U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation

ICLR 2020

이 정 수

DAVIAN

Data and Visual Analytics Lab

Motivation

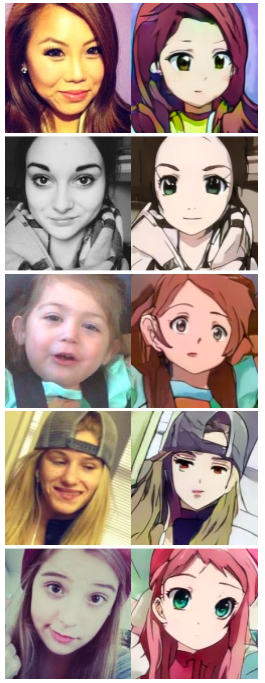
Previous models show performance differences depending on amount of change in texture/shape between domains (datasets)

- 1) Style transfer (photo2vangogh, photo2portrait) good <-> larger shape change (selfie2anime, cat2dog) bad
- 2) DRIT fails with 1) preserving shape (horse2zebra) **and** 2) changing shape (cat2dog) with fixed network architecture and hyper-parameters

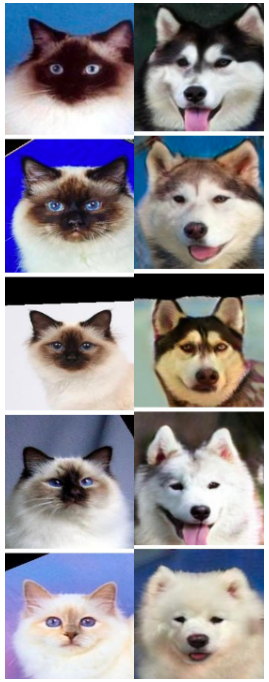
-> flexibly control the amount of change in shape and texture without modifying the model architecture or the hyper-parameters : 1) new **attention module** & 2) new **normalization method**

photo2vangogh

horse2zebra



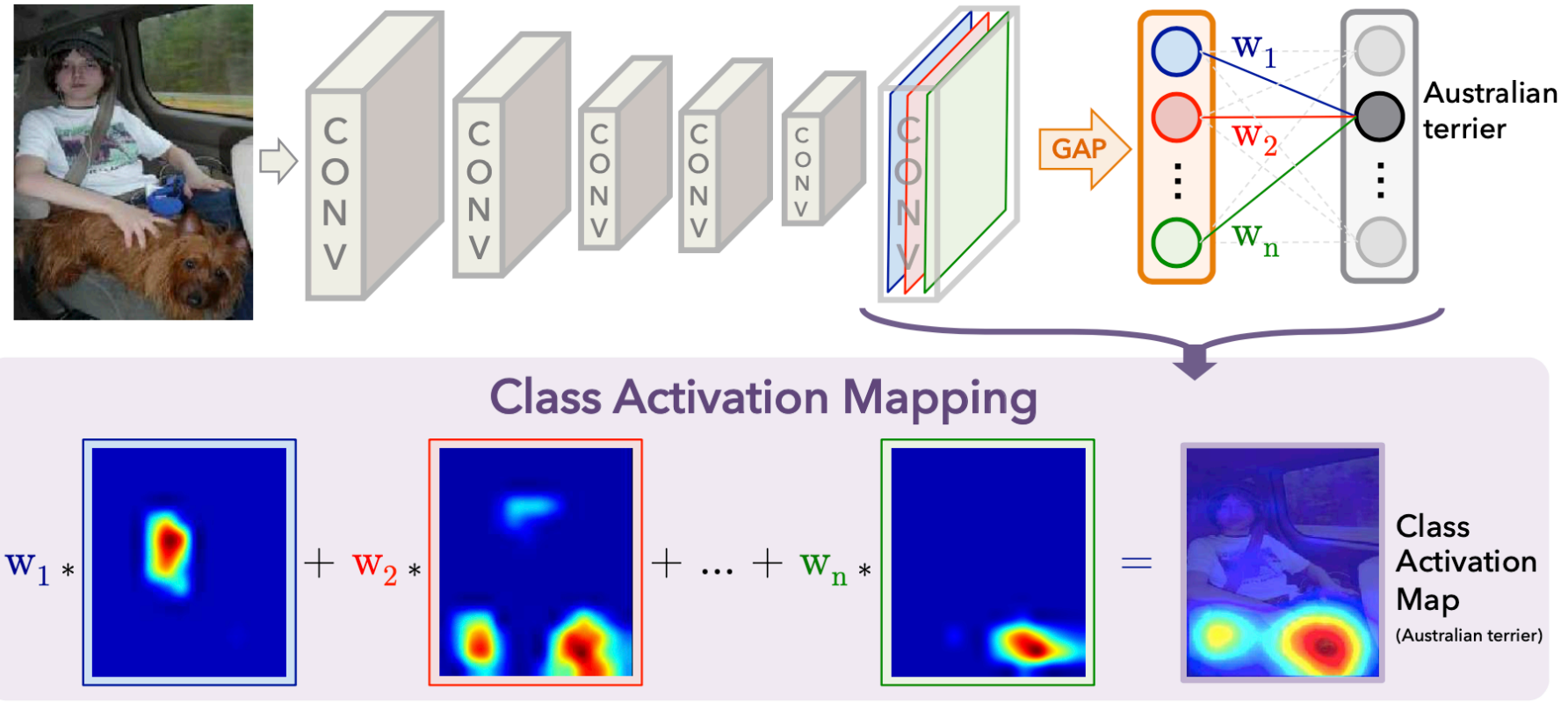
selfie2anime



cat2dog

Review

CAM Review



Review

Normalization Review

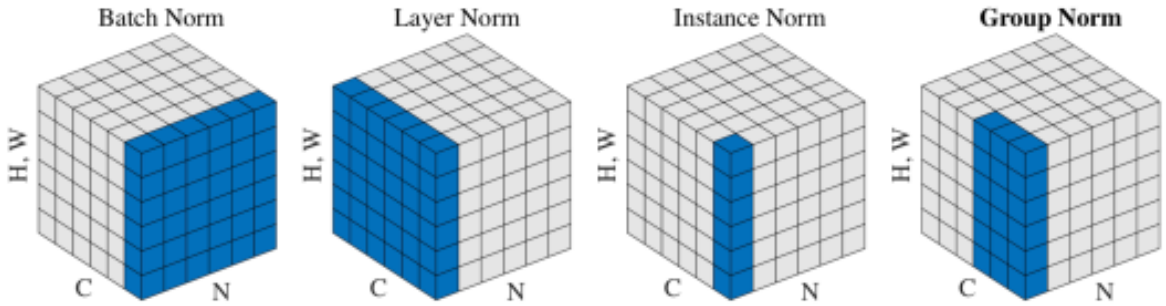


Figure 2. **Normalization methods**. Each subplot shows a feature map tensor, with N as the batch axis, C as the channel axis, and (H, W) as the spatial axes. The pixels in blue are normalized by the same mean and variance, computed by aggregating the values of these pixels.

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β
Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

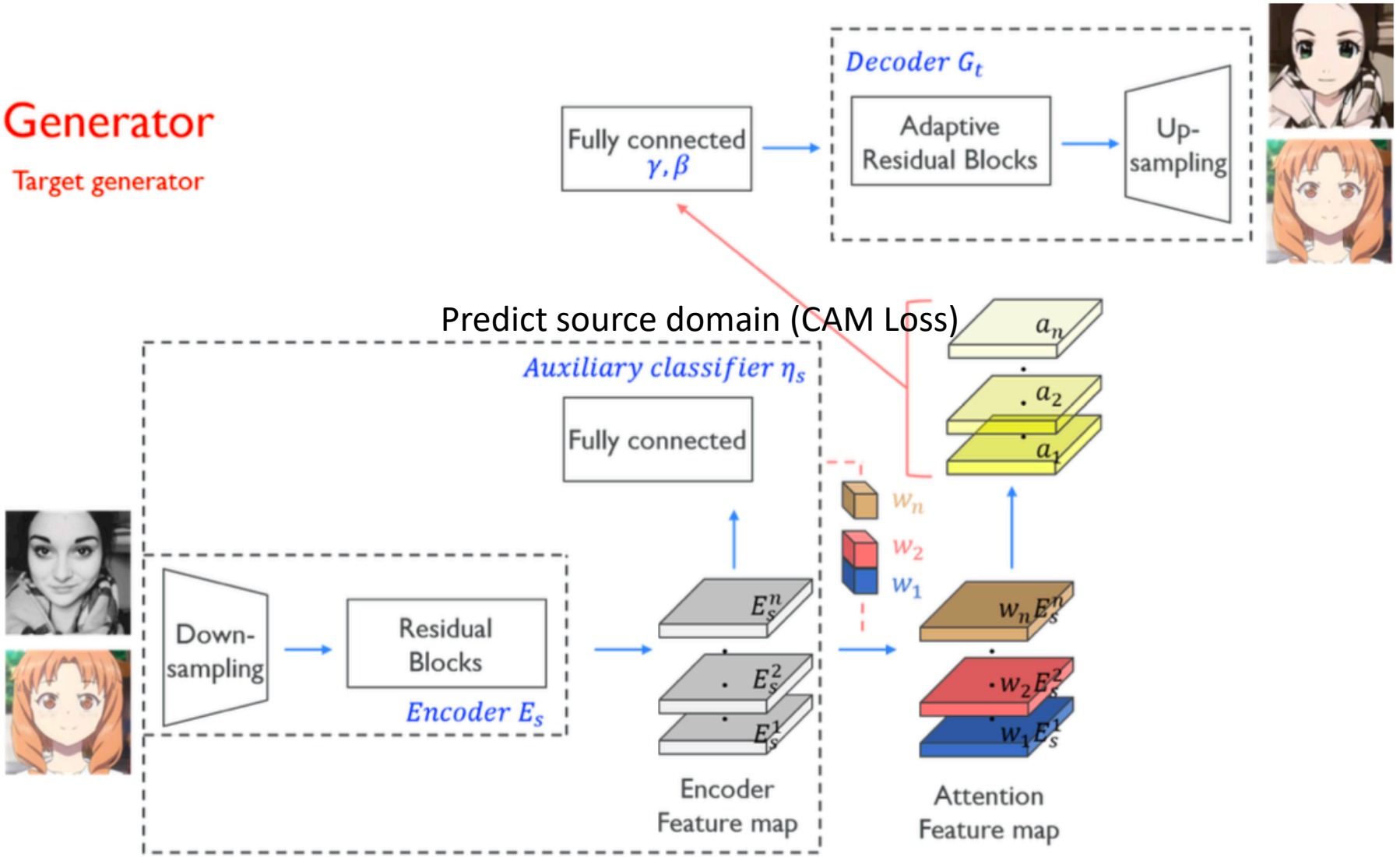
$$\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

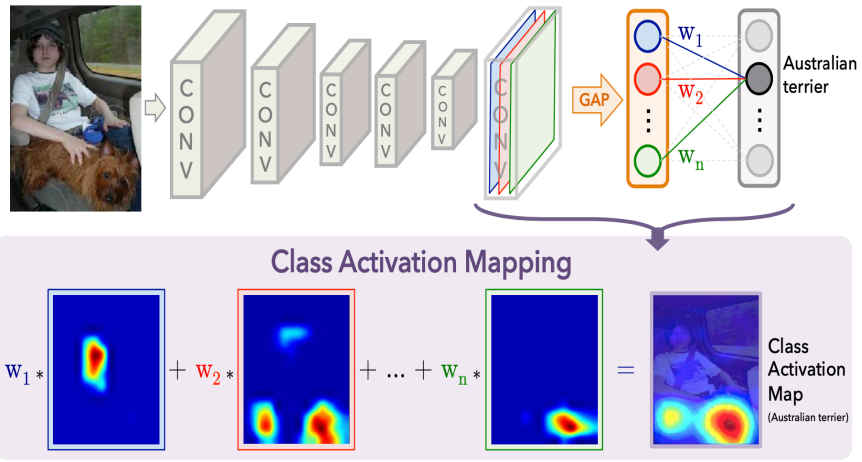
Batch Normalization example

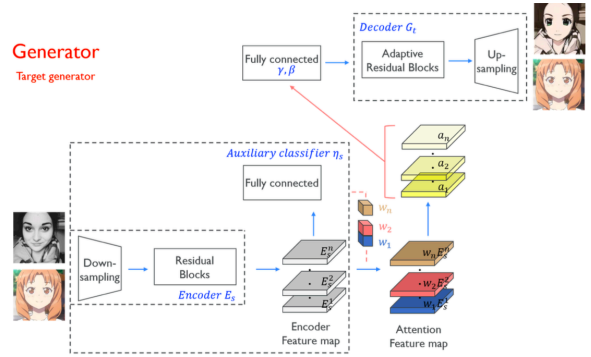
Method


Generator
Target generator
Decoder G_t
Fully connected γ, β
Adaptive Residual Blocks
Up-sampling
Predict source domain (CAM Loss)
Auxiliary classifier η_s
Fully connected
a_n
a_2
a_1
w_n
w_2
w_1
Down-sampling
Residual Blocks
Encoder E_s
E_s^n
E_s^2
E_s^1
$w_n E_s^n$
$w_2 E_s^2$
$w_1 E_s^1$
Encoder Feature map
Attention Feature map

Method

CAM

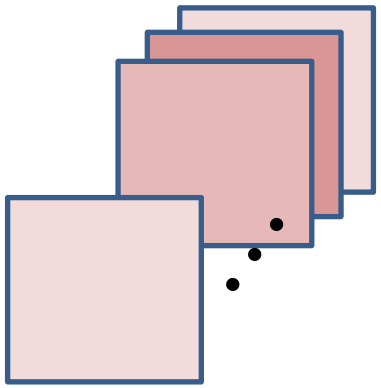




Predict source domain (CAM Loss)

256 -> 1

Encoder & Residual Block



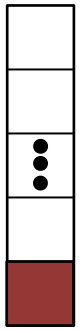
Last layer of residual block

E_s^k

256 * 64 * 64

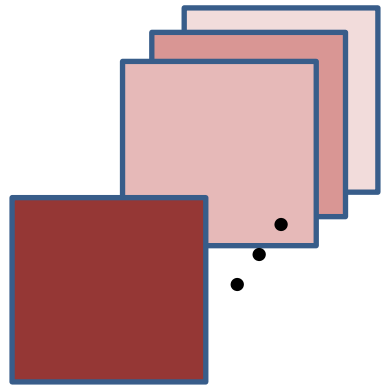
Global Average Pooling

256 * 1 * 1



Fully Connected

w_s^k



$w_s^k E_s^k$

256 * 64 * 64

decoder

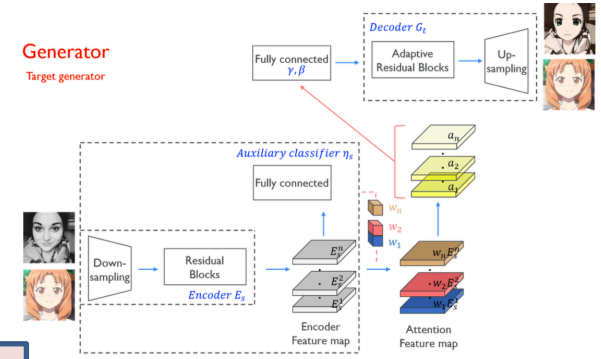
Method

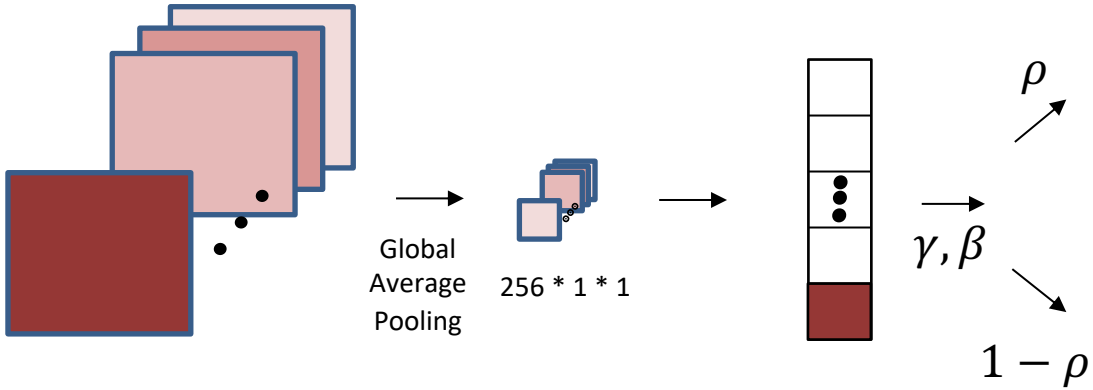
AdaLIN

$$AdaLIN(a, \gamma, \beta) = \gamma \cdot (\rho \cdot \hat{a_I} + (1 - \rho) \cdot \hat{a_L}) + \beta$$

$$\hat{a_I} = \frac{a - \mu_I}{\sqrt{\sigma_I^2 + \epsilon}}, \hat{a_L} = \frac{a - \mu_L}{\sqrt{\sigma_L^2 + \epsilon}},$$

$$\rho \leftarrow clip_{[0,1]}(\rho - \tau \Delta \rho)$$

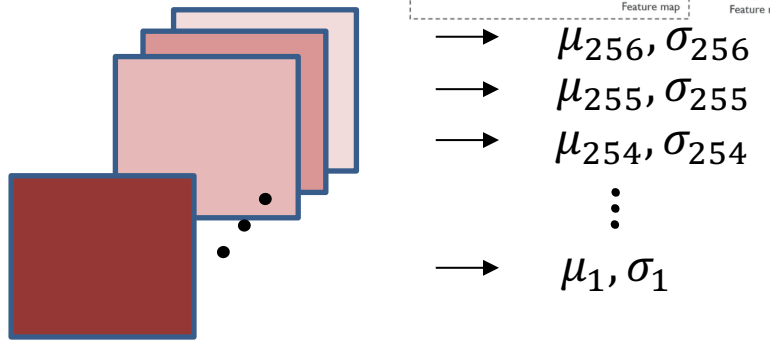




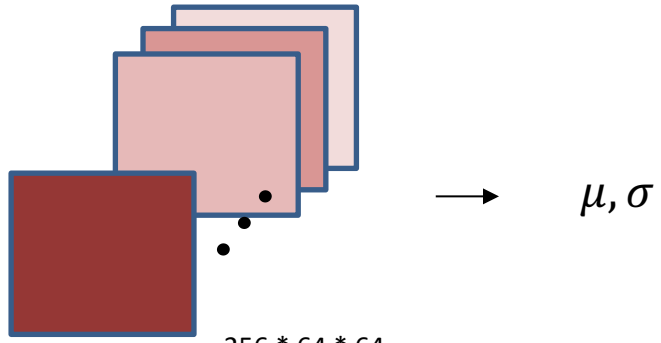
$w_s^k E_s^k (a_s^k)$

256 * 64 * 64

Fully Connected



Instance Normalization: features of source domain well preserved



256 * 64 * 64

Layer Normalization: target domain style transferred

*Related Work

Batch-Instance Normalization

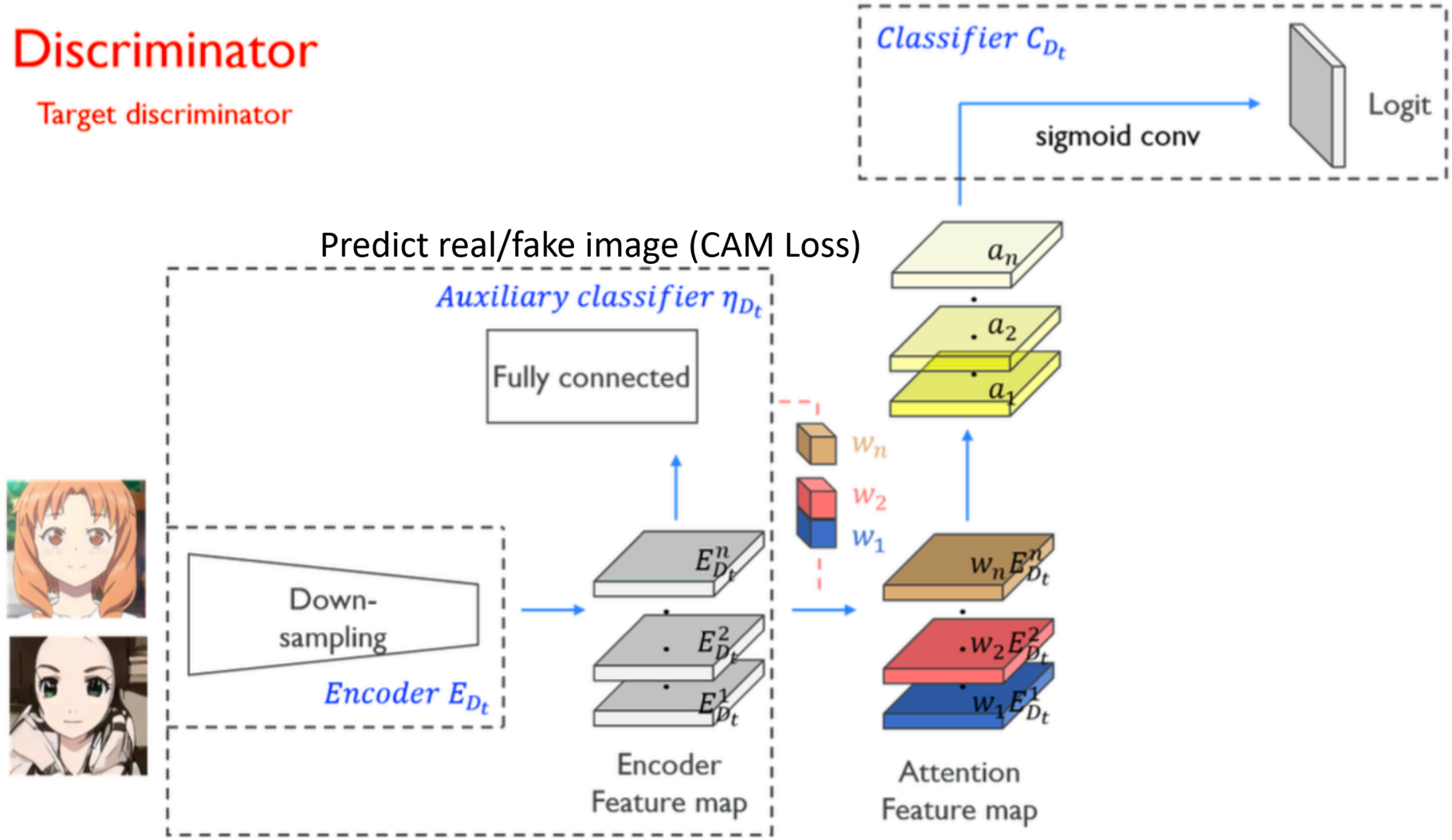
Irrelevant style / disturb image translation task: BIN suppress style using IN
Important style: BIN preserve style using BN

$$\mathbf{y} = \left(\rho \cdot \hat{\mathbf{x}}^{(B)} + (1-\rho) \cdot \hat{\mathbf{x}}^{(I)}\right) \cdot \gamma + \beta,$$

$$\rho \leftarrow \mathrm{clip}_{[0,1]}\left(\rho - \eta \Delta \rho\right),$$

Batch-Instance Normalization for Adaptively Style-Invariant Neural Networks, NIPS 2018

Method


Discriminator
Target discriminator
Classifier C_{D_t}
sigmoid conv
Logit
Predict real/fake image (CAM Loss)
Auxiliary classifier η_{D_t}
Fully connected
Down-sampling
Encoder E_{D_t}
$E^n_{D_t}$
$E^2_{D_t}$
$E^1_{D_t}$
Encoder Feature map
w_n
w_2
w_1
a_n
a_2
a_1
$w_n E^n_{D_t}$
$w_2 E^2_{D_t}$
$w_1 E^1_{D_t}$
Attention Feature map

Method

Loss

Adversarial Loss

$$L_{lsgan}^{s\to t} = (\mathbb{E}_{x\sim X_t}[(D_t(x))^2] + \mathbb{E}_{x\sim X_s}[(1 - D_t(G_{s\to t}(x)))^2]).$$

Cycle Loss

$$L_{cycle}^{s\to t} = \mathbb{E}_{x\sim X_s}[|x - G_{t\to s}(G_{s\to t}(x))|_1].$$

Identity Loss

$$L_{identity}^{s\to t} = \mathbb{E}_{x\sim X_t}[|x - G_{s\to t}(x)|_1].$$

CAM Loss

$$L_{cam}^{s\to t} = -(\mathbb{E}_{x\sim X_s}[log(\eta_s(x))] + \mathbb{E}_{x\sim X_t}[log(1 - \eta_s(x))])$$

$$L_{cam}^{D_t} = \mathbb{E}_{x\sim X_t}[(\eta_{D_t}(x))^2] + \mathbb{E}_{x\sim X_s}[(1 - \eta_{D_t}(G_{s\to t}(x))^2]$$

Total Loss

$$\min_{G_{s\to t}, G_{t\to s}, \eta_s, \eta_t} \max_{D_s, D_t, \eta_{D_s}, \eta_{D_t}} \lambda_1 L_{lsgan} + \lambda_2 L_{cycle} + \lambda_3 L_{identity} + \lambda_4 L_{cam}$$

where $\lambda_1 = 1, \lambda_2 = 10, \lambda_3 = 10, \lambda_4 = 1000$. Here, $L_{lsgan} = L_{lsgan}^{s\to t} + L_{lsgan}^{t\to s}$ and the other losses are defined in the similar way (L_{cycle}, $L_{identity}$, and L_{cam})

Experiment Results

CAM Analysis

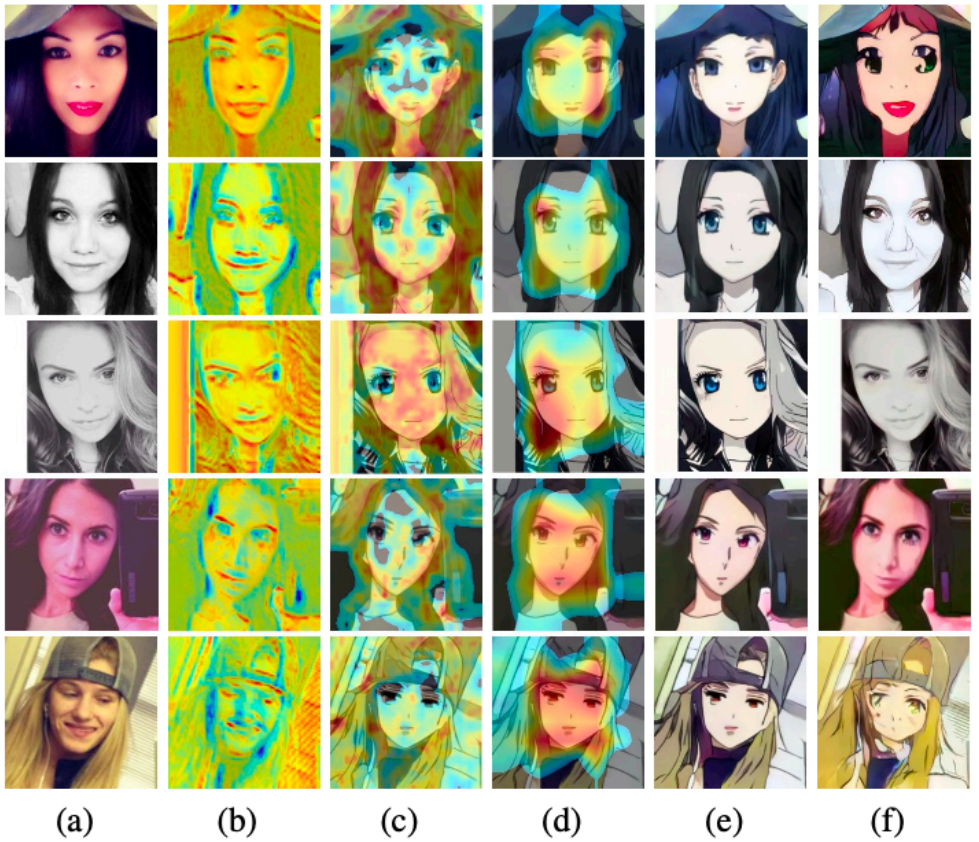


Figure 2: Visualization of the attention maps and their effects shown in the ablation experiments: (a) Source images, (b) Attention map of the generator, (c-d) Local and global attention maps of the discriminator, respectively. (e) Our results with CAM, (f) Results without CAM.

Experiment Results

ADALIN Analysis

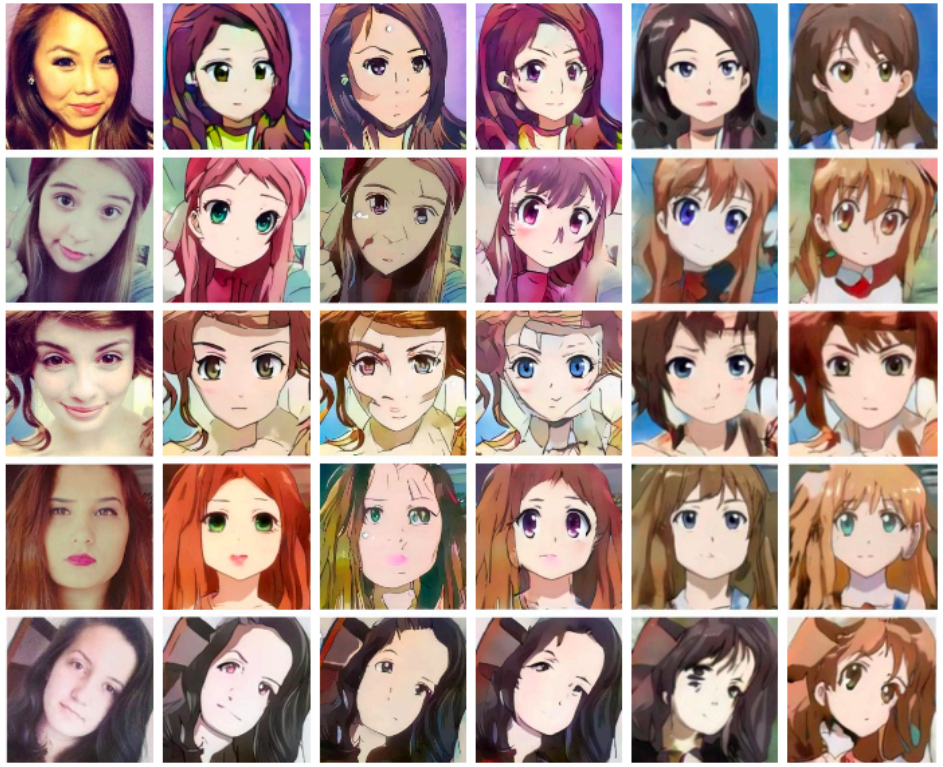


Figure 3: Comparison of the results using each normalization function: (a) Source images, (b) Our results, (c) Results only using IN in decoder with CAM, (d) Results only using LN in decoder with CAM, (e) Results only using AdaIN in decoder with CAM, (f) Results only using GN in decoder with CAM.

Experiment Results

Comparison with baseline models

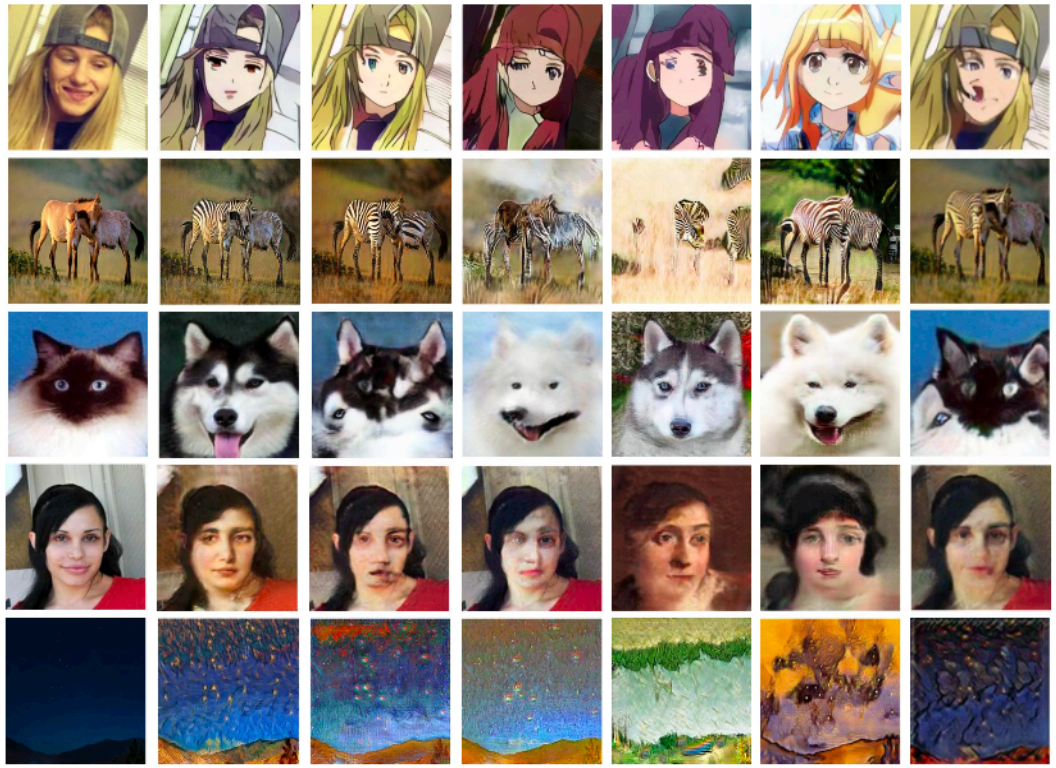


Figure 4: Visual comparisons on the five datasets. From top to bottom: selfie2anime, horse2zebra, cat2dog, photo2portrait, and photo2vangogh. (a)Source images, (b)U-GAT-IT, (c)CycleGAN, (d)UNIT, (e)MUNIT, (f)DRIT, (g)AGGAN

Table 3: Kernel Inception Distance×100±std.×100 for difference image translation mode. Lower is better.

Model	selfie2anime	horse2zebra	cat2dog	photo2portrait	photo2vangogh
U-GAT-IT	**11.61 ± 0.57**	**7.06 ± 0.8**	**7.07 ± 0.65**	1.79 ± 0.34	4.28 ± 0.33
CycleGAN	13.08 ± 0.49	8.05 ± 0.72	8.92 ± 0.69	1.84 ± 0.34	5.46 ± 0.33
UNIT	14.71 ± 0.59	10.44 ± 0.67	8.15 ± 0.48	**1.20 ± 0.31**	**4.26 ± 0.29**
MUNIT	13.85 ± 0.41	11.41 ± 0.83	10.13 ± 0.27	4.75 ± 0.52	13.08 ± 0.34
DRIT	15.08 ± 0.62	9.79 ± 0.62	10.92 ± 0.33	5.85 ± 0.54	12.65 ± 0.35
AGGAN	14.63 ± 0.55	7.58 ± 0.71	9.84 ± 0.79	2.33 ± 0.36	6.95 ± 0.33
CartoonGAN	15.85 ± 0.69	-	-	-	-
Model	anime2selfie	zebra2horse	dog2cat	portrait2photo	vangogh2photo
U-GAT-IT	**11.52 ± 0.57**	**7.47 ± 0.71**	**8.15 ± 0.66**	1.69 ± 0.53	5.61 ± 0.32
CycleGAN	11.84 ± 0.74	8.0 ± 0.66	9.94 ± 0.36	1.82 ± 0.36	**4.68 ± 0.36**
UNIT	26.32 ± 0.92	14.93 ± 0.75	9.81 ± 0.34	**1.42 ± 0.24**	9.72 ± 0.33
MUNIT	13.94 ± 0.72	16.47 ± 1.04	10.39 ± 0.25	3.30 ± 0.47	9.53 ± 0.35
DRIT	14.85 ± 0.60	10.98 ± 0.55	10.86 ± 0.24	4.76 ± 0.72	7.72 ± 0.34
AGGAN	12.72 ± 1.03	8.80 ± 0.66	9.45 ± 0.64	2.19 ± 0.40	5.85 ± 0.31

Experiment Results

Comparison with baseline models - anime2selfie (Supplementary Materials)

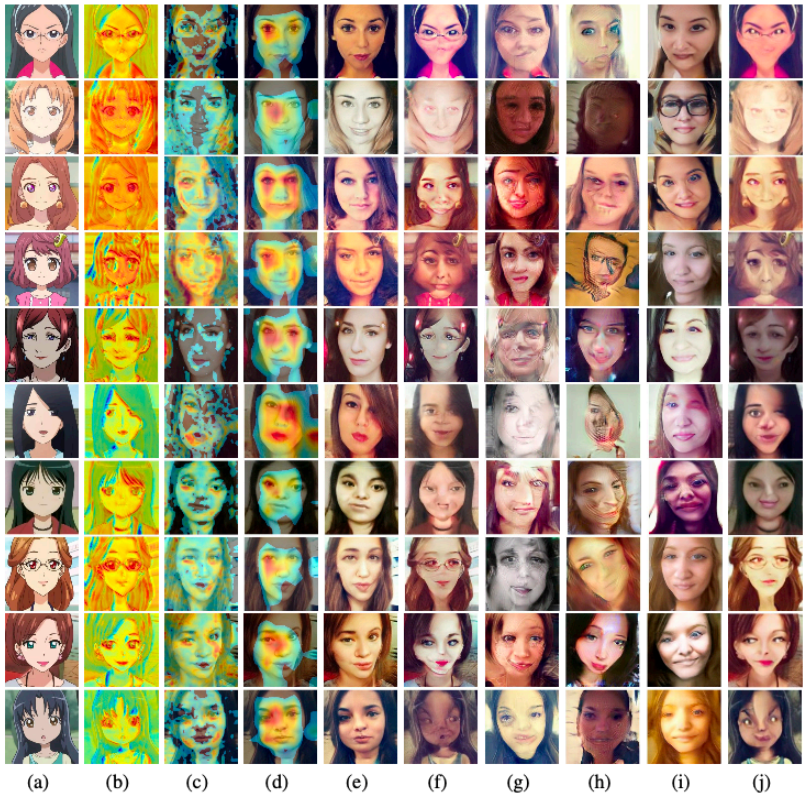


Figure 6: Visual comparisons of the anime2selfie with attention features maps. (a) Source images, (b) Attention map of the generator, (c-d) Local and global attention maps of the discriminators, (e) Our results, (f) CycleGAN (Zhu et al. (2017)), (g) UNIT (Liu et al. (2017)), (h) MUNIT (Huang et al. (2018)), (i) DRIT (Lee et al. (2018)), (j) AGGAN (Mejjati et al. (2018)).

Experiment Results

CAM & ADALIN ablation study

Table 1: Kernel Inception Distance×100±std.×100 for ablation our model. Lower is better. There are some notations; GN: Group Normalization, G_CAM: CAM of generator, D_CAM: CAM of discriminator

Model	selfie2anime	anime2selfie
U-GAT-IT	**11.61 ± 0.57**	**11.52 ± 0.57**
U-GAT-IT w/ IN	13.64 ± 0.76	13.58 ± 0.8
U-GAT-IT w/ LN	12.39 ± 0.61	13.17 ± 0.8
U-GAT-IT w/ AdaIN	12.29 ± 0.78	11.81 ± 0.77
U-GAT-IT w/ GN	12.76 ± 0.64	12.30 ± 0.77
U-GAT-IT w/o CAM	12.85 ± 0.82	14.06 ± 0.75
U-GAT-IT w/o G_CAM	12.33 ± 0.68	13.86 ± 0.75
U-GAT-IT w/o D_CAM	12.49 ± 0.74	13.33 ± 0.89

Experiment Results

Comparison with baseline models - User Study

Table 2: Preference score on translated images by user study.

Model	selfie2anime	horse2zebra	cat2dog	photo2portrait	photo2vangogh
U-GAT-IT	**73.15**	**73.56**	**58.22**	30.59	**48.96**
CycleGAN	20.07	23.07	6.19	26.59	27.33
UNIT	1.48	0.85	18.63	**32.11**	11.93
MUNIT	3.41	1.04	14.48	8.22	2.07
DRIT	1.89	1.48	2.48	2.48	9.70