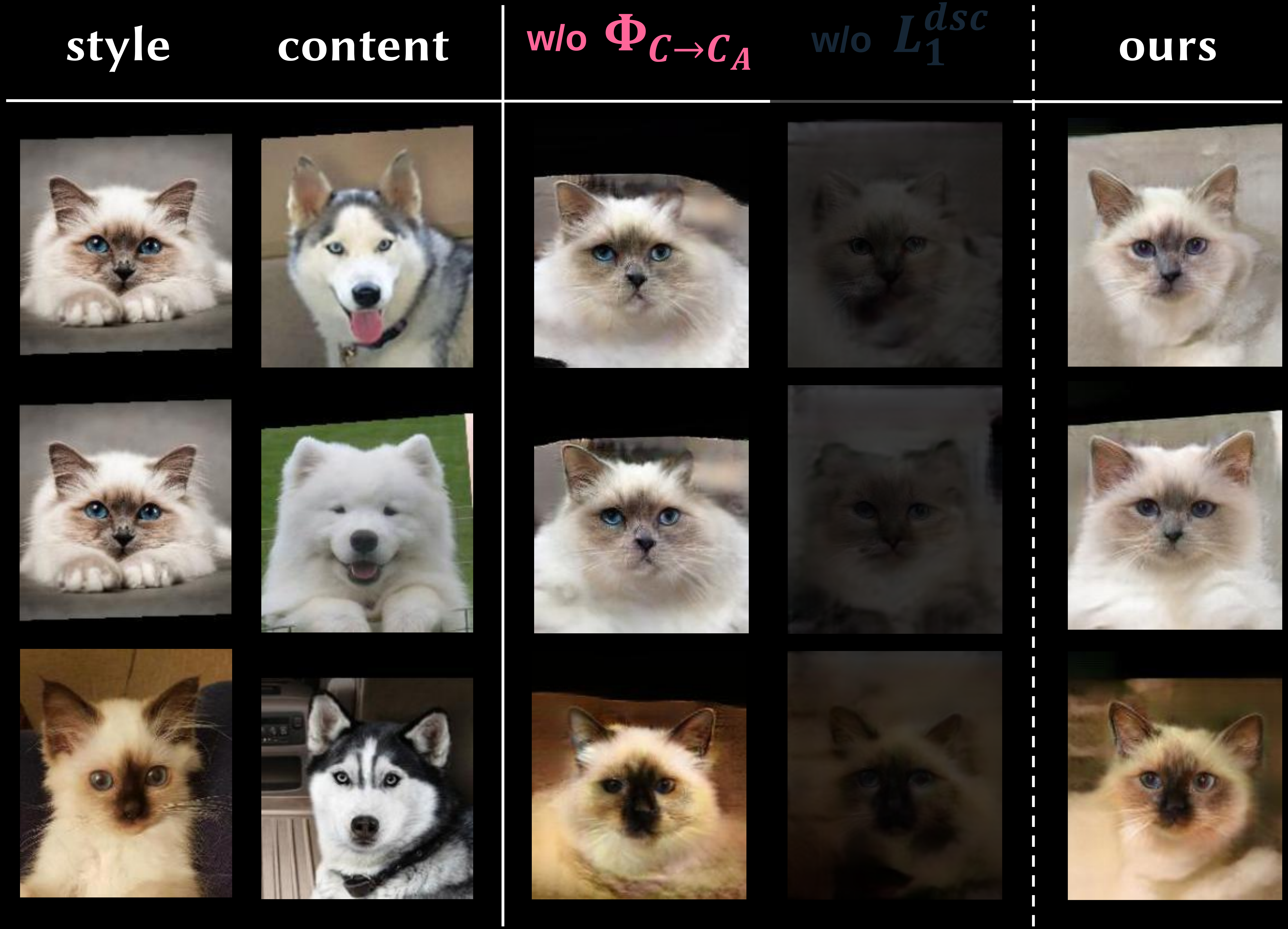
Quantitative Comparison- User study

1. Which one preserves content information (identity, shape, semantic) better?
2. Which one performs better style translation (in terms of color, pattern)?
3. Which one is more likely to be a member of the domain B?



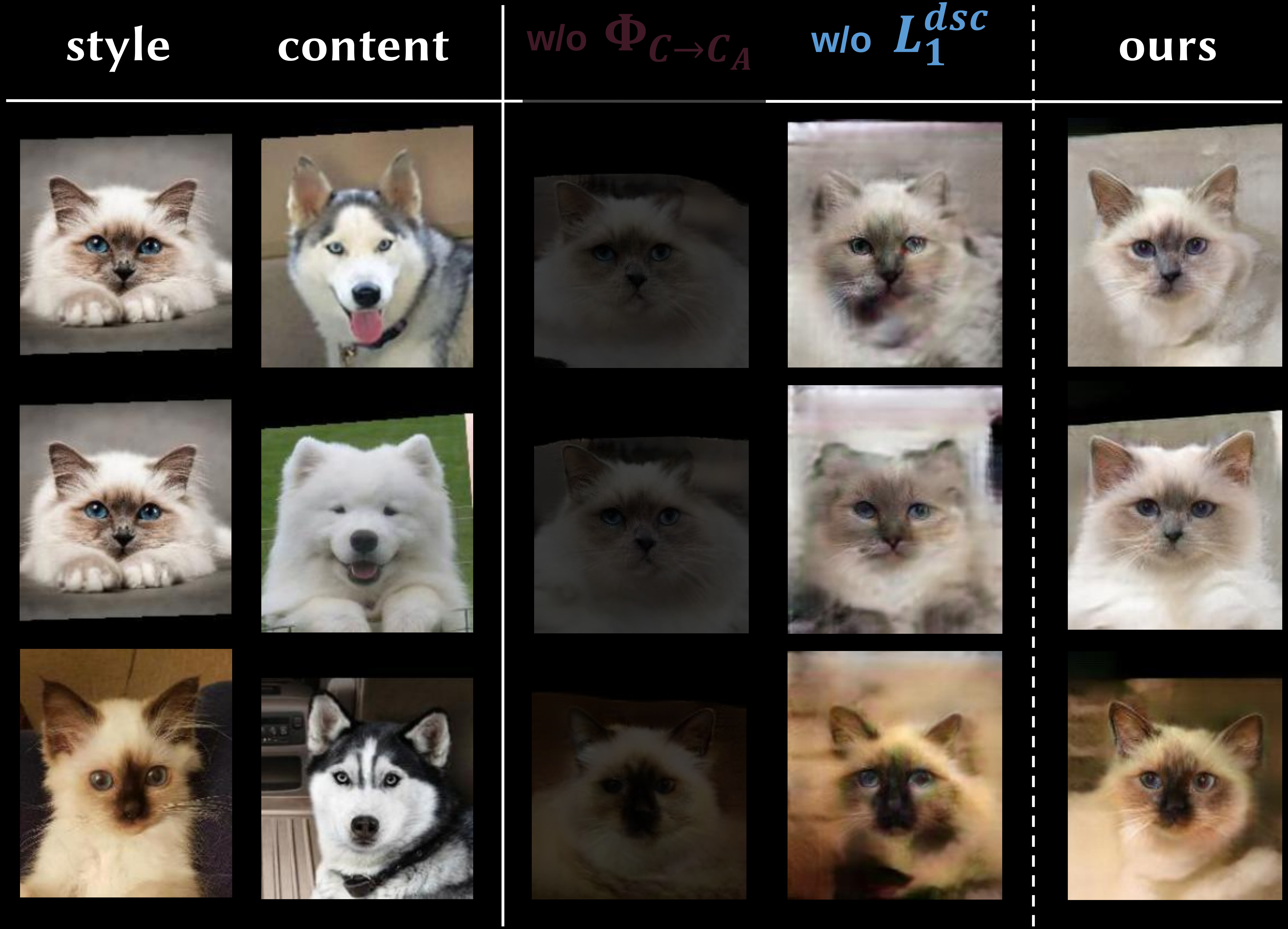
The results show that we can perform style translation well while preserving content information.
Note that MUNIT can preserve content well but does very little on transferring styles.

Ablation Study



Without $\Phi_{C \to C_A}$ (DS map), the spatial layouts of the content images can not be preserved well.

Ablation Study



The proposed loss L_1^{dsc} (DS loss), ensures the remapped feature resembles the domain-specific feature h_A.

Failure Case

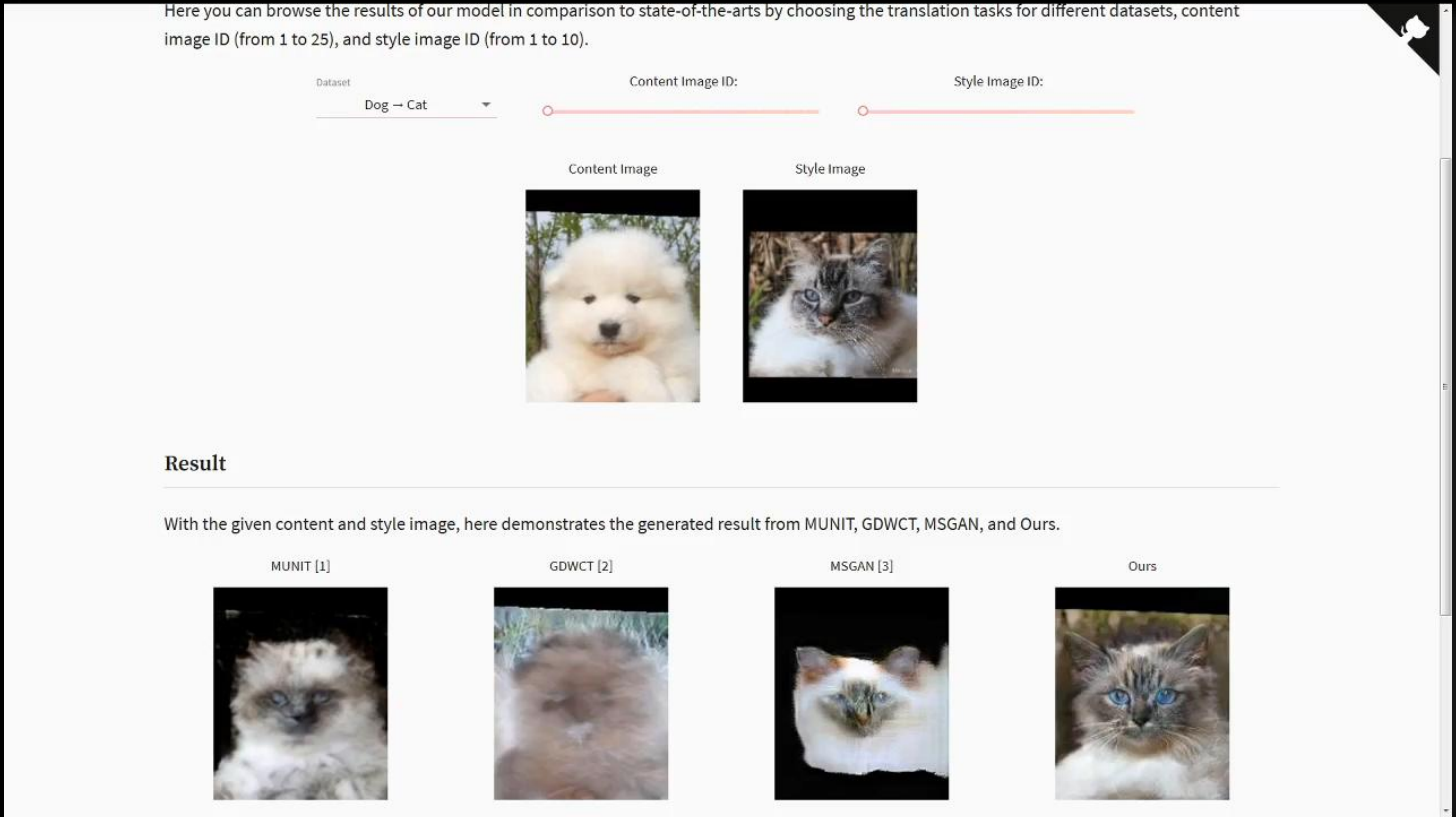
The figure gives examples in which our method is less successful.
In this case, the poses are rare in the training set.
Thus, the content is not preserved as well as other examples.

Failure Case

For the second case, the target domains are photographs.
They are more challenging, and our method could generate less realistic images.
However, our results are still much better than those of other methods.

Code & Demo page

https://acht7111020.github.io/DSMAP-demo/



More results can be found in our website and Github page!

Thanks!

Audio from Google Text-to-Speech!
https://cloud.google.com/text-to-speech