Two-Level Adversarial Visual-Semantic Coupling for Generalized Zero-shot Learning

Reporter：陈思玉

2023.4.1

Chandhok S, Balasubramanian V N. Two-level adversarial visual-semantic coupling for generalized zero-shot learning[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021: 3100-3108.

Zero-Shot Learning

ZSL 目标

ZSL 旨在训练个模型，该模型能够通过语义信息的辅助，利用从 seen classes 中学到的知识来对 unseen classes 进行分类。

ZSL 所用数据

- seen classes：X^s（图像特征），Y^s（类别标签），A^s（语义信息）
- unseen classes：A^u（语义信息）

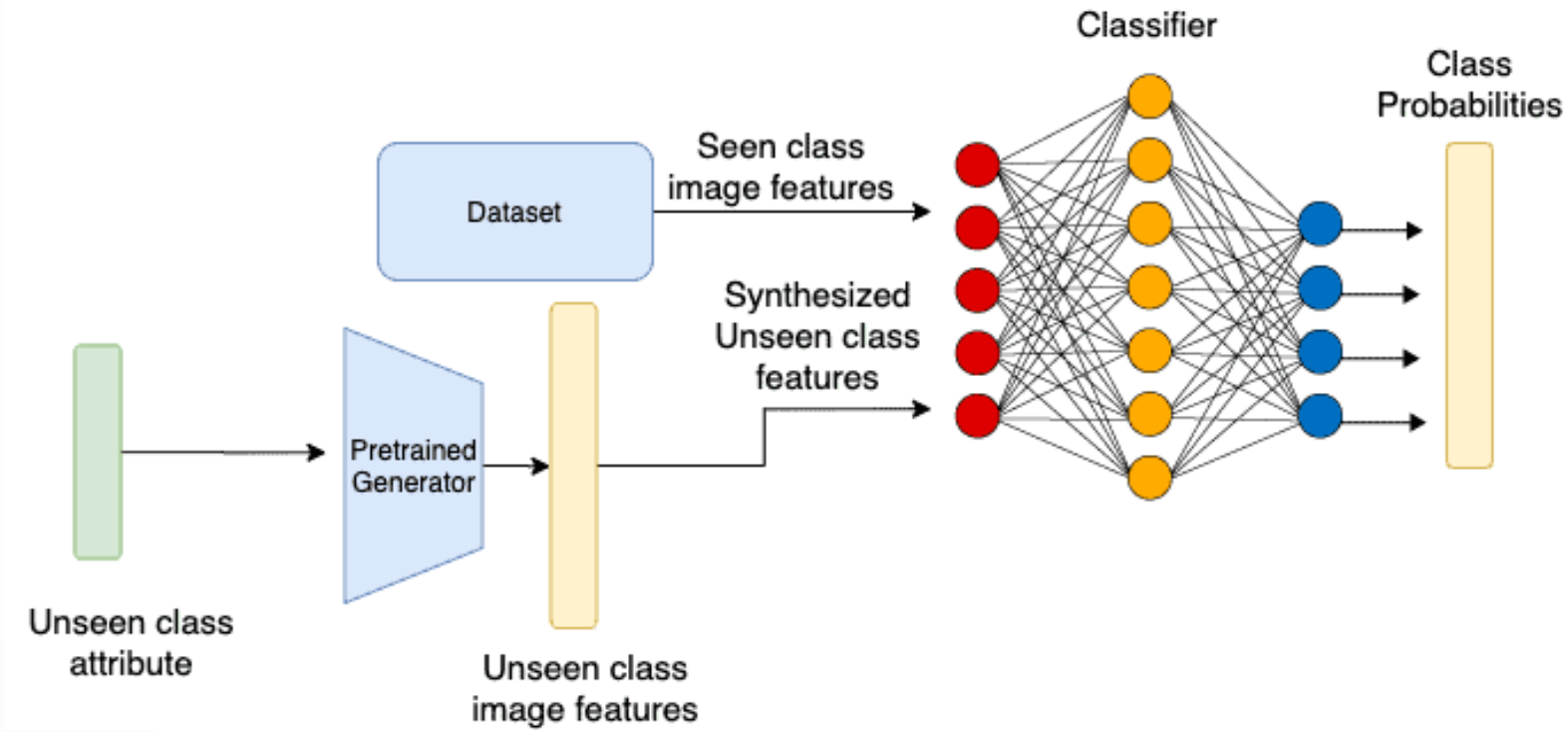
举例说明

1. 训练集有马、老虎、熊猫的图片
2. 语义信息有形状、条纹、颜色等属性
3. 给出斑马的定义：马的形状、老虎的条纹、熊猫的颜色
4. 输入斑马的图像，分类器能输出斑马的类别

Generative-based Methods

主要思想

1. 训练一个生成模型，该模型能够使用语义信息进行条件生成
2. 向训练好的模型输入 unseen classes 的语义信息，从而生成 unseen 的样本
3. 将训练集的 seen 样本和生成的 unseen 的样本组合成数据集
4. 将数据集输入分类器进行学习，从而使得分类器能对 seen 和 unseen classes 进行分类

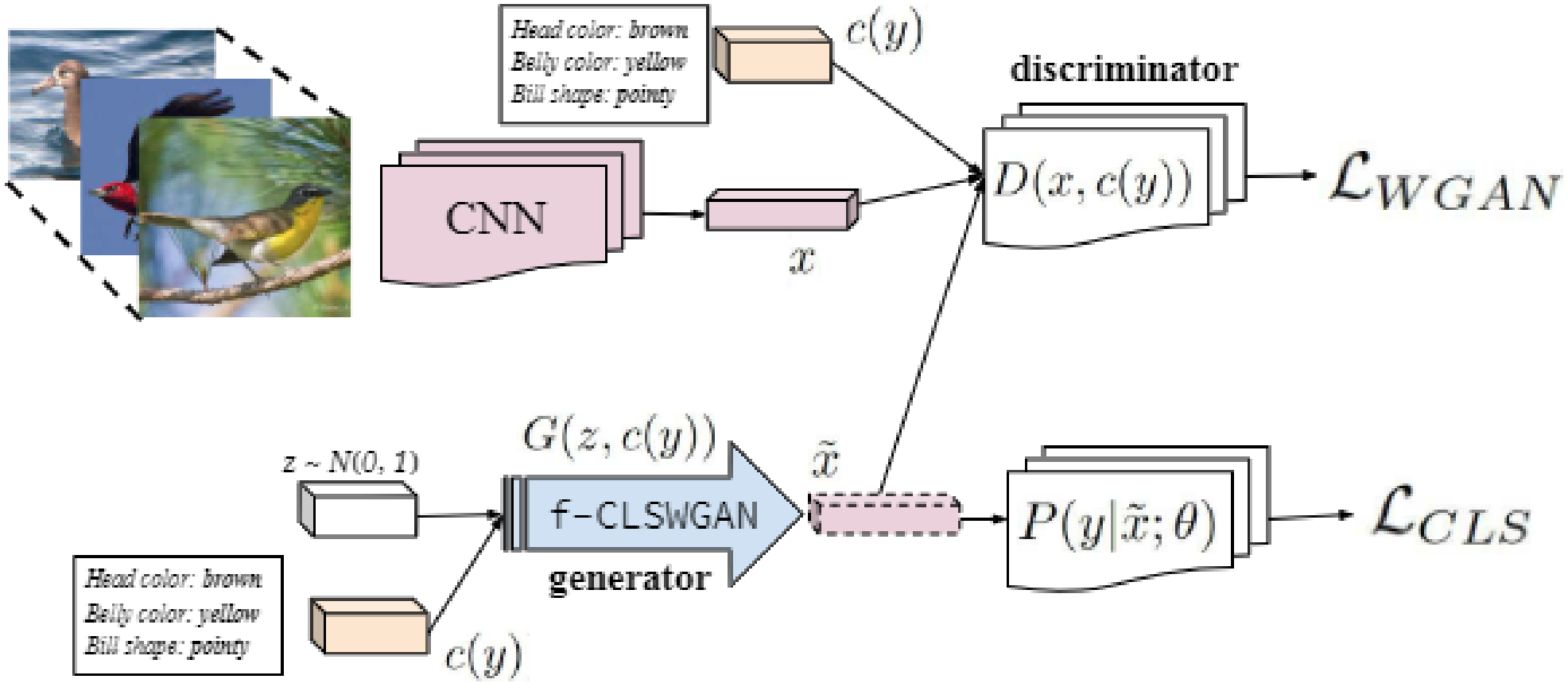


GAN 模型结构

“ Xian Y, Lorenz T, Schiele B, et al. Feature generating networks for zero-shot learning. ”

$$\mathcal{L}_{WGAN} = E[D(x, c(y))] - E[D(\tilde{x}, c(y))] - \lambda E[(\|\nabla_{\hat{x}} D(\hat{x}, c(y))\|_2 - 1)^2]$$

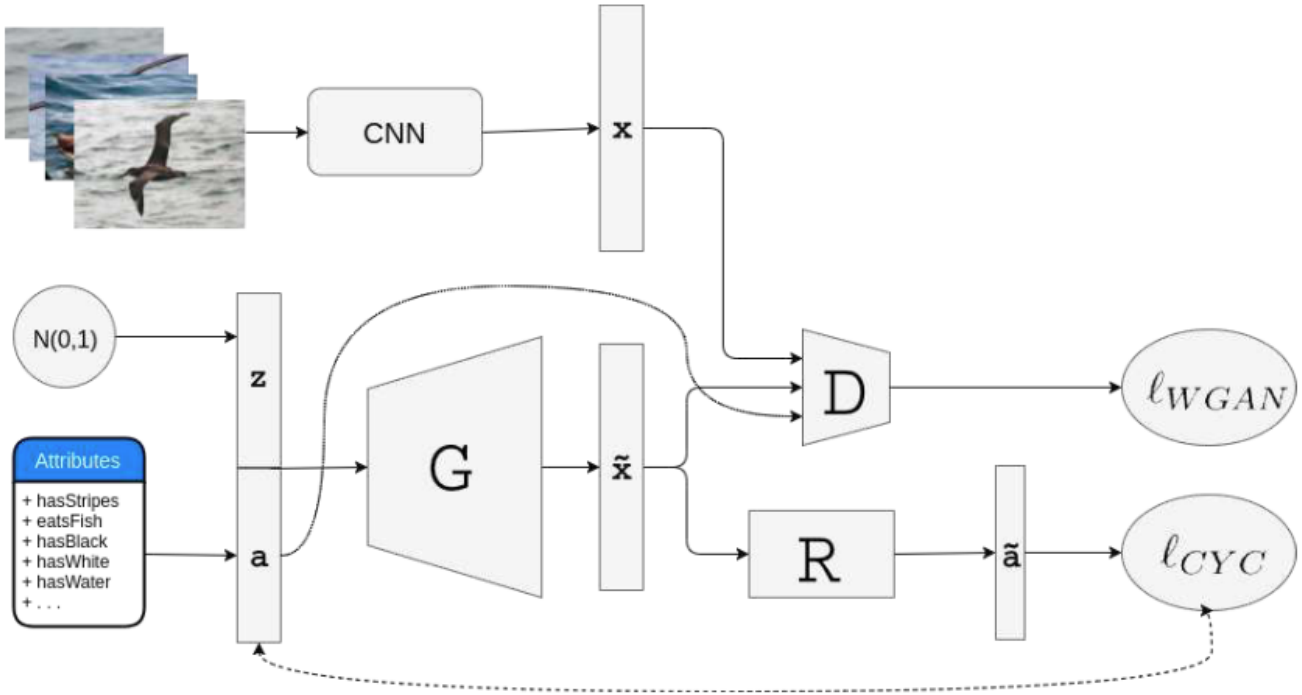
$$\mathcal{L}_{cls} = -E_{\tilde{x} \sim p_{\tilde{x}}}[\log P(y|\tilde{x}; \theta)]$$



Cycle-Consistent Loss

> Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks.
> Felix R, Reid I, Carneiro G. Multi-modal cycle-consistent generalized zero-shot learning.

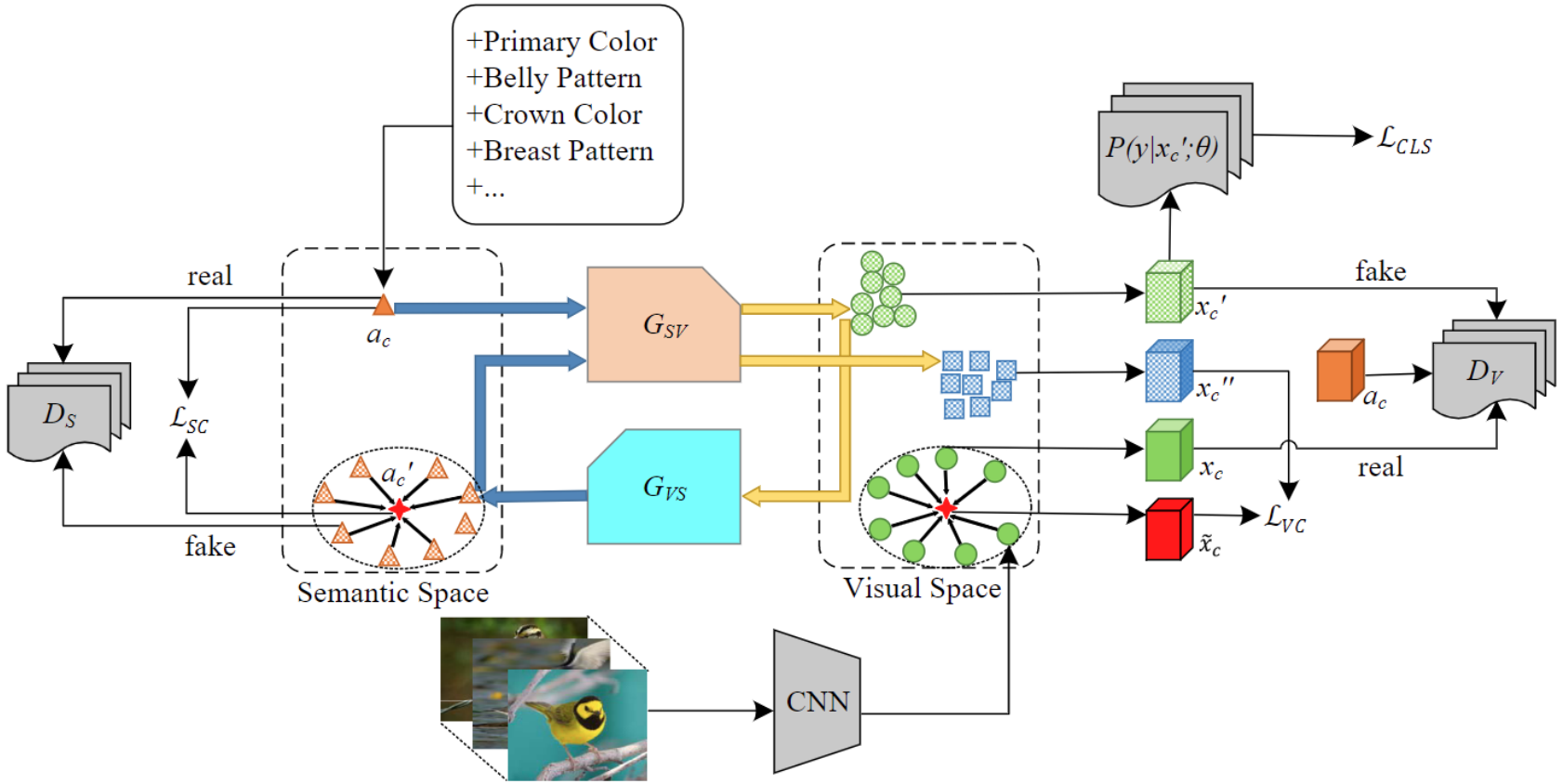
$$\ell_{CYC}(\theta_R, \theta_G) = \mathbb{E}_{\mathbf{a}\sim\mathbb{P}_S^a, \mathbf{z}\sim\mathcal{N}(0,\mathbf{I})}[\|\mathbf{a} - R(G(\mathbf{a}, \mathbf{z}; \theta_G); \theta_R)\|_2^2] + \mathbb{E}_{\mathbf{a}\sim\mathbb{P}_U^a, \mathbf{z}\sim\mathcal{N}(0,\mathbf{I})}[\|\mathbf{a} - R(G(\mathbf{a}, \mathbf{z}; \theta_G); \theta_R)\|_2^2]$$



双 WGAN 其一

“ Ni J, Zhang S, Xie H. Dual adversarial semantics-consistent network for generalized zero-shot learning. ”

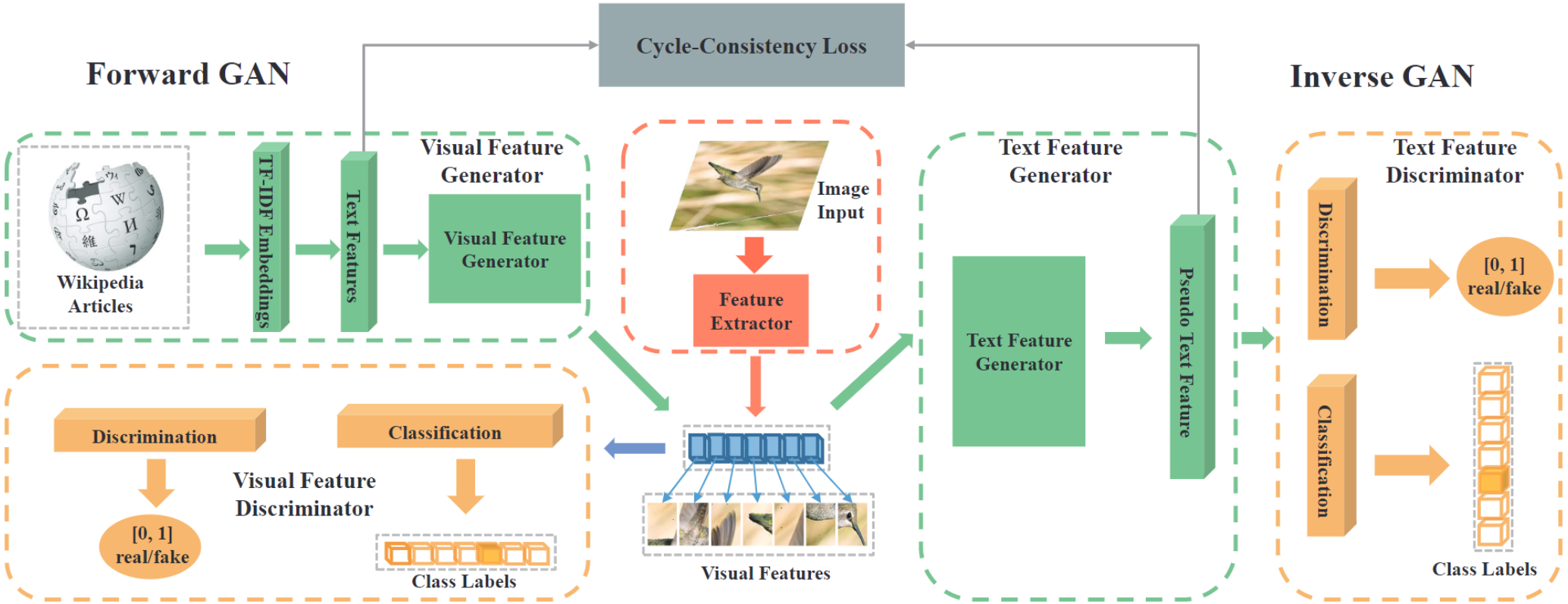
$$\mathcal{L}_{SC} = \frac{1}{C}\sum_{c=1}^{C}\|E_{a_c' \sim p_{a'}^c}[a_c'] - a_c\|_2 \qquad L_{VC} = \frac{1}{C}\sum_{c=1}^{C}\|E_{x_c'' \sim P_{x''}^c}[x_c''] - E_{x_c \sim P_x^c}[x_c]\|_2$$



双 WGAN 其二

> Chen Z, Li J, Luo Y, et al. Canzsl: Cycle-consistent adversarial networks for zero-shot learning from natural language.

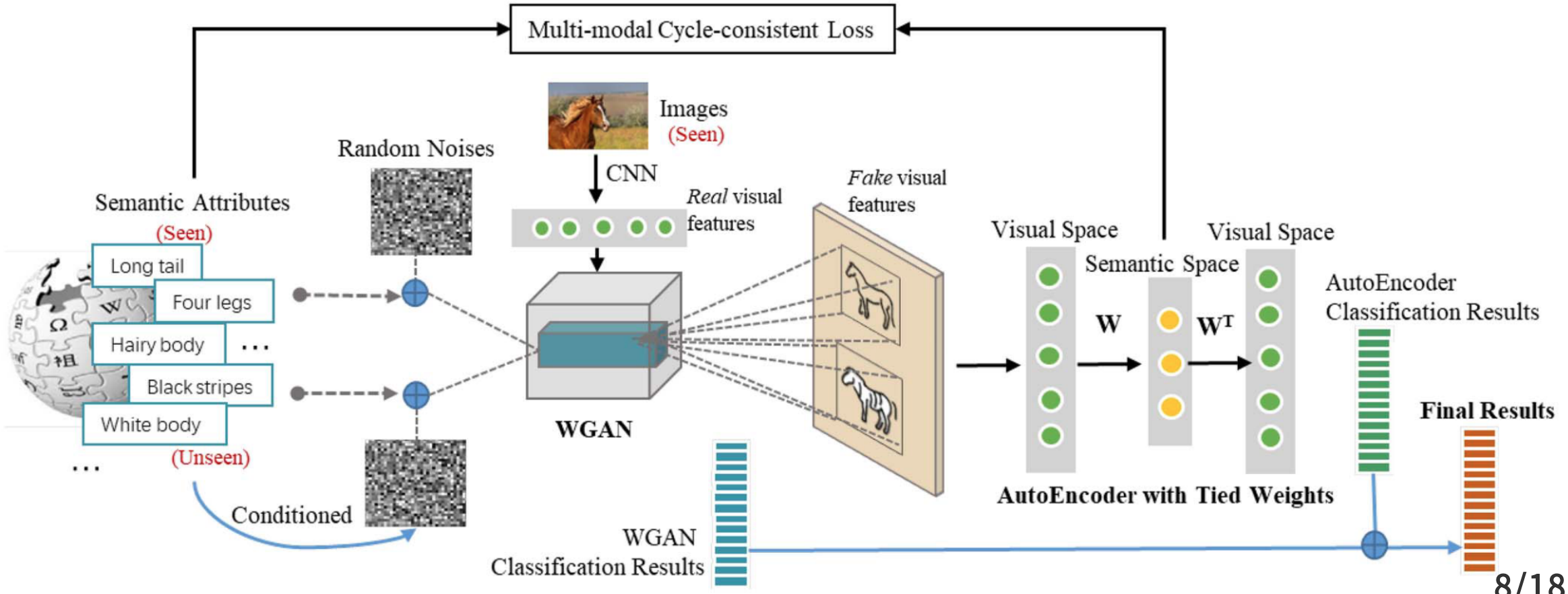
$$\mathcal{L}_{cyc} = \lambda \frac{1}{N^b} \sum_{n=1}^{N^b} \| G_2(G_1(\alpha, z, \theta), z, \delta) - s \|$$



Autoencoder

“ Li J, Jing M, Lu K, et al. Investigating the bilateral connections in generative zero-shot learning. ”

$$\min_{W} \|WX - A\|_F^2 + \eta \|X - W^T A\|_F^2 \qquad \ell_{cyc} = \mathbb{E}\left[\|a - W(G(z,a))\|_2^2\right]$$



双 WGAN 其三

“ Chandhok S, Balasubramanian V N. Two-level adversarial visual-semantic coupling for generalized zero-shot learning. ”

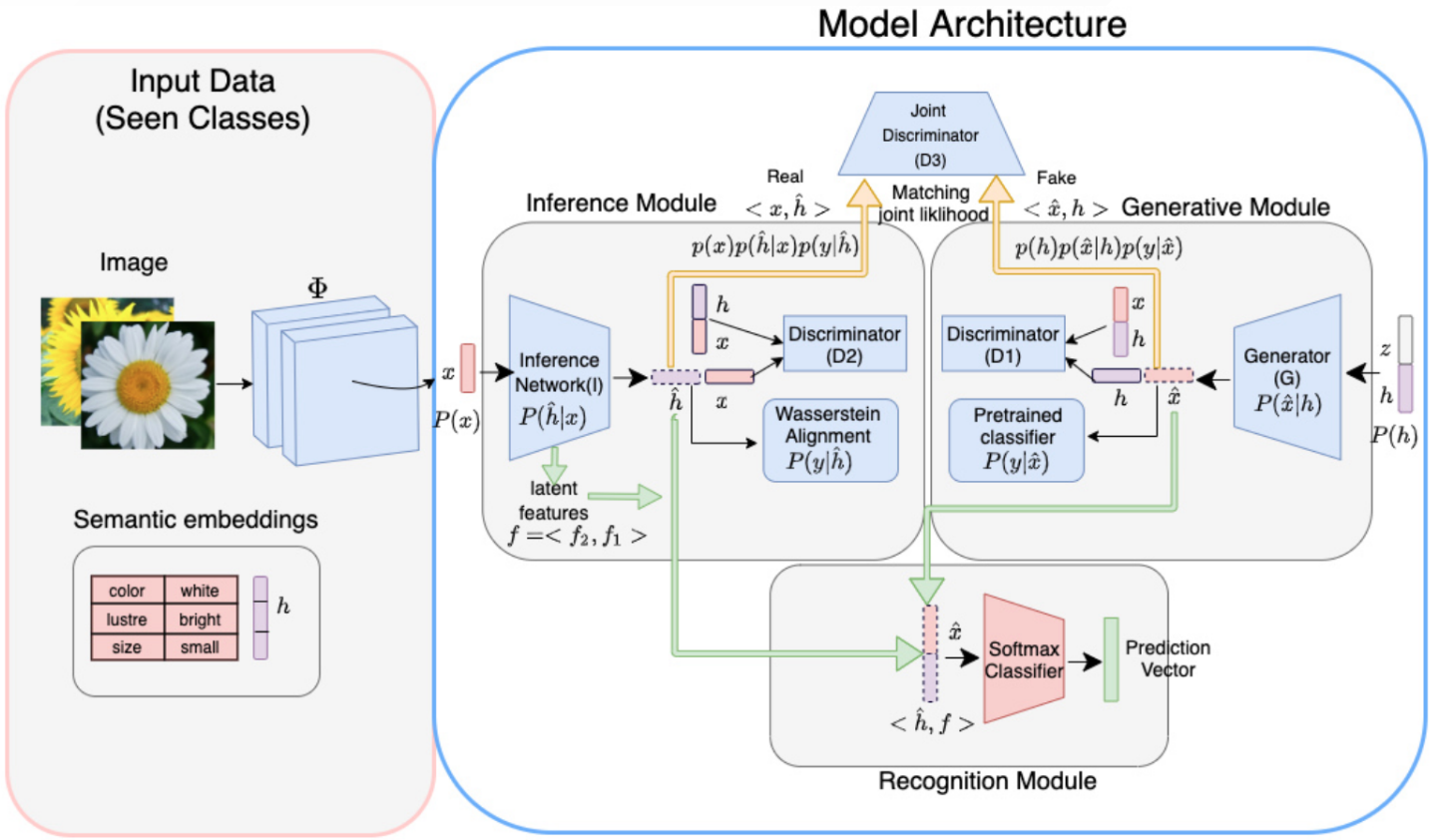


Figure 1: Network architecture for our proposed methodology. The proposed pipeline consists of a Generative module, Inference module, Recognition module and a Joint Discriminator. The model is trained on seen class visual features and semantic attributes. The feature extractor backbone network Φ is used to extract visual features from images. The vectors generated by our model are shown with a dotted outline. The final softmax classifier is trained on synthesized features $\hat{x}$ and representations from the inference network($< \hat{h}, f >$) as shown.

问题与贡献

> The performance of generative zero-shot methods mainly depends on the quality of generated features and how well the model facilitates knowledge transfer between visual and semantic domains.

- train GAN with an inference network to maximize the joint likelihood of visual and semantic features and capture the underlying modes of the data distribution better.
- use an adversarial joint-maximization loss to enhance the visual-semantic coupling and facilitate better cross-domain information transfer.
- use a novel Wasserstein semantic alignment loss that model the joint distribution of visual and semantic features better, and ensures that the generated semantic features are distributionally aligned with real semantic features.
- use the discriminative information in latent layers of the inference network to train our final recognition model, which helps provide the final recognition module with representations from both generative and inference modules, and thus enhances performance.

GZSL 建模

$$\mathbf{F1}: p(\mathbf{x}, \mathbf{h}, y) = p(\mathbf{x})p(\mathbf{h}|\mathbf{x})p(y|\mathbf{h})$$
$$\mathbf{F2}: p(\mathbf{x}, \mathbf{h}, y) = p(\mathbf{h})p(\mathbf{x}|\mathbf{h})p(y|\mathbf{x})$$

对于 $\mathbf{F1}$，
$p(\mathbf{x})$ 已知，$p(\mathbf{h}|\mathbf{x})$ 即给定输入 $\mathbf{x}$ 时 $\mathbf{h}$ 的条件概率，由推理网络建模

$$\begin{aligned}\mathbf{F1}: p(\mathbf{x}, \mathbf{h}, y) &\approx p(\mathbf{x})p(\hat{\mathbf{h}}|\mathbf{x})p(y|\hat{\mathbf{h}}) \\ &= p(\mathbf{x})p(\hat{\mathbf{h}}|\mathbf{x})p(y|\mathbf{h})p(\mathbf{h}|\hat{\mathbf{h}}) \\ &= p(\mathbf{x})p(\hat{\mathbf{h}}|\mathbf{x})p(\mathbf{h}|\hat{\mathbf{h}})\end{aligned}$$

对于 $\mathbf{F2}$，
$p(\mathbf{h})$ 已知，$p(\mathbf{x}|\mathbf{h})$ 即给定输入 $\mathbf{h}$ 时 $\mathbf{x}$ 的条件概率，由生成网络建模
$p(y|\mathbf{x})$ 即为 $\mathcal{L}_{CLS}$

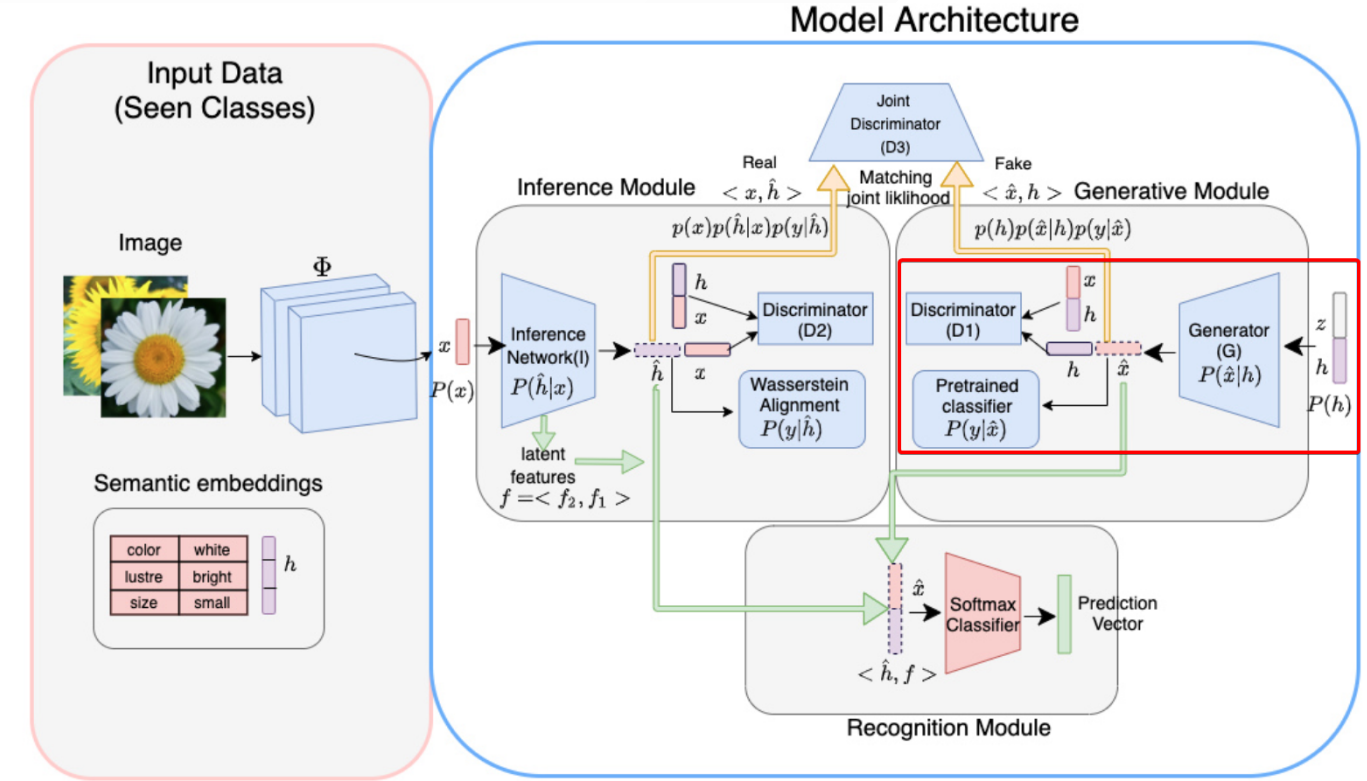
$$\mathbf{F2}: p(\mathbf{x}, \mathbf{h}, y) \approx p(\mathbf{h})p(\hat{\mathbf{x}}|\mathbf{h})p(y|\hat{\mathbf{x}})$$

生成网络 V S

$$\mathcal{L}_{WGAN_1} = \mathbb{E}[D_1(\mathbf{x}, \mathbf{h}(y))] - \mathbb{E}[D_1(\hat{\mathbf{x}}, \mathbf{h}(y))] - \lambda \mathbb{E}[(\|\nabla_{\tilde{\mathbf{x}}} D_1(\tilde{\mathbf{x}}, \mathbf{h}(y))\|_2 - 1)^2]$$

$$\mathcal{L}_{CLS} = -\mathbb{E}_{\hat{\mathbf{x}} \sim p_{\hat{\mathbf{x}}}}[\log P(y|\hat{\mathbf{x}}; \theta)]$$

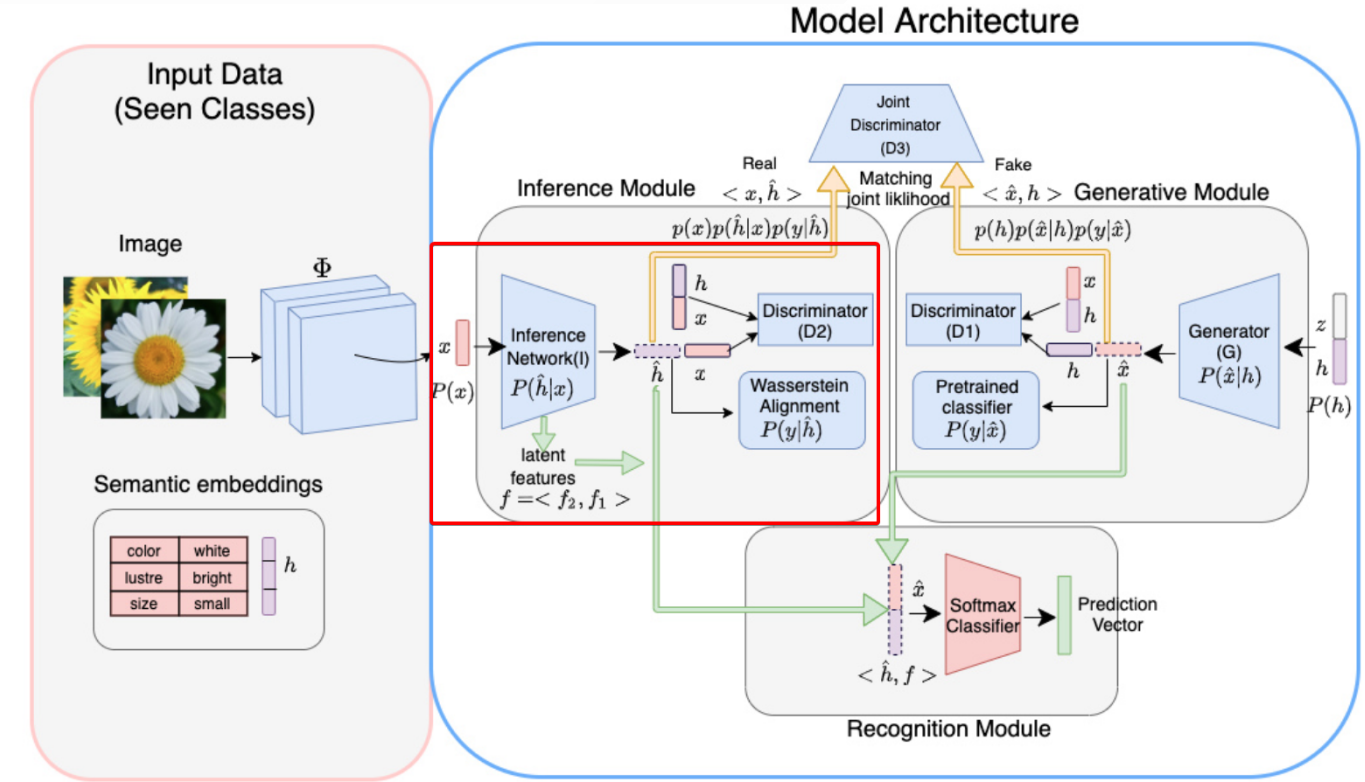
$$L_{gen} = \min_G \max_D \mathcal{L}_{WGAN_1} + \beta \mathcal{L}_{CLS}$$



推理网络 S V

$$\mathcal{L}_{WGAN_2} = \mathbb{E}[D_2(\mathbf{h}(y), x)] - \mathbb{E}[D_2(\hat{\mathbf{h}}, \mathbf{x})] - \lambda \mathbb{E}[\left(\|\nabla_{\mathbf{h}(\tilde{y})} D_2(\tilde{\mathbf{h}}, \mathbf{x})\|_2 - 1\right)^2]$$

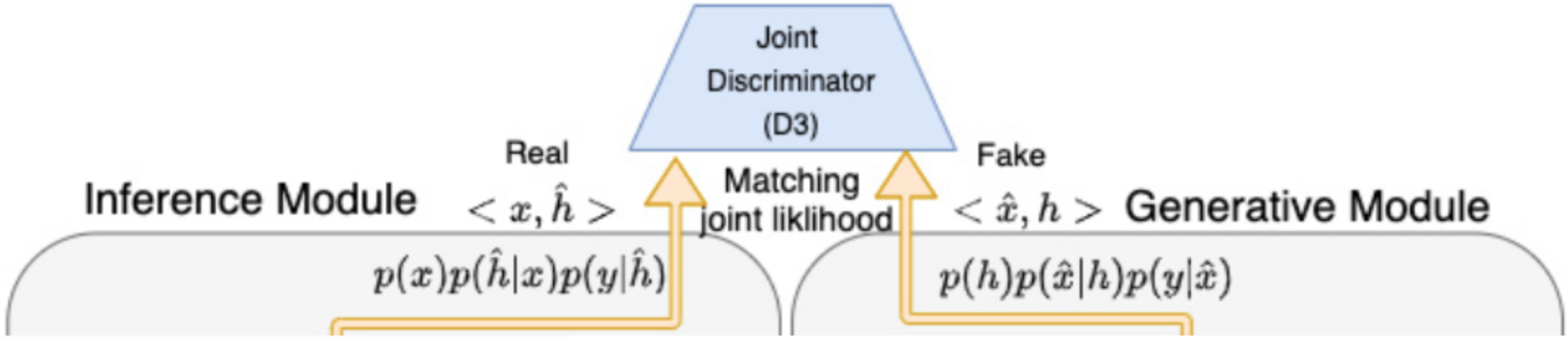
$$L_{inf} = \min_I \max_D \mathcal{L}_{WGAN_2} + \gamma \mathcal{L}_{wasserstein}$$



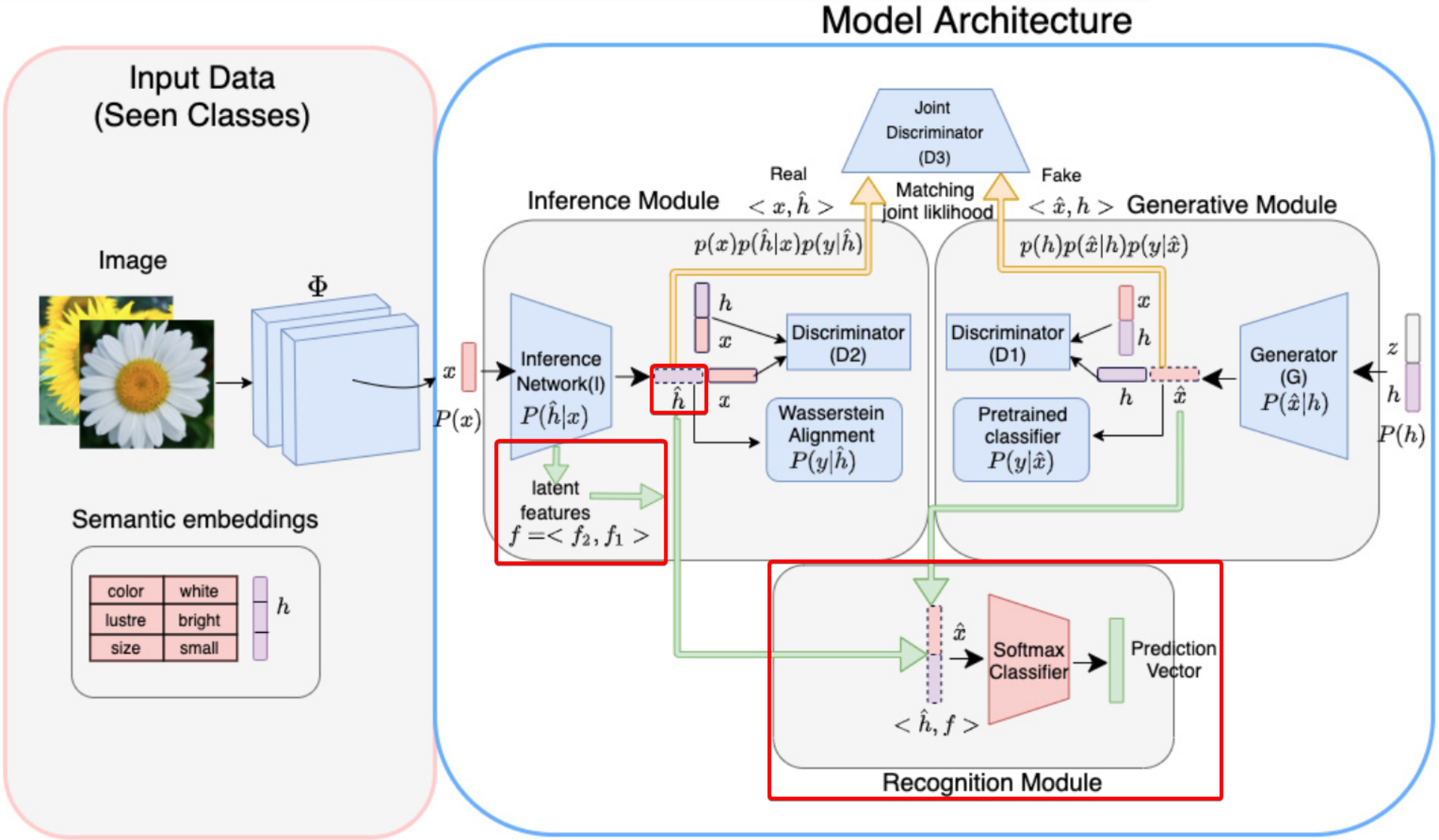
联合优化

$$\mathcal{L}_{joint-discriminator} = \mathbb{E}[D_3(\mathbf{x}, \hat{\mathbf{h}}(y))] - \mathbb{E}[D_3(\hat{\mathbf{x}}, \mathbf{h}(y))] - \lambda\mathbb{E}[\left(\|\nabla_{\tilde{x}} D_3(\tilde{\mathbf{x}}, \mathbf{h}(\tilde{y})), \nabla_{\tilde{\mathbf{h}}} D_3(\tilde{\mathbf{x}}, \mathbf{h}(\tilde{y}))\|_2 - 1\right)^2]$$

$$\mathcal{L}_{joint-max} = -(E[D_3(\mathbf{x}, \hat{\mathbf{h}}(y))] - \mathbb{E}[D_3(\hat{\mathbf{x}}, \mathbf{h}(y))])$$



分类


Model Architecture
Input Data
(Seen Classes)
Image
Φ
Semantic embeddings
color
white
lustre
bright
size
small
h
x
P(x)
Inference Module
Inference
Network(I)
P(ĥ|x)
latent
features
f =< f2, f1 >
Joint
Discriminator
(D3)
Real
< x, ĥ >
Matching
joint liklihood
Fake
< x̂, h > Generative Module
p(x)p(ĥ|x)p(y|ĥ)
p(h)p(x̂|h)p(y|x̂)
Discriminator
(D2)
Wasserstein
Alignment
P(y|ĥ)
Discriminator
(D1)
Pretrained
classifier
P(y|x̂)
Generator
(G)
P(x̂|h)
z
P(h)
Softmax
Classifier
Prediction
Vector
< ĥ, f >
Recognition Module

Experiments

Dataset	CUB			FLO			AWA1			AWA2		
Methods	U	S	H	U	S	H	U	S	H	U	S	H
DEM(CVPR'17)[32]	19. 6	57. 9	29. 2	-	-	-	32. 8	**84. 7**	47. 3	30. 5	**86. 4**	45. 1
ZSKL(CVPR'18)[10]	21. 6	52. 8	30. 6	-	-	-	18. 3	79. 3	29. 8	18. 9	82. 7	30. 8
DCN(NIPS'18)[14]	28. 4	60. 7	38. 7	-	-	-	-	-	-	25. 5	**84. 2**	39. 1
ALE(TPAMI'13)[1]	23. 7	62. 8	34. 4	13. 3	61. 6	21. 9	16. 8	76. 1	27. 5	**81. 8**	14. 0	23. 9
DEVISE(NIPS'13)[9]	23. 8	53. 0	32. 8	9. 9	44. 2	16. 2	13. 4	68. 7	22. 4	74. 7	17. 1	27. 8
ESZSL(ICML'15)[20]	12. 6	63. 8	21. 0	11. 4	56. 8	19. 0	6. 6	75. 6	12. 1	77. 8	5. 9	11. 0
SYNC(CVPR'16)[5]	11. 5	**70. 9**	19. 8	-	-	-	8. 9	**87. 3**	16. 2	**90. 5**	10. 0	18. 0
LATEM(CVPR'16)[27]	15. 2	57. 3	24. 0	-	-	-	7. 3	71. 7	13. 3	77. 3	11. 5	20. 0
SJE(CVPR'15)[2]	23. 5	59. 2	33. 6	13. 9	47. 6	21. 5	**74. 6**	11. 3	19. 6	73. 9	8. 0	14. 4
CLSWGAN(CVPR'18)[29]	43. 7*	57. 7*	49. 7*	59. 0	73. 8	**65. 6**	-	-	-	57. 9	61. 4	59. 6
CADA-VAE(CVPR'19)[21]	53. 5	51. 6	52. 4*	-	-	-	**72. 8**	57. 3	**64. 1**	75. 0	55. 8	63. 9
VSE(CVPR'19)[19]	39.5*	68.9*	50.2*	-	-	-	-	-	-	45.6	88.7	60.2
GZLOCD(CVPR'20)[12]	44. 8*	59. 9*	51. 3*	-	-	-	-	-	-	59.5	73.4	65.7
GDAN(NIPS'19)[11]	39. 3*	66. 7*	49. 5*	-	-	-	-	-	-	32. 1	67. 5	43. 5
DASCN(NIPS'19)[16]	45. 9*	59. 0*	51. 6*	-	-	-	59. 3	68. 0	63. 4	-	-	-
SGAL(NIPS'19)[31]	40. 9*	55. 3*	47. 0*	-	-	-	52. 7	75. 7	62. 2	55. 1	81. 2	65. 6
SE-GZSL(CVPR'18)[23]	41. 5	53. 3	46. 7	-	-	-	56. 3	67. 8	61. 5	58. 3	68. 1	62. 8
CycWGAN(ECCV'18)[8]	47. 9	59. 3	53. 0	**61. 6**	69. 2	65. 2	59. 6	63. 4	59. 8	59. 6	63. 4	59. 8
f-VAEGAN(CVPR'19)[30]	48. 4	60. 1	53. 6	56. 8	**74. 9**	64. 6	-	-	-	57. 6	70. 6	63. 5
ZSML(AAAI'20)[24]	60. 0	52. 1	55. 7	-	-	-	57.4	71.1	63.5	58.9	74.6	65.8
TACO-GZSL	**61.2**	57.7	**59.4**	**60. 6**	**81. 1**	**69. 4**	60.5	71.9,	**65.7**	59.4,	74.2,	**66.0**
TACO-GZSL(312)	51.8*	60. 0*	**55. 6***	**60. 6**	**81. 1**	**69. 4**	60.5	71.9,	**65.7**	59.4,	74.2,	**66.0**
TACO-GZSL(using Φ_2)	**64.7**	65.9	**65.35**	-	-	-	-	-,	-	62.6	75.6	**68.5**
TACO-GZSL(312)(using Φ_2)	55.6*	**67.50***	**61.0***	-	-	-	-	-,	-	62.6	75.6	**68.5**

Table 1: GZSL performance comparison with several baseline and state-of-the-art methods. For fair comparison, all results reported here are *without fine-tuning* the backbone ResNet101 feature extractor. We measure Top-1 accuracy on Unseen(U), Seen(S) classes and their Harmonic mean(H). Best results are highlighted in bold. * indicates result on CUB dataset with only 312 dim attributes (included for fair comparison with other work that use this setting)

Experiments

Model	CUB	AWA1
$S1$ = Baseline Generative Module	51.9	61.1
$S2 = S1$ + Inference module + Joint maximization	52.7	62.5
$S3 = S2$ + Additional features for recognition module	54.4	65.4
$S4 = S3$ + Wasserstein alignment	**55.6**	**65.7**

Table 2: Ablation study of different components of our framework on CUB and AWA1. Result reported is harmonic mean accuracy.

对比

	AWA1				CUB				FLO				SUN		
	U	S	H		U	S	H		U	S	H		U	S	H
f-CLSWGAN	57.9	61.4	59.6		43.7	57.7	49.7		59.0	73.8	65.6		42.6	36.6	39.4
cycle-CLSWGAN	56.9	64.0	60.2		45.7	61.0	52.3		59.2	72.5	65.1		49.4	33.6	40.0
DASCN	59.3	68.0	63.4		45.9	59.0	51.6						42.4	38.5	40.3
Boomerang-GAN	50.3	73.6	59.8		52.3	58.6	55.3		61.3	78.2	68.7		49.3	35.1	41.0
TACO-GZSL	60.5	71.9	65.7		51.8	60.0	55.6		60.6	81.1	69.4				