Single-Frame Unsupervised Video-To-Video Translation

Wenjing Wang, Jiazhan Feng

Deep Generation Model

OUTLINE

- Motivation
- Proposal
- Experiment
- Discussion

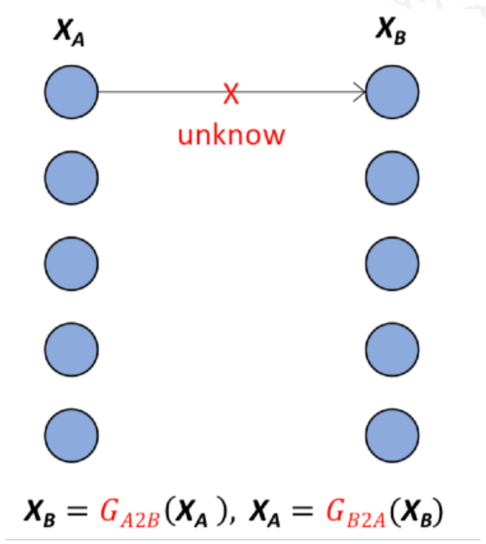
OUTLINE

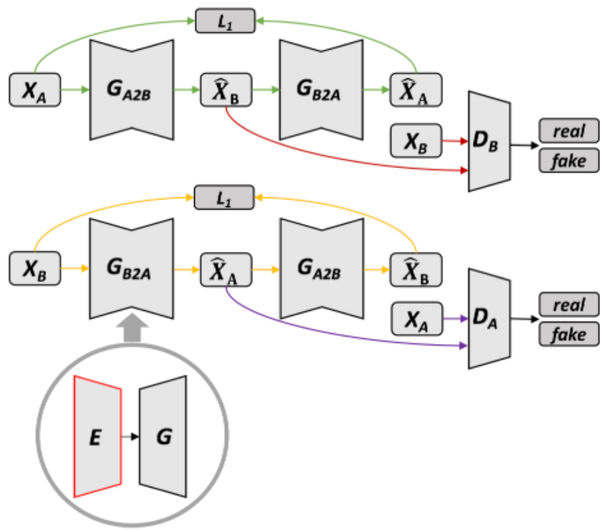
- ➤ Motivation
- ➤ Proposal
- ➤ Experiment
- ➤ Discussion

MOTIVATION

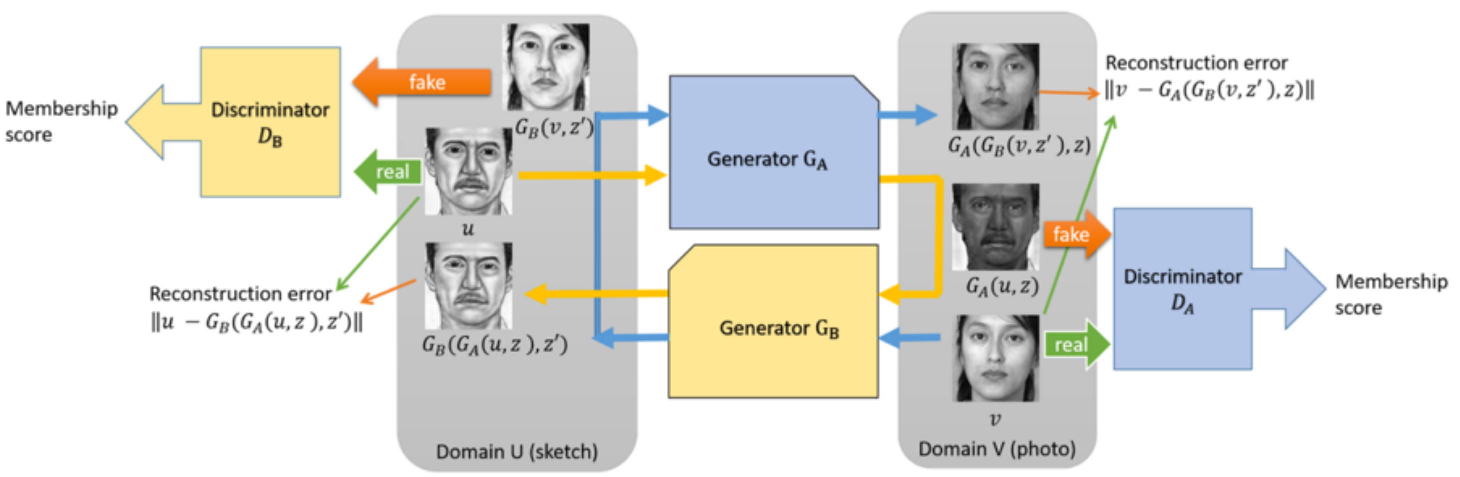
➤ Unsupervised image-to-image translation

- Source domain → target domain
- Unsupervised: without seeing any examples of corresponding image pairs





CycleGAN[1]

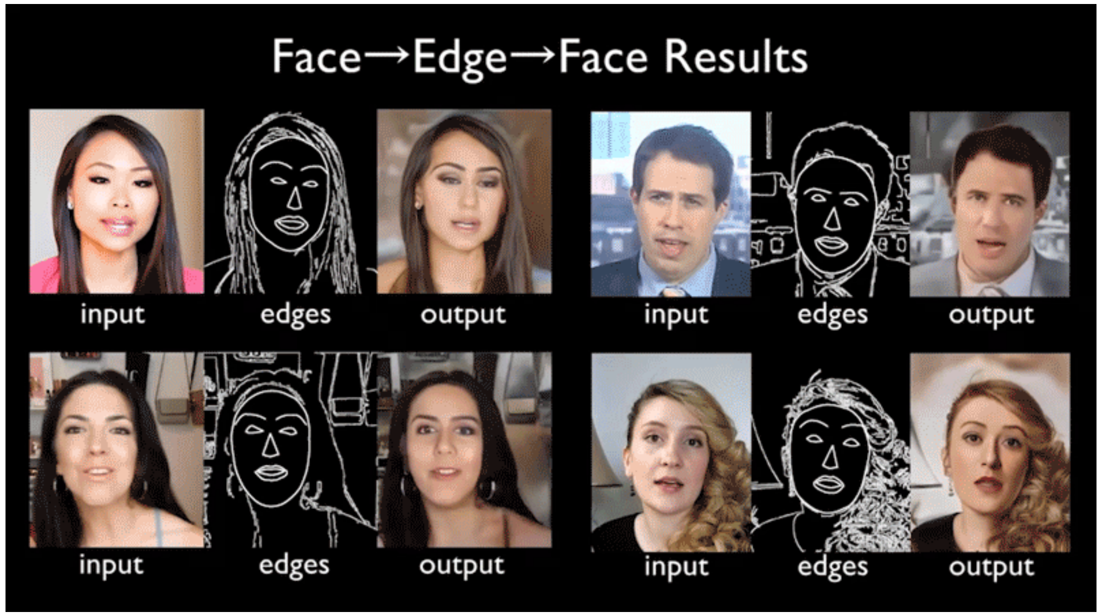


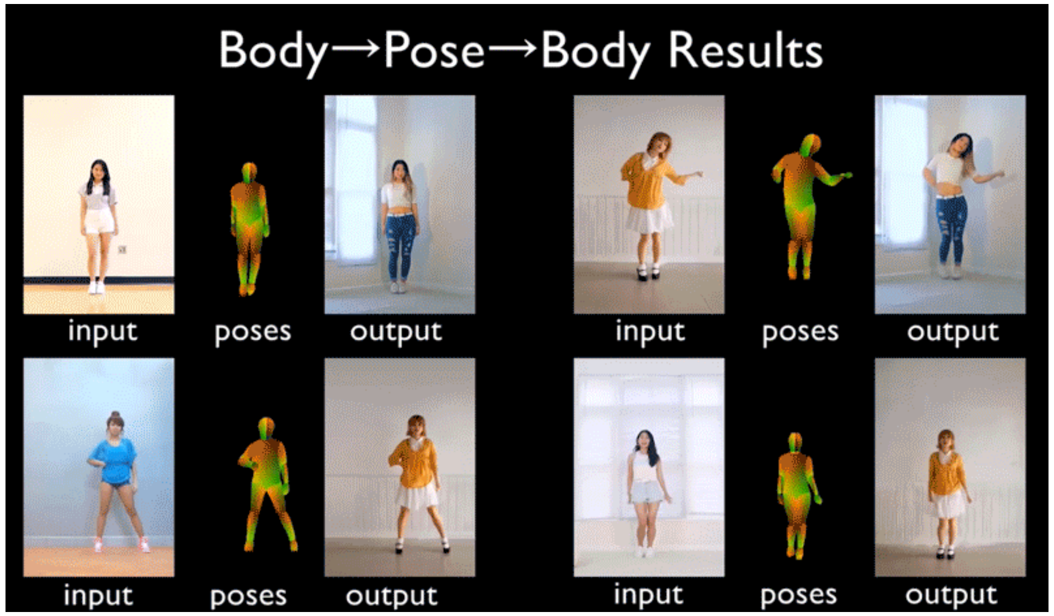
DualGAN[2]

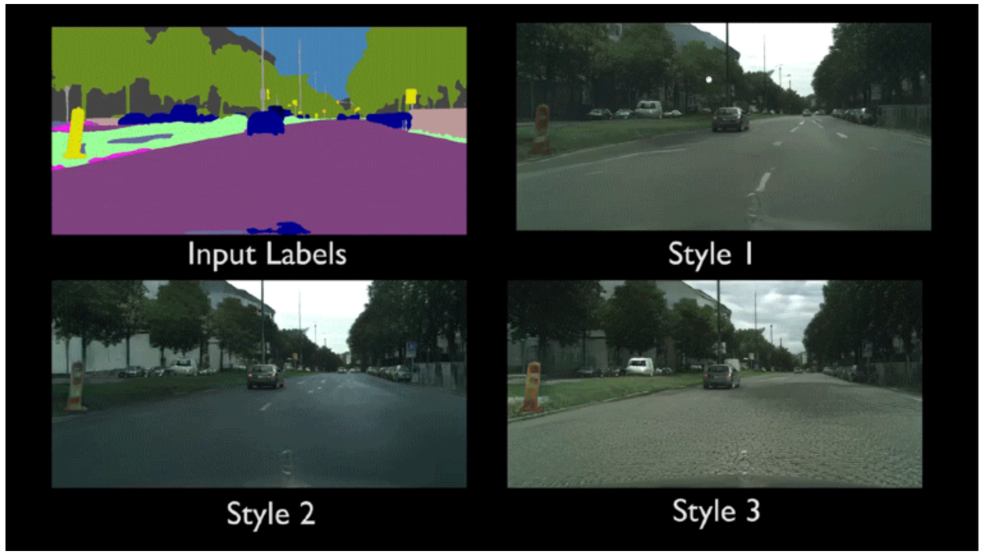
[1] Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks (ICCV 2017)
[2] DualGAN: Unsupervised Dual Learning for Image-to-Image Translation (ICCV 2017)

MOTIVATION

➤ Video-to-video translation





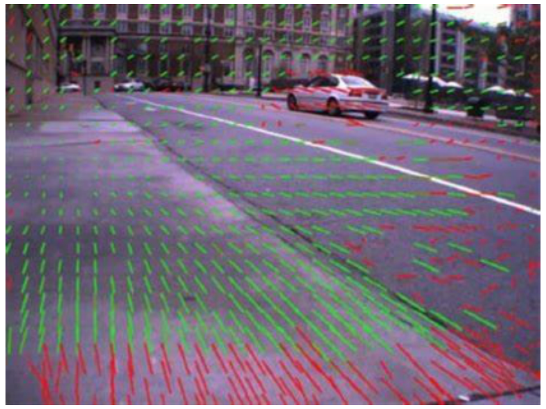


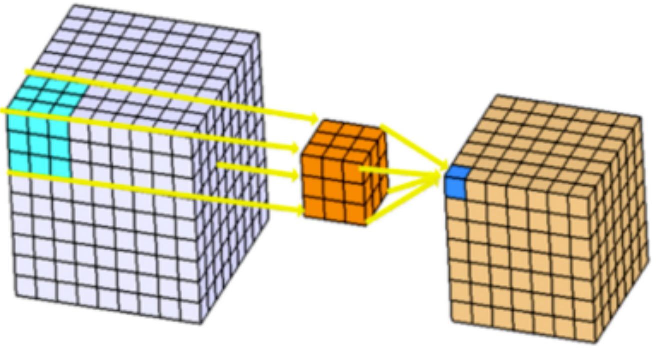
[1] Video-to-Video Synthesis (NIPS 2018)

MOTIVATION

➤ Inter-frame relationship

- Optical flow, 3D-convolutional network





- Drawbacks
  - Computational complexity
  - Require video training data

[1] Video-to-Video Synthesis (NIPS 2018)

MOTIVATION

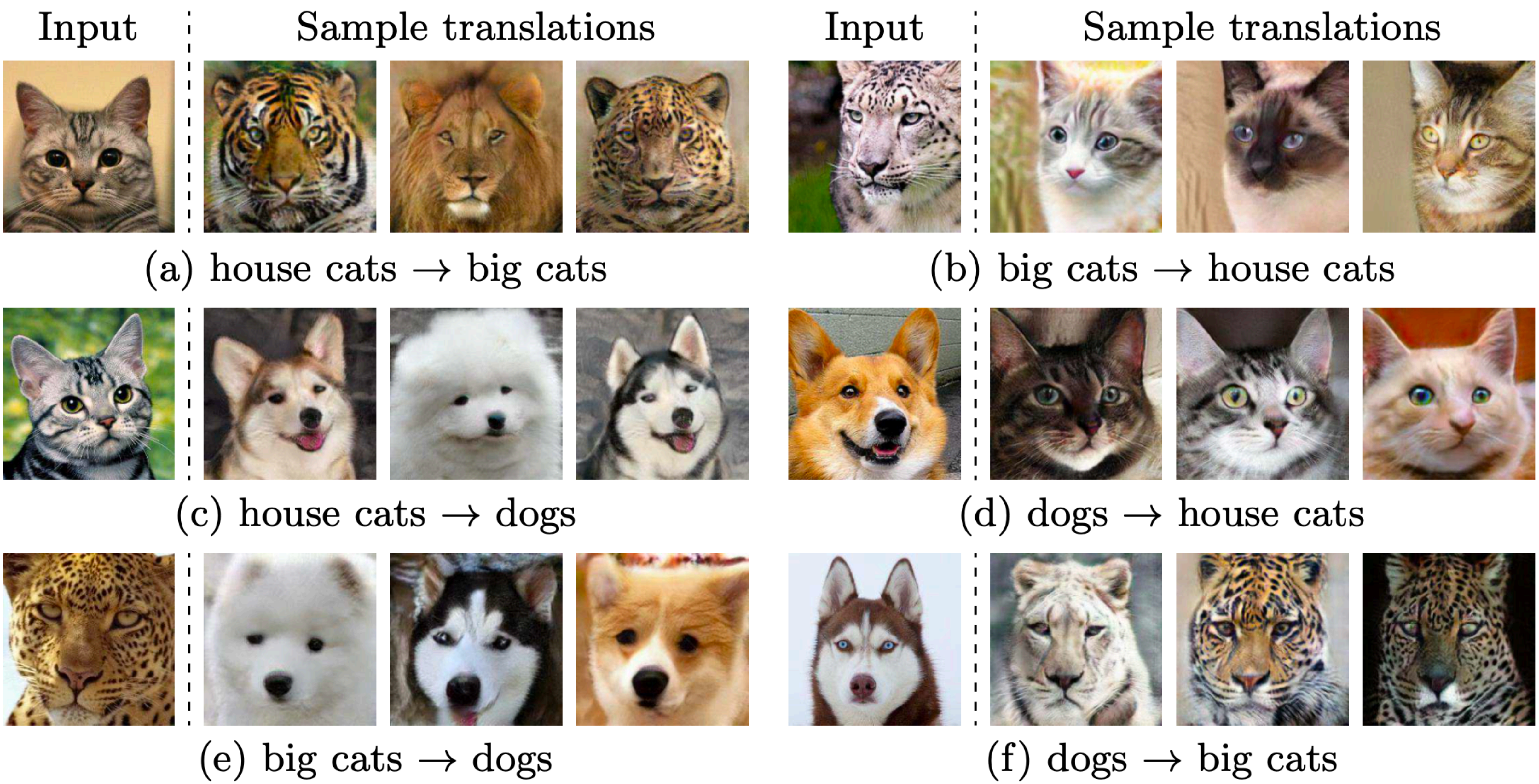
- Single-frame unsupervised video-to-video translation
- Training without video data

OUTLINE

- Motivation
- Proposal
- Experiment
- Discussion

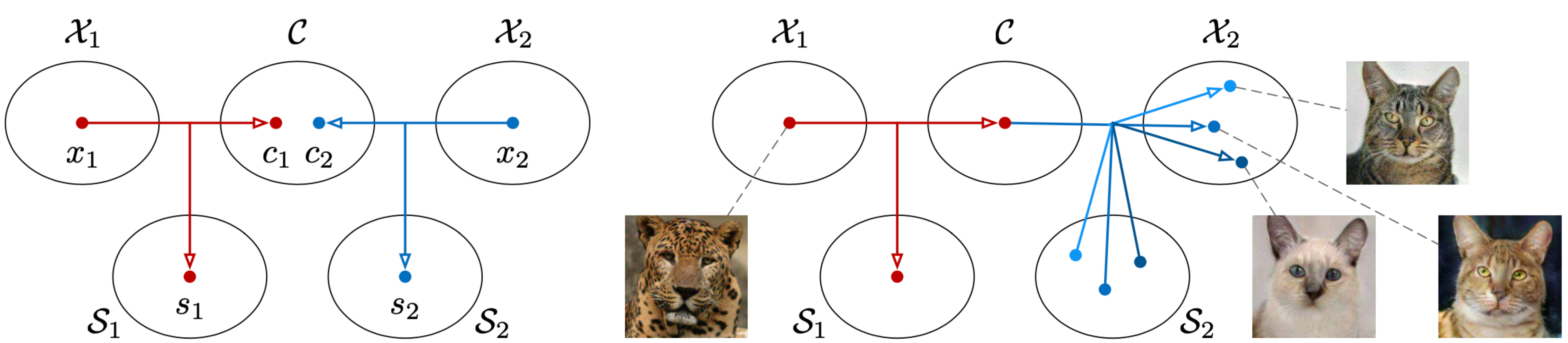
PROPOSAL

- Multimodel Unsupervised Image-to-Image Translation



PROPOSAL

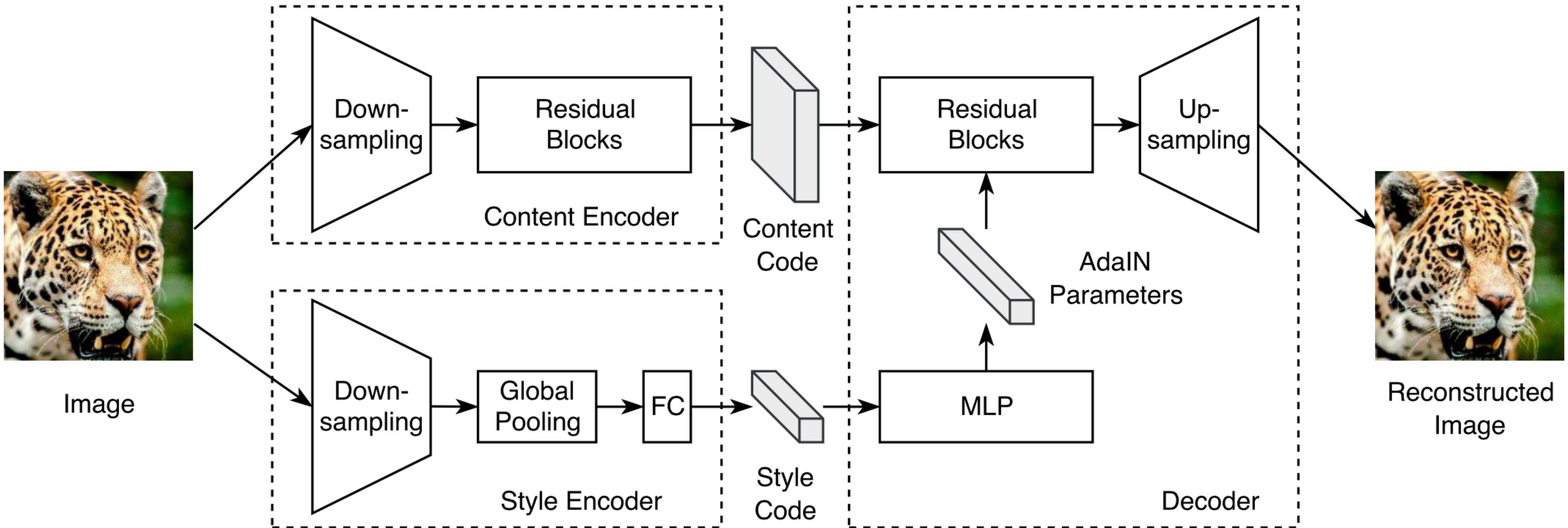
- Multimodel
  - Existing techniques: assume a *deterministic/unimodal* mapping.
  - However, the cross-domain mapping of interest is *multimodal*.
- Assumption
  - The latent space of images = a content space + a style space.



(a) Auto-encoding (b) Translation

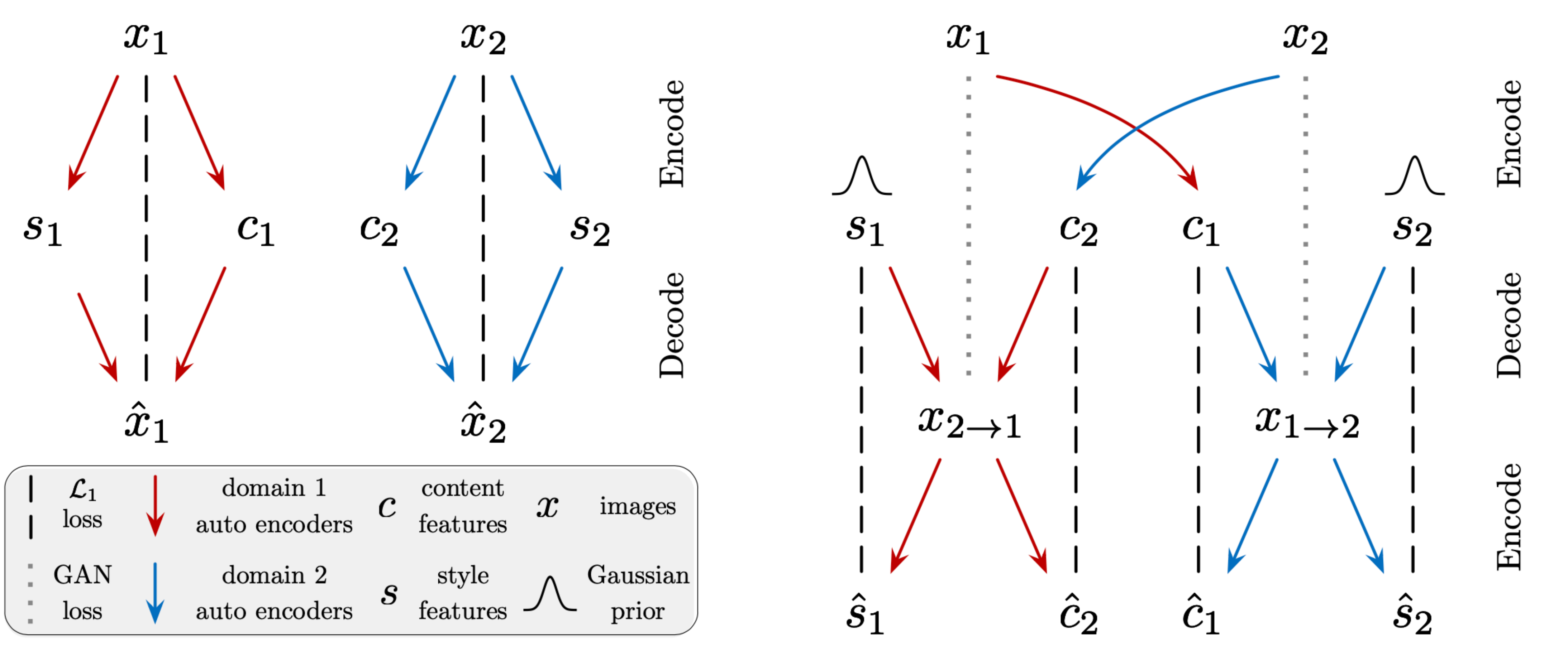
PROPOSAL

➤ Encoder-Decoder + GAN



- Loss = bidirectional reconstruction + adversarial

PROPOSAL

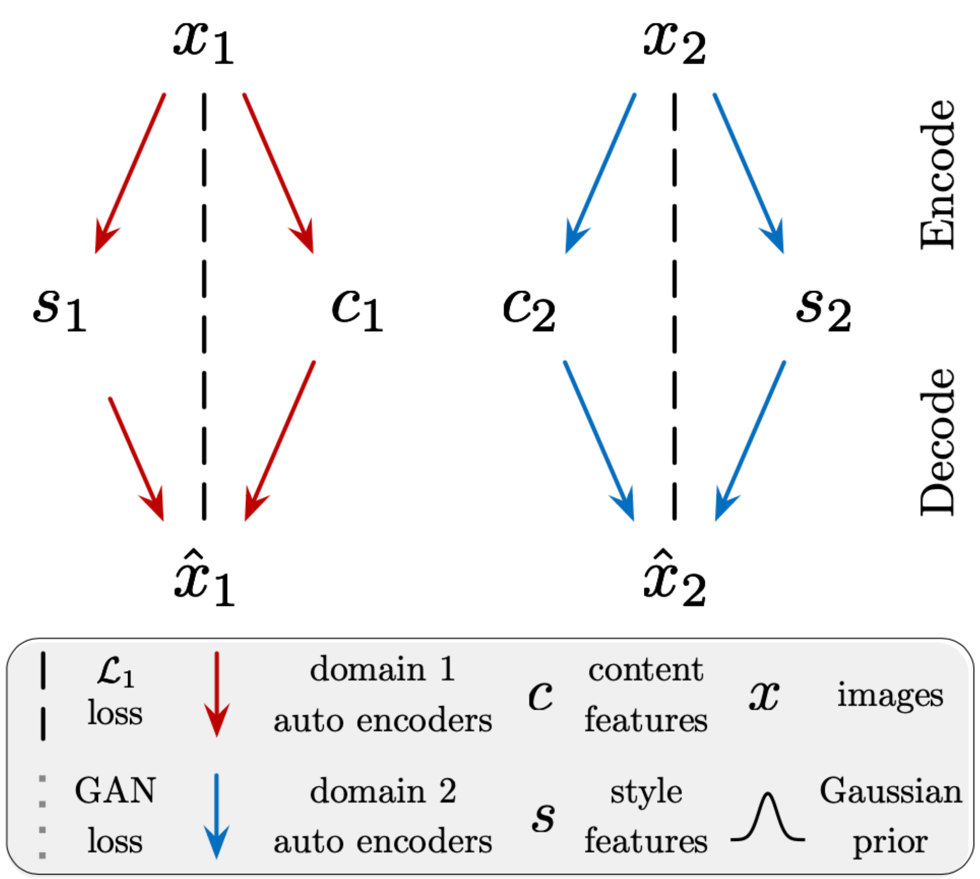


(a) Within-domain reconstruction (b) Cross-domain translation

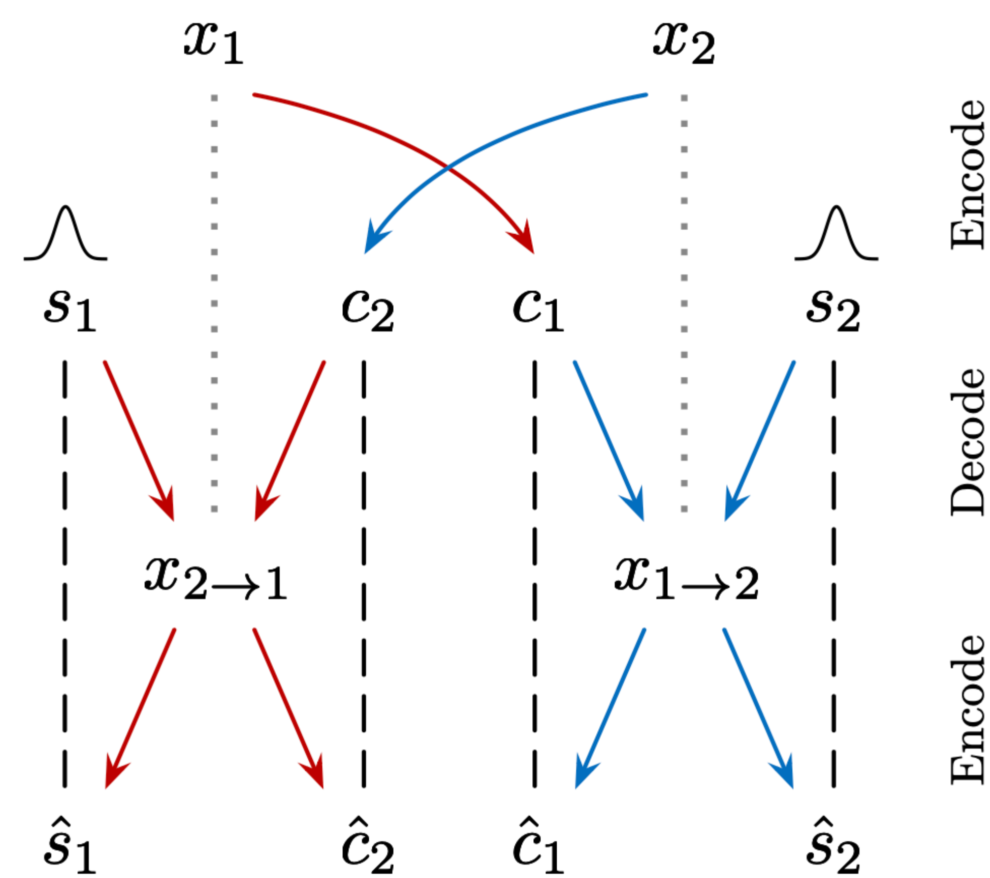
➤ Image Reconstruction

- $L_{recon}^{x_1} = ||\tilde{x}_1 - x_1||, \tilde{x}_1 = G_1(E_1^c(x_1), E_1^s(x_1))$

PROPOSAL



(a) Within-domain reconstruction

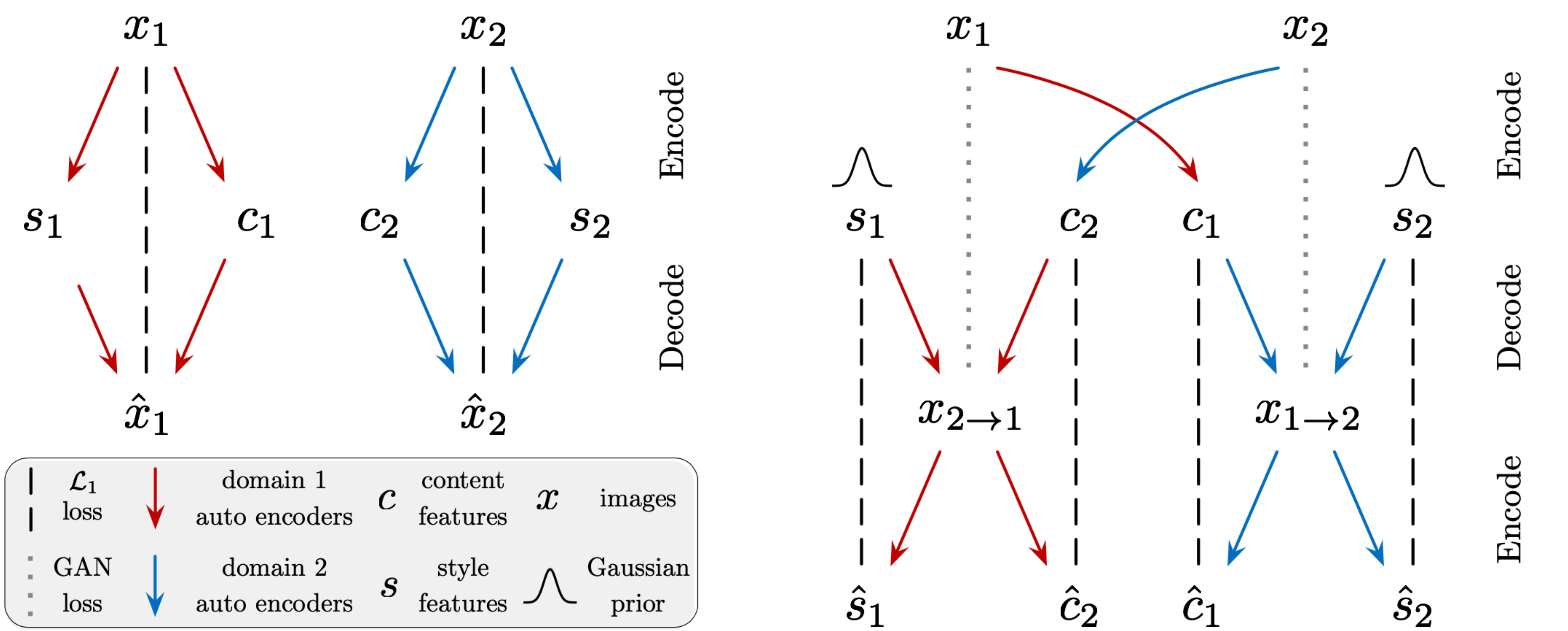


(b) Cross-domain translation

- Latent Reconstruction

  - $L_{recon}^{c_1} = ||E_2^c(G_2(c_1, s_2)) - c_1||$, $L_{recon}^{s_2} = ||E_2^s(G_2(c_1, s_2)) - s_2||$

PROPOSAL



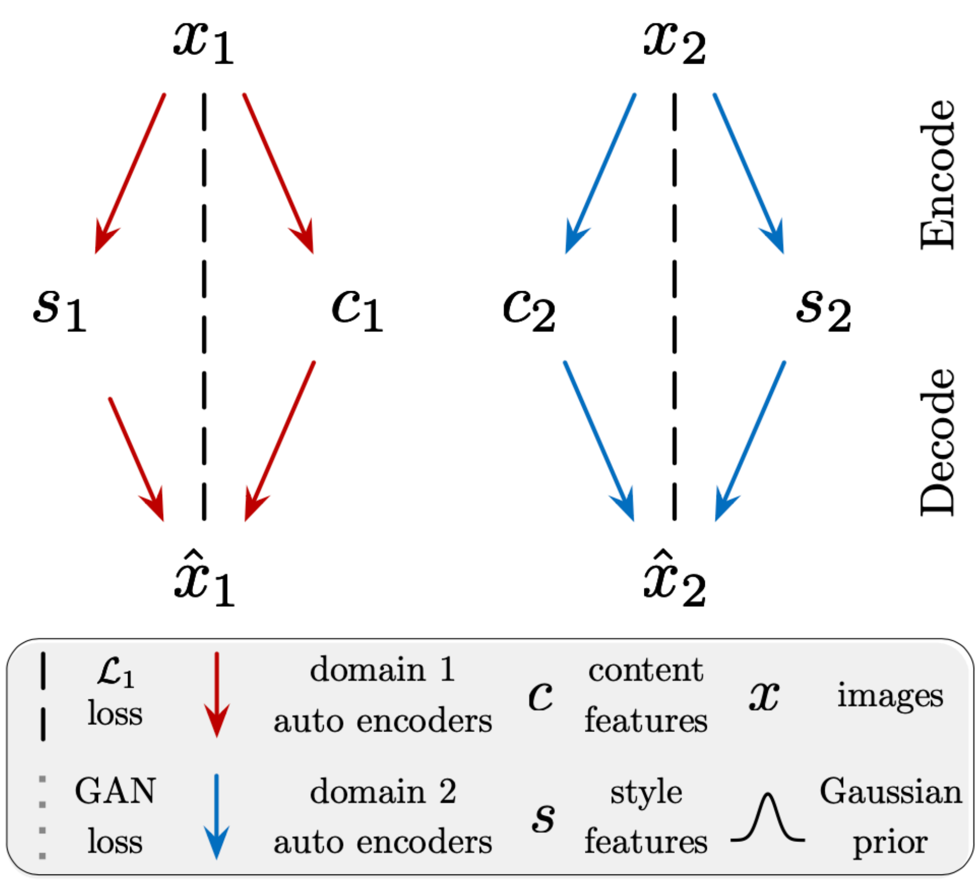
(a) Within-domain reconstruction

(b) Cross-domain translation

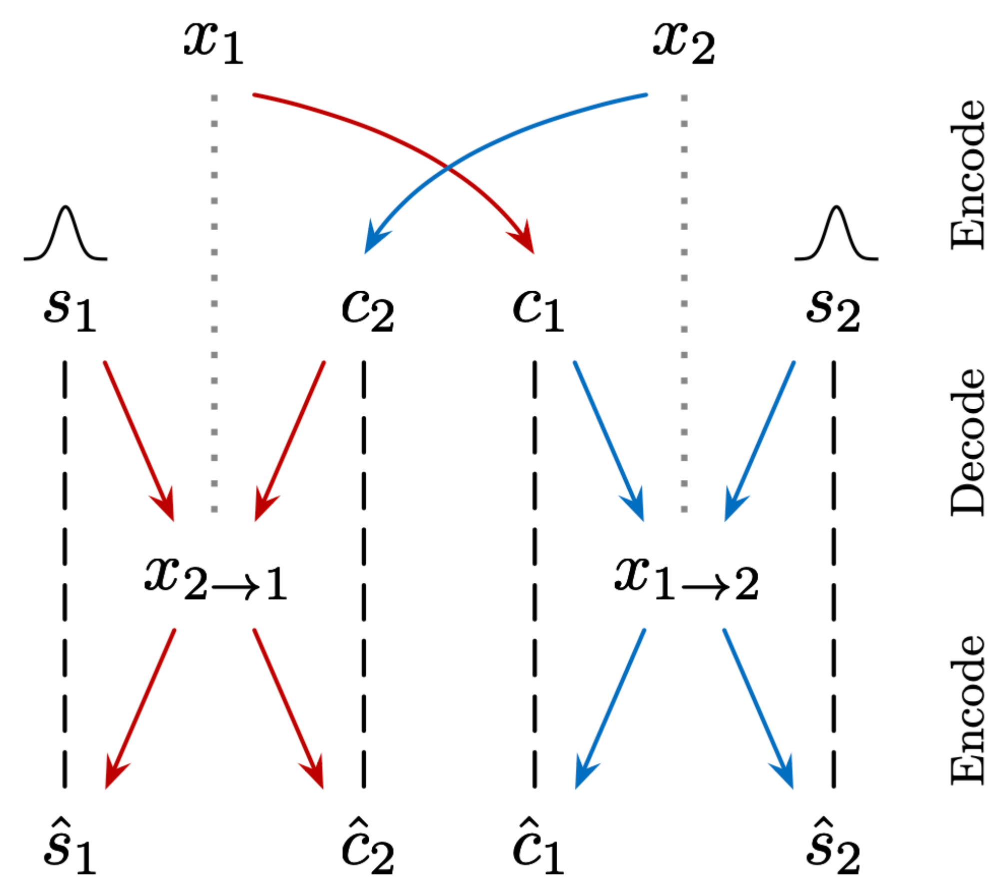
➤ Adversarial loss

- $L_{GAN}^{x_1} = \text{LSGAN}(D_1, G_1)$

PROPOSAL



(a) Within-domain reconstruction



(b) Cross-domain translation

- Cycle Image Reconstruction
  - $L^{x_1}_{cyc\_recon} = ||\hat{x}_1 - x_1||$ with $G_1(E^c_1(\bar{x}_2), E^s_1(x_1)), \bar{x}_2 = G_2(E^c_1(x_1), E^s_2(x_2))$

PROPOSAL

➤ Optical-flow-based training

Frame X^{n-1}

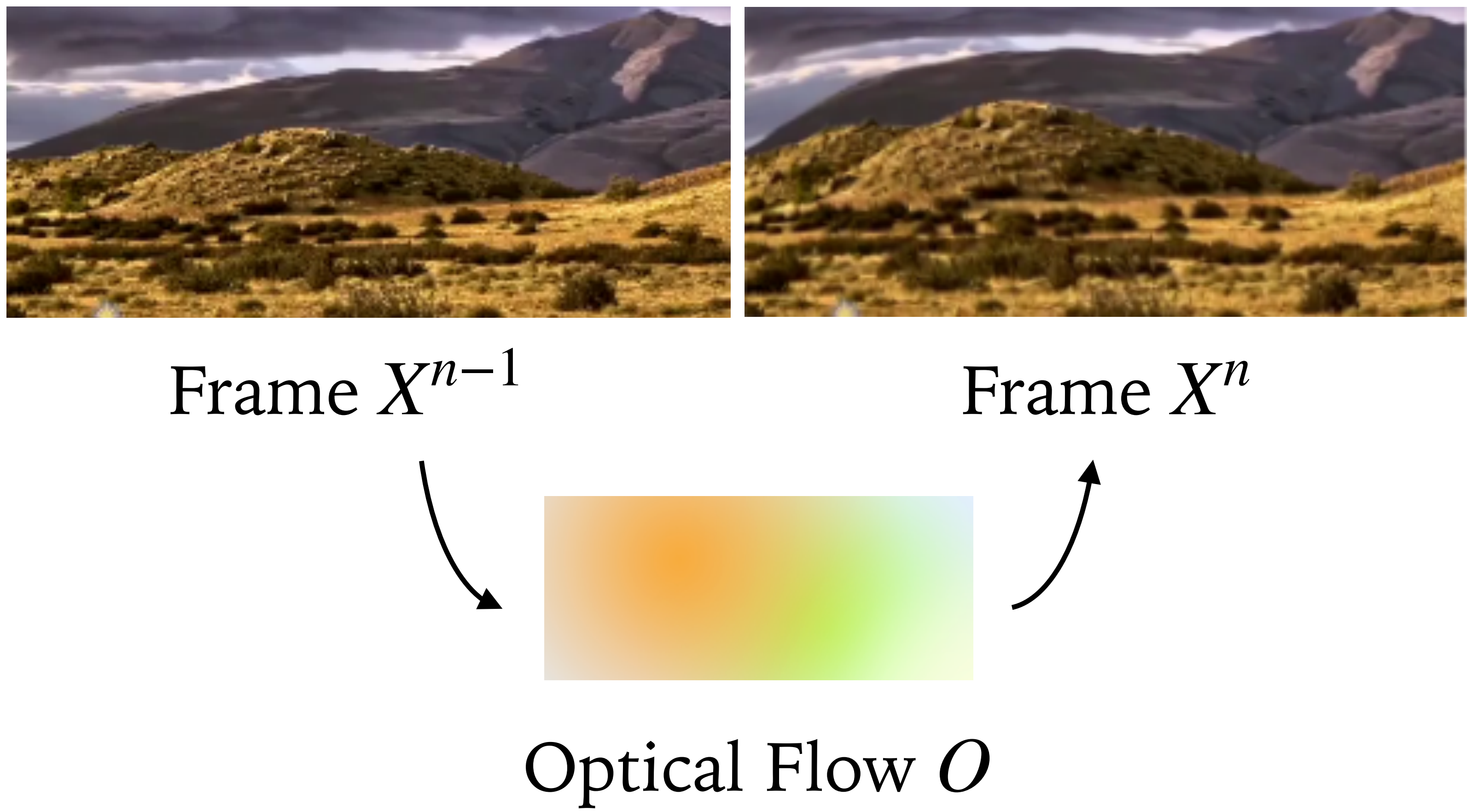
Frame X^{n}

Optical Flow O

- Both X^{n-1}, X^{n} and O are provided by the dataset

PROPOSAL

➤ Our proposal



- Only X^{n-1} is provided by the dataset
- O is randomly synthesized, and $X^n = Warp(X^{n-1}, O)$

PROPOSAL

- Temporal loss
  - $L_t^{x_1} = ||\tilde{x}_1^N - \mathrm{Warp}(\tilde{x}_1^{N-1}, O)|| + ||\hat{x}_1^N - \mathrm{Warp}(\hat{x}_1^{N-1}, O)||$
  - Works as a training loss function term
  - No video/optical flow for both training/testing

- Temporal loss plus noise
  - $x_1^N = Warp(x_1^{N-1}, O) + \delta$

PROPOSAL

➤ Loss functions

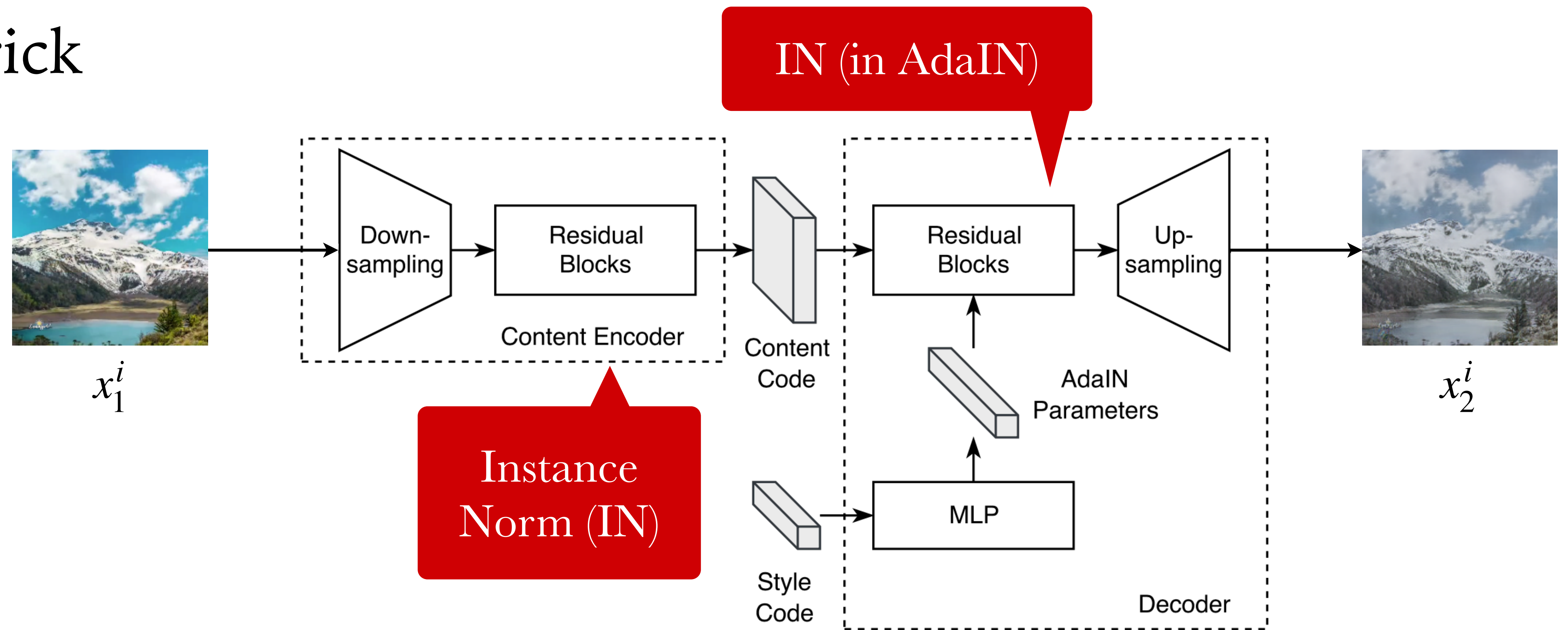
- $$\min_{E_1,E_2,G_1,G_2} \max_{D_1,D_2} L = L_{GAN}^{x_1} + L_{GAN}^{x_2} + L_{recon}^{x_1} + L_{recon}^{x_2}$$

$$L_{recon}^{c_1} + L_{recon}^{c_2} + L_{recon}^{s_1} + L_{recon}^{s_2}$$

$$L_{cyc\_recon}^{x_1} + L_{cyc\_recon}^{x_2} + L_t^{x_1} + L_t^{x_2}$$

PROPOSAL

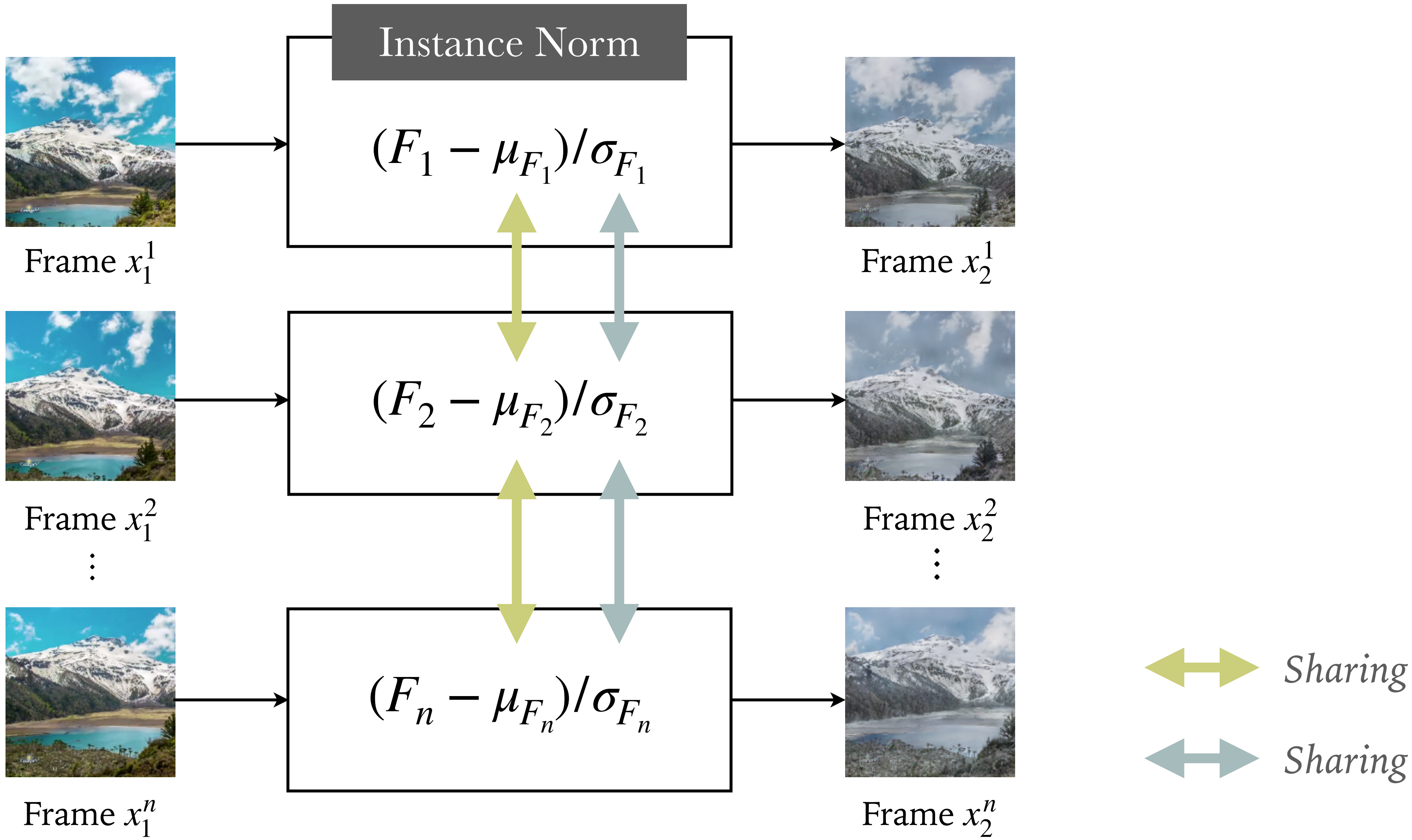
➤ Trick



- Mean and variance in IN of each frame are different
- We share them across frames → inter-frame relationship

PROPOSAL

➤ Trick

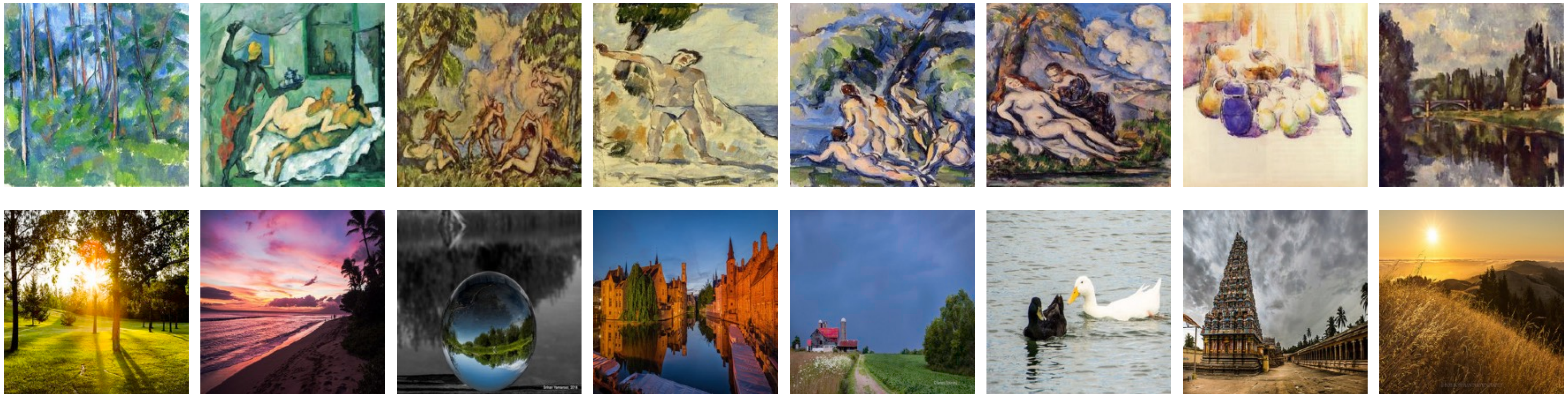


- $\mathrm{IN}(F) = \mathrm{Clamp}((F - \mu_{seq})/\sigma_{seq},\ \mathrm{MIN}_{seq},\ \mathrm{MAX}_{seq})$

EXPERIMENT

➤ Dataset

- We collect 7038 photographies and 3401 oil painting images from a series of CycleGAN datasets



- Resolution: 256×256
- Split: 6287/2559 for training, 842/751 for testing

EXPERIMENT

➤ Quantitative Comparison

- Metric: Temporal loss

$$L = ||M \odot Warp(X^{n-1}, O) - X^{n}||$$

  - Data: 16 scenery videos*, optical flow by PWC-Net[1]

- Metric: FID

$$L = ||\mu_{data} - \mu_{g}|| + tr(\Sigma_{data} + \Sigma_{g} - (\Sigma_{data}\Sigma_{g})^{\frac{1}{2}})$$

  - Data: testing set of our collected data

* Collected from bilibili.com

[1] PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume (CVPR 2018)

EXPERIMENT

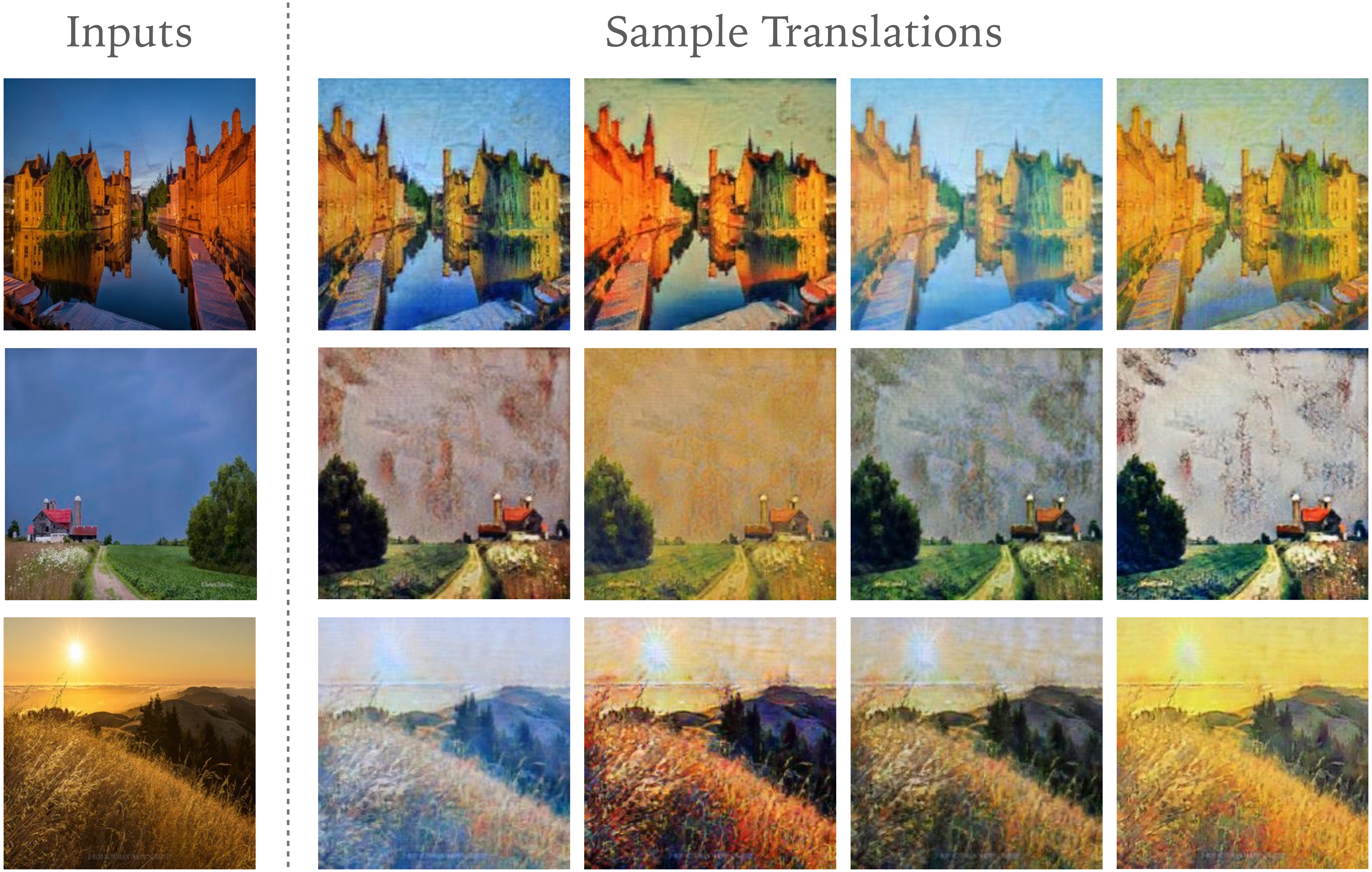
➤ Quantitative Comparison

	Temporal loss ↓	FID ↓
$\lambda_t = 0$	0.0479	**125.10**
$\lambda_t = 10$ (Ours)	0.0416	126.54
$\lambda_t = 15$	0.0312	153.37
$\lambda_t = 20$	**0.0307**	154.35

Input | $\lambda_t = 0$ | $\lambda_t = 10$ | $\lambda_t = 15$ | $\lambda_t = 20$

EXPERIMENT

➤ Qualitative Results



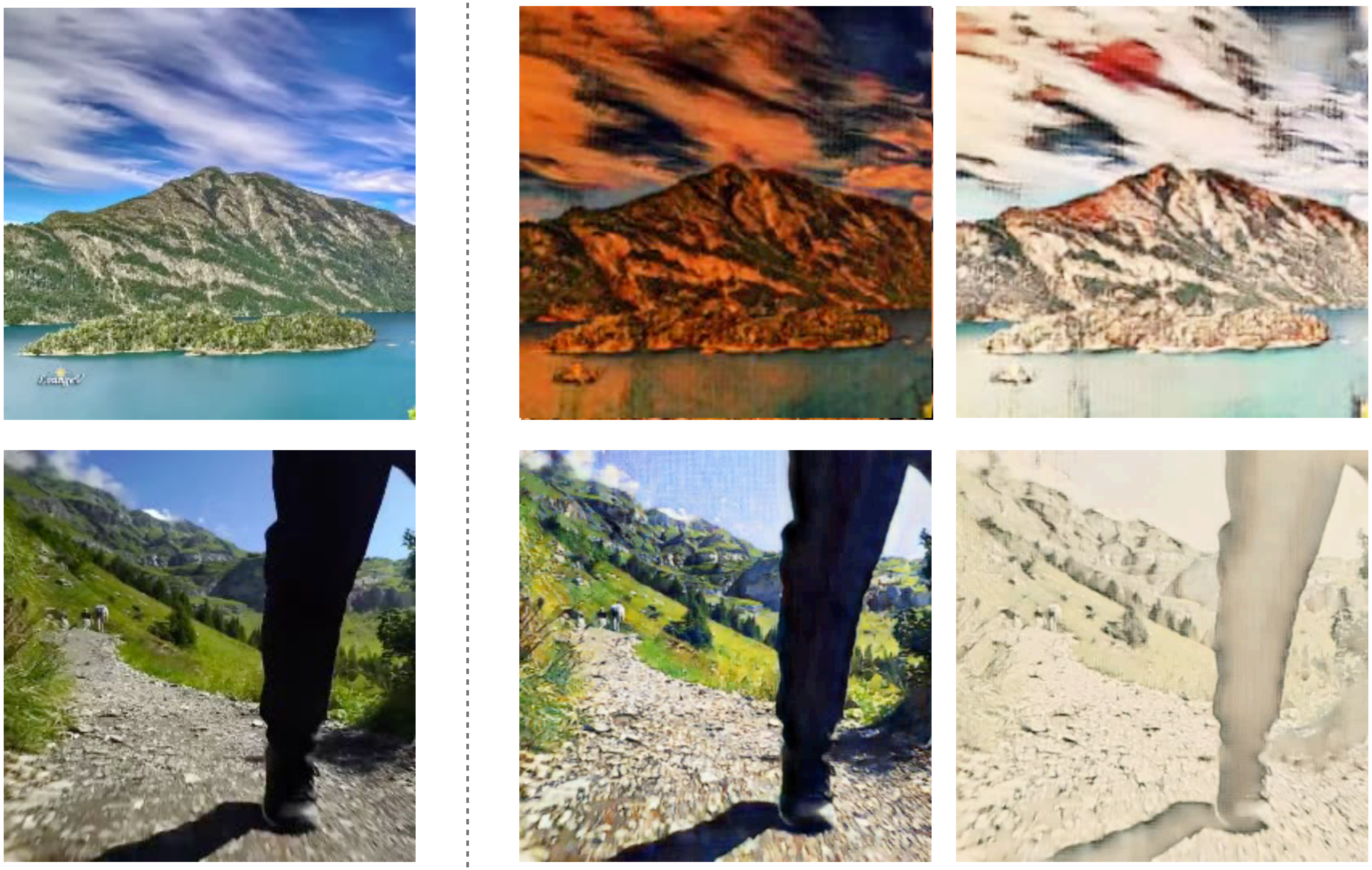
EXPERIMENT

Input

Sample Translations

256×256

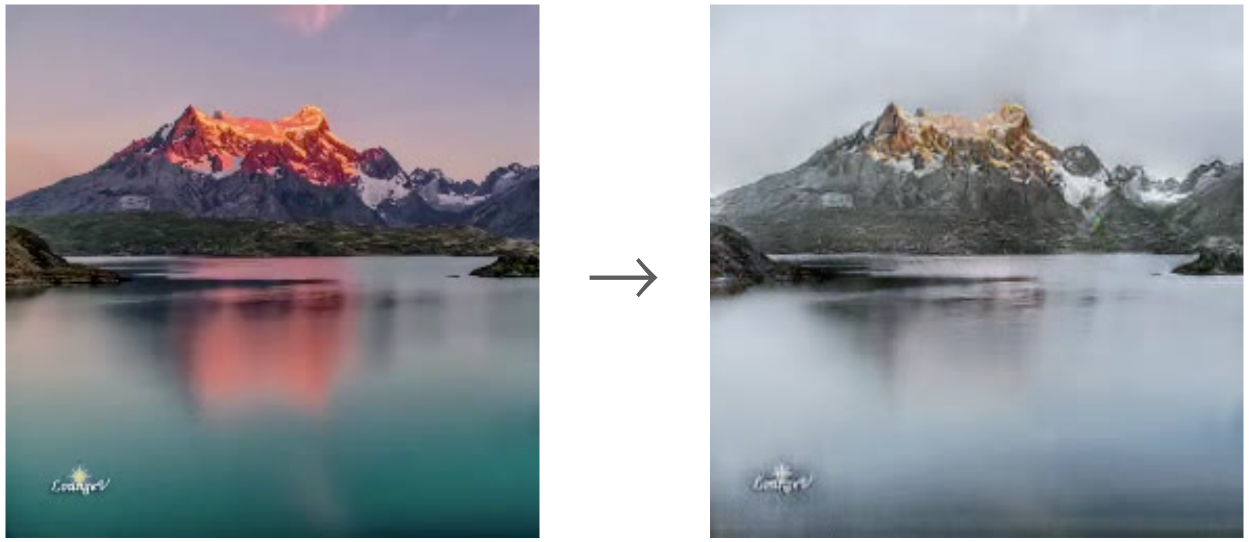
512×512



EXPERIMENT

- Other applications
  - Flower translation

  - Season translation



Input Result

CONCLUSION

- ➤ Single-Frame Unsupervised Video-to-video Translation
- ➤ Temporal loss
- $L_t^{x_1} = ||\tilde{x}_1^N - \mathrm{Warp}(\tilde{x}_1^{N-1}, O)|| + ||\hat{x}_1^N - \mathrm{Warp}(\hat{x}_1^{N-1}, O)||$
- Works as a training loss function term
- No video/optical flow for both training/testing