
Dual Adversarial Semantics-Consistent Network for Generalized Zero-Shot Learning

Jian Ni¹ Shanghang Zhang² Haiyong Xie^{3,4,1}
nj1@mail.ustc.edu.cn shanghaz@andrew.cmu.edu haiyong.xie@ieee.org

¹University of Science and Technology of China, Anhui 230026, China

²Carnegie Mellon University, Pittsburgh, PA 15213, USA

³Advanced Innovation Center for Human Brain Protection, Capital Medical University, Beijing 100054, China

⁴National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data (NEL-PSRPC), Beijing 100041, China

Abstract

Generalized zero-shot learning (GZSL) is a challenging class of vision and knowledge transfer problems in which both seen and unseen classes appear during testing. Existing GZSL approaches either suffer from semantic loss and discard discriminative information at the embedding stage, or cannot guarantee the visual-semantic interactions. To address these limitations, we propose a *Dual Adversarial Semantics-Consistent Network* (referred to as *DASCN*), which learns both primal and dual Generative Adversarial Networks (GANs) in a unified framework for GZSL. In *DASCN*, the primal GAN learns to synthesize inter-class discriminative and semantics-preserving visual features from both the semantic representations of seen/unseen classes and the ones reconstructed by the dual GAN. The dual GAN enforces the synthetic visual features to represent prior semantic knowledge well via semantics-consistent adversarial learning. To the best of our knowledge, this is the first work that employs a novel dual-GAN mechanism for GZSL. Extensive experiments show that our approach achieves significant improvements over the state-of-the-art approaches.

1 Introduction

In recent years, tremendous progress has been achieved across a wide range of computer vision and machine learning tasks with the introduction of deep learning. However, conventional deep learning approaches rely on large amounts of labeled data, thus may suffer from performance decay in problems where only limited training data are available. The reasons are two folds. On the one hand, objects in the real world have a long-tailed distribution, and obtaining annotated data is expensive. On the other hand, novel categories of objects arise dynamically in nature, which fundamentally limits the scalability and applicability of supervised learning models for handling this dynamic scenario when labeled examples are not available.

Tackling such restrictions, zero-shot learning (ZSL) has been researched widely recently, recognized as a feasible solution [16, 24]. ZSL is a learning paradigm that tries to fulfill the ability to correctly categorize objects from previous unseen classes without corresponding training samples. However, conventional ZSL models are usually evaluated in a restricted setting where test samples and the search space are limited to the unseen classes only, as shown in Figure 1. To address the shortcomings of ZSL, GZSL has been considered in the literature since it not only learns information that can be transferred to an unseen class but can also generalize to new data from seen classes well.

ZSL approaches typically adopt two commonly used strategies. The first strategy is to convert tasks into visual-semantic embedding problems [4, 23, 26, 33]. They try to learn a mapping function from

40 the visual space to the semantic space (note that all the classes reside in the semantic space), or to a
 41 latent intermediate space, so as to transfer knowledge from the seen classes to the unseen classes.
 42 However, the ability of these embedding-based ZSL models to transfer semantic knowledge is limited
 43 by the semantic loss and the heterogeneity gap [4]. Meanwhile, since the ZSL model is only trained
 44 with the labeled data from the seen classes, it is highly biased towards predicting the seen classes [3].
 45 The second strategy ZSL approaches typically adopt is to use generative methods to generate various
 46 visual features conditioned on semantic feature vectors [7, 10, 19, 29, 35], which circumvents the
 47 need for labeled samples of unseen classes and boosts the ZSL classification accuracy. Nevertheless,
 48 the performance of these methods is limited either by capturing the visual distribution information
 49 via only a unidirectional alignment from the class semantics to the visual feature only, or by adopting
 50 just a single Euclidean distance as the constraint to preserve the semantic information between the
 51 generated high-level visual features and real semantic features. Recent work has shown that the
 performance of most ZSL approaches drops significantly in the GZSL setting [28].

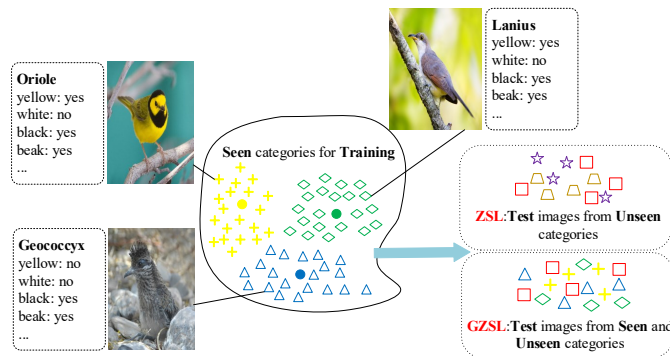


Figure 1: Problem illustration of zero-shot learning (ZSL) and generalized zero-shot learning (GZSL).

52
 53 To address these limitations, we propose a novel Dual Adversarial Semantics-Consistent Network
 54 (referred to as DASCN) for GZSL. DASCN is based on the Generative Adversarial Networks (GANs),
 55 and is characterized by its dual structure which enables bidirectional synthesis by allowing both
 56 the common visual features generation and the corresponding semantic features reconstruction, as
 57 shown in Figure 2. Such bidirectional synthesis procedures available in DASCN boost these two tasks
 58 jointly and collaboratively, preserving the visual-semantic consistency. This results in two advantages
 59 as follows. First, our generative model synthesizes inter-class discrimination visual features via
 60 a classification loss constraint, which makes sure that synthetic visual features are discriminative
 61 enough among different classes. Second, our model encourages the synthesis of visual features that
 62 represent their semantic features well and are of a highly discriminative semantic nature from the
 63 perspectives of both form and content. From the form perspective, the semantic reconstruction error
 64 between the synthetic semantic features (reconstructed by the dual GAN from the pseudo visual
 65 features generated by the primal GAN) and the real corresponding semantic features is minimized to
 66 ensure that the reconstructed semantic features are tightly centered around the real corresponding
 67 class semantics. From the content perspective, the pseudo visual features (generated via the primal
 68 GAN by further exploiting the reconstructed semantic features as input) are constrained to be as close
 69 as possible to their respective real visual features in the data distribution. Therefore, our approach
 70 can ensure that the reconstructed semantic features are consistent with the relevant real semantic
 71 knowledge, thus avoiding semantic loss to a large extent.

72 We summarize our contributions as follows. First, we propose a novel generative dual adversarial
 73 architecture for GZSL, which preserves semantics-consistency effectively with a bidirectional align-
 74 ment, alleviating the issue of semantic loss. To the best of our knowledge, DASCN is the first network
 75 to employ the dual-GAN mechanism for GZSL. Second, by combining the classification loss and
 76 the semantics-consistency adversarial loss, our model generates high-quality visual features with
 77 inter-class separability and a highly discriminative semantic nature, which is crucial to the generative
 78 approaches used in GZSL. Last but not least, we conduct comprehensive experiments demonstrating
 79 that DASCN is highly effective and outperforms the state-of-the-art GZSL methods consistently.

80 The remainder of this paper is structured as follows. We discuss the related work in Section 2, present
 81 our DASCN model in Section 3, evaluate the proposed model in Section 4, and conclude in Section 5.

82 2 Related Work

83 2.1 Zero-Shot Learning

84 Some of the early ZSL works make use of the primitive attributes prediction and classification, such as
85 DAP [15], and IAP [16]. Recently, the attribute-based classifier has evolved into the embedding-based
86 framework, which now prevails due to its simple and effective paradigm [1, 23, 24, 25, 33]. The core
87 of such approaches is to learn a projection from visual space to semantic space spanned by class
88 attributes [23, 24], or conversely [33], or jointly learn an appropriate compatibility function between
89 the visual space and the semantic space [1, 25].

90 The main disadvantage of the above methods is that the embedding process suffers from semantic loss
91 and the lack of visual training data for unseen classes, thus biasing the prediction towards the seen
92 classes and undermining seriously the performance of models in the GZSL setting. More recently,
93 generative approaches are promising for GZSL setting by generating labeled samples for the seen
94 and unseen classes. [10] synthesize samples by approximating the class conditional distribution
95 of the unseen classes based on learning that of the seen classes. [29, 35] apply GAN to generate
96 visual features conditioned on class descriptions or attributes, which ignore the semantics-consistency
97 constraint and allow the production of synthetic visual features that may be too far from the actual
98 distribution. [7] consider minimizing L2 norm between real semantics and reconstructed semantics
99 produced by a pre-trained regressor, which is rather weak and unreliable to preserve high-level
100 semantics via the Euclidean distance.

101 DASCN differs from the above approaches in that it learns the semantics effectively via multi-
102 adversarial learning from both the form and content perspectives. Note that ZSL is also closely
103 related to domain adaptation and image-to-image translation tasks, where all of them assume the
104 transfer between source and target domains. Our approach is motivated by, and is similar in spirit
105 to, recent work on synthesizing samples for GZSL [29] and unpaired image-to-image translation
106 [11, 30, 34]. DASCN preserves the visual-semantic consistency by employing dual GANs to capture
107 the visual and semantic distributions, respectively.

108 2.2 Generative Adversarial Networks

109 As one of the most promising generative models, GANs have achieved a series of impressive results.
110 The idea behind GANs is to learn a generative model to capture an arbitrary data distribution via a
111 max-min training procedure, which consists of a generator and a discriminator that work against each
112 other. DCGAN [22] extends GAN by leveraging deep convolution neural networks. InfoGAN [5]
113 maximizes the mutual information between the latent variables and generator distribution. In our
114 work, given stabilizing training behavior and eliminating model collapse as much as possible, we
115 utilize WGANs [9] as basic models in a dual structure.

116 3 Methodology

117 In this section, we first formalize the GZSL task in Section 3.1. Then we present our model and
118 architecture in Section 3.2. We then describe in detail our model’s objective, training procedures and
119 generalized zero-shot recognition in Section 3.3, Section 3.4 and Section 3.5, respectively.

120 3.1 Formulation

121 We denote by $D^{Tr} = \{(x, y, a) | x \in \mathcal{X}, y \in \mathcal{Y}^s, a \in \mathcal{A}\}$ the set of N^s training instances of the
122 seen classes. Note that $x \in \mathcal{X} \subseteq \mathbb{R}^K$ represents K -dimensional visual features extracted from
123 convolution neural networks, \mathcal{Y}^s denotes the corresponding class labels, and $a \in \mathcal{A} \subseteq \mathbb{R}^L$ denotes
124 semantic features (e.g., the attributes of seen classes). In addition, we have a disjoint class label set
125 $\mathcal{U} = \{(y, a) | y \in \mathcal{Y}^u, a \in \mathcal{A}\}$ of unseen classes, where visual features are missing. Given D^{Tr} and
126 \mathcal{U} , in GZSL, we learn a prediction: $\mathcal{X} \rightarrow \mathcal{Y}^s \cup \mathcal{Y}^u$. Note that our method is of the inductive school
127 where model has access to neither visual nor semantic information of unseen classes in the training
128 phase.

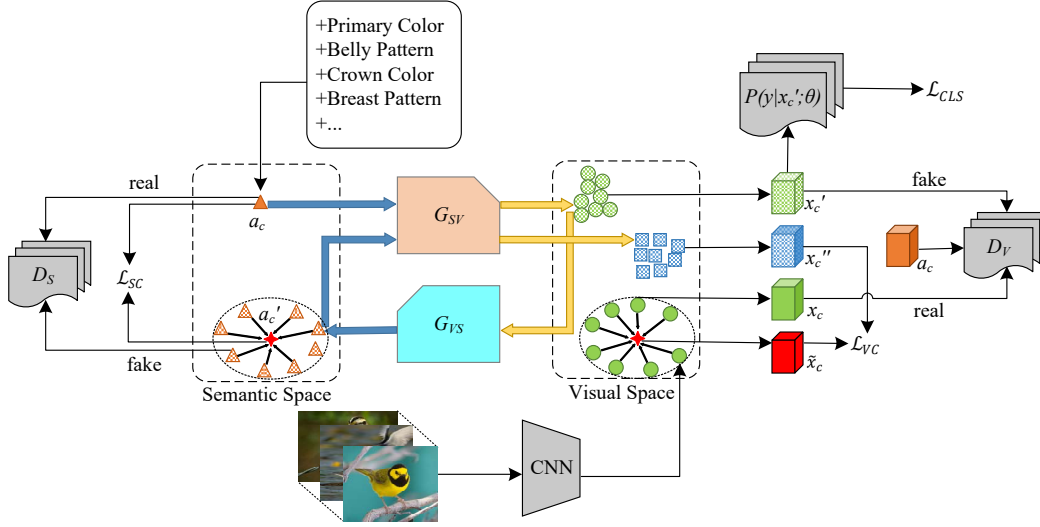


Figure 2: Network architecture of DASCN. The semantic feature of class c , represented as a_c , and a group of randomly sampled noise vectors are utilized by generator G_{SV} to synthesize pseudo visual features x_c' . Then the synthesized visual features are used by generator G_{VS} and discriminator D_V simultaneously to perform semantics-consistency constraint in the perspective of form and content and distinguish between real visual features x_c and synthesized visual features x_c' . D_S denotes the discriminator that distinguishes between a_c , and the reconstructed semantic feature a_c' generated from corresponding x_c' . x_c'' are produced by generator G_{SV} taking a_c' and sampled noise as input to perform visual consistency constraint. Please zoom to view better.

129 3.2 Model Architecture

130 Given the training data D^{Tr} of the seen classes, the primal task of DASCN is to learn a generator
 131 $G_{SV}: \mathcal{Z} \times \mathcal{A} \rightarrow \mathcal{X}$ that takes the random Gaussian noise $z \in \mathcal{Z}$ and semantic attribute $a \in \mathcal{A}$ as input
 132 to generate the visual feature $x' \in \mathcal{X}$, while the dual task is to train an inverse generator $G_{VS}: \mathcal{X} \rightarrow \mathcal{A}$.
 133 Once the generator G_{SV} learns to generate visual features of the seen classes conditioned on the seen
 134 class-level attributes, it can also generate that of the unseen classes. To realize this, we employ two
 135 WGANs, the primal GAN and the dual GAN. The primal GAN consists of the generator G_{SV} and the
 136 discriminator D_V that discriminates between fake visual features generated by G_{SV} and real visual
 137 features. Similarly, the dual GAN learns a generator G_{VS} and a discriminator D_S that distinguishes
 138 the fake semantic features generated by G_{VS} from the real data.

139 The overall architecture and data flow are illustrated in Figure 2. In the primal GAN, we hallucinate
 140 pseudo visual features $x_c' = G_{SV}(a_c, z)$ of the class c using G_{SV} based on corresponding class
 141 semantic features a_c and then put the real visual features and synthetic features from G_{SV} into D_V to
 142 be evaluated. To ensure that G_{SV} generates inter-class discrimination visual features, inspired by that
 143 work [29], we design a classifier trained on the real visual features and minimize the classification
 144 loss over the generated features. It is formulated as:

$$\mathcal{L}_{CLS} = -E_{x' \sim P_{x'}}[\log P(y|x'; \theta)] \quad (1)$$

145 where x' represents the generated visual feature, y is the class label of x' , the conditional probability
 146 $P(y|x'; \theta)$ is computed by a linear softmax classifier parameterized by θ .

147 Following that is one of the main innovations in our work that we guarantee semantics-consistency
 148 in both form and content perspectives thanks to dual structure. In form, $G_{SV}(a, z)$ are translated
 149 back to semantic space using G_{VS} , which outputs $a' = G_{VS}(G_{SV}(a, z))$ as the reconstruction of
 150 a . To achieve the goal that the generated semantic features of each class are distributed around the
 151 corresponding true semantic representation, we design the centroid regularization that regularizes the
 152 mean of generated semantic features of each class to approach respectively real semantic embeddings

153 so as to maintain semantics-consistency to a large extent. The regularization is formulated as:

$$\mathcal{L}_{SC} = \frac{1}{C} \sum_{c=1}^C \left\| E_{a_{c'} \sim P_{a'}^c} [a_{c'}] - a_c \right\|_2 \quad (2)$$

154 where C is the number of seen classes, a_c is the semantic feature of class c , $P_{a'}^c$ denotes the conditional
155 distribution of generated semantic features of class c , $a_{c'}$ are the generated features of class c and the
156 centroid is formulated as:

$$E_{a_{c'} \sim P_{a'}^c} [a_{c'}] = \frac{1}{N_c^s} \sum_{i=1}^{N_c^s} G_{VS}(G_{SV}(a_c, z_i)) \quad (3)$$

157 where N_c^s is the number of generated semantic features of class c . We employ the centroid regulariza-
158 tion to encourage G_{VS} to reconstruct semantic features of each seen class that statistically match real
159 features of that class. From the content point of view, the question of how well the pseudo semantic
160 features a' are reconstructed can be translated into the evaluation of the visual features obtained by
161 G_{SV} taking a' as input. Motivated by the observation that visual features have a higher intra-class
162 similarity and relatively lower inter-class similarity, we introduce the visual consistency constraint:

$$L_{VC} = \frac{1}{C} \sum_{c=1}^C \left\| E_{x_{c''} \sim P_{x''}^c} [x_{c''}] - E_{x_c \sim P_x^c} [x_c] \right\|_2 \quad (4)$$

163 where x_c is the visual features of class c , $x_{c''}$ is the pseudo visual feature generated by generator
164 G_{SV} employing $G_{VS}(G_{SV}(a_c, z))$ as input, P_x^c and $P_{x''}^c$ are conditional distributions of real and
165 synthetic features respectively and the centroid of $x_{c''}$ is formulated as:

$$E_{x_{c''} \sim P_{x''}^c} [x_{c''}] = \frac{1}{N_c^s} \sum_{i=1}^{N_c^s} G_{SV}(G_{VS}(G_{SV}(a_c, z_i)), z_i') \quad (5)$$

166 It is worth nothing that our model is constrained in terms of both form and content aspects to achieve
167 the goal of retaining semantics-consistency and achieves superior results in extensive experiments.

168 3.3 Objective

169 Given the issue that the Jenson-Shannon divergence optimized by the traditional GAN leads to
170 instability during training, our model is based on two WGANs that leverage the Wasserstein distance
171 between two distributions as the objectives. The corresponding loss functions used in the primal
172 GAN are defined as follows. First,

$$L_{DV} = E_{x' \sim P_{x'}} [D_V(x', a)] - E_{x \sim P_{data}} [D_V(x, a)] + \lambda_1 E_{\hat{x} \sim P_{\hat{x}}} \left[\left(\left\| \nabla_{\hat{x}} D_V(\hat{x}, a) \right\|_2 - 1 \right)^2 \right] \quad (6)$$

173 where $\hat{x} = \alpha x + (1 - \alpha)x'$ with $\alpha \sim U(0, 1)$, λ_1 is the penalty coefficient, the first two terms
174 approximate Wasserstein distance of the distributions of fake features and real features, the third term
175 is the gradient penalty. Second, the loss function of the generator of the primal GAN is formulated as:

$$L_{G_{SV}} = -E_{x' \sim P_{x'}} [D_V(x', a)] - E_{a' \sim P_{a'}} [D_V(x'', a')] + \lambda_2 L_{CLS} + \lambda_3 L_{VC} \quad (7)$$

177 where the first two terms are Wasserstein loss, the third term is the classification loss corresponding
178 to class labels, the fourth term is visual consistency constraint introduced before, and $\lambda_1, \lambda_2, \lambda_3$ are
179 hyper-parameters.

180 Similarly, the loss functions of the dual GAN are formulated as:

$$L_{DS} = E_{a' \sim P_{a'}} [D_S(a')] - E_{a \sim P_a} [D_S(a)] + \lambda_4 E_{\hat{y} \sim P_{\hat{y}}} \left[\left(\left\| \nabla_{\hat{y}} D_S(\hat{y}) \right\|_2 - 1 \right)^2 \right] \quad (8)$$

181

$$L_{G_{VS}} = -E_{a' \sim P_{a'}} [D_S(a')] + \lambda_5 L_{SC} + \lambda_6 L_{VC} \quad (9)$$

182 In Eq. (8) and Eq. (9), $\hat{y} = \beta a + (1 - \beta)a'$ is the linear interpolation of the real semantic feature a
183 and the fake a' , and $\lambda_4, \lambda_5, \lambda_6$ are hyper-parameters weighting the constraints.

Table 1: Datasets used in our experiments, and their statistics

Dataset	Semantics/Dim	# Image	# Seen Classes	# Unseen Classes
CUB	A/312	11788	150	50
SUN	A/102	14340	645	72
AWA1	A/85	30475	40	10
aPY	A/64	15339	20	12

184 3.4 Training Procedure

185 We train the discriminators to judge features as real or fake and optimize the generators to fool the
 186 discriminator. To optimize the DASCN model, we follow the training procedure proposed in WGAN
 187 [9]. The training procedure of our framework is summarized in Algorithm 1. In each iteration, the
 188 discriminators D_V , D_S are optimized for n_1 , n_2 steps using the loss introduced in Eq. (6) and Eq.
 189 (8) respectively, and then one step on generators with Eq. (7) and Eq. (9) after the discriminators
 190 have been trained. According to [30], such a procedure enables the discriminators to provide more
 191 reliable gradient information. The training for traditional GANs suffers from the issue that the
 192 sigmoid cross-entropy is locally saturated as discriminator improves, which may lead to vanishing
 193 gradient and need to balance discriminator and generator carefully. Compared to the traditional
 194 GANs, the Wasserstein distance is differentiable almost everywhere and demonstrates its capability
 195 of extinguishing mode collapse. We put the detailed algorithm for training DASCN model in the
 196 supplemental material.

197 3.5 Generalized Zero-Shot Recognition

198 With the well-trained generative model, we can elegantly generate labeled exemplars of any class
 199 by employing the unstructured component z resampled from random Gaussian noise and the class
 200 semantic attribute a_c into the G_{SV} . An arbitrary number of visual features can be synthesized and
 201 those exemplars are finally used to train any off-the-shelf classification model. For simplicity, we
 202 adopt a softmax classifier. Finally, the prediction function for an input test visual feature v is:

$$f(v) = \arg \max_{y \in \tilde{\mathcal{Y}}} P(y|v; \theta') \quad (10)$$

203 where $\tilde{\mathcal{Y}} = \mathcal{Y}^s \cup \mathcal{Y}^u$ for GZSL.

204 4 Experiments

205 4.1 Datasets and Evaluation Metrics

206 To test the effectiveness of the proposed model for GZSL, we conduct extensive evaluations on
 207 four benchmark datasets: CUB [27], SUN [21], AWA1 [15], aPY [6] and compare the results with
 208 state-of-the-art approaches. Statistics of the datasets are presented in Table 1. For all datasets, we
 209 extract 2048 dimensional visual features via the 101-layered ResNet from the entire images, which is
 210 the same as [29]. For fair comparison, we follow the training/validation/testing split as described in
 211 [28].

212 During test time, in the GZSL setting, the search space includes both the seen and unseen classes, i.e.
 213 $\mathcal{Y}^u \cup \mathcal{Y}^s$. To evaluate the GZSL performance over all classes, the following measures are applied. (1)
 214 ts: average per-class classification accuracy on test images from the unseen classes with the prediction
 215 label set being $\mathcal{Y}^u \cup \mathcal{Y}^s$. (2) tr: average per-class classification accuracy on test images from the seen
 216 classes with the prediction label set being $\mathcal{Y}^u \cup \mathcal{Y}^s$. (3) H: the harmonic mean of above defined tr and
 217 ts, which is formulated as $H = (2 \times ts \times tr) / (ts + tr)$ and quantities the aggregate performance
 218 across both seen and unseen test classes. We hope that our model is of high accuracy on both seen
 219 and unseen classes.

220 4.2 Implementation Details

221 Our implementation is based on PyTorch. DASCN consists of two generators and two discriminators:
 222 G_{SV} , G_{VS} , D_V , D_S . We train specific models with appropriate hyper-parameters. Due to the space

Table 2: Evaluations on four benchmark datasets. *indicates that Cycle-WGAN employs 1024-dim per-class sentences as class semantic rather than 312-dim per-class attributes on CUB, whose results on CUB may not be directly comparable with others.

Method	AWA1			SUN			CUB			aPY		
	ts	tr	H	ts	tr	H	ts	tr	H	ts	tr	H
CMT [24]	0.9	87.6	1.8	8.1	21.8	11.8	7.2	49.8	12.6	1.4	85.2	2.8
DEVISE [8]	13.4	68.7	22.4	16.9	27.4	20.9	23.8	53.0	32.8	4.9	76.9	9.2
ESZSL [23]	6.6	75.6	12.1	11.0	27.9	15.8	12.6	63.8	21.0	2.4	70.1	4.6
SJE [1]	11.3	74.6	19.6	14.7	30.5	19.8	23.5	59.2	33.6	3.7	55.7	6.9
SAE [13]	1.8	77.1	3.5	8.8	18.0	11.8	7.8	54.0	13.6	0.4	80.9	0.9
LESAE [18]	19.1	70.2	30.0	21.9	34.7	26.9	24.3	53.0	33.3	12.7	56.1	20.1
SP-AEN [4]	-	-	-	24.9	38.2	30.3	34.7	70.6	46.6	13.7	63.4	22.6
RN [25]	31.4	91.3	46.7	-	-	-	38.1	61.1	47.0	-	-	-
TRIPLE [31]	27	67.9	38.6	22.2	38.3	28.1	26.5	62.3	37.2	-	-	-
f-CLSWGAN [29]	57.9	61.4	59.6	42.6	36.6	39.4	43.7	57.7	49.7	-	-	-
KERNEL [32]	18.3	79.3	29.8	19.8	29.1	23.6	19.9	52.5	28.9	11.9	76.3	20.5
PSR [2]	-	-	-	20.8	37.2	26.7	24.6	54.3	33.9	13.5	51.4	21.4
DCN [17]	25.5	84.2	39.1	25.5	37	30.2	28.4	60.7	38.7	14.2	75.0	23.9
SE-GZSL [14]	56.3	67.8	61.5	40.9	30.5	34.9	41.5	53.3	46.7	-	-	-
GAZSL [35]	25.7	82.0	39.2	21.7	34.5	26.7	23.9	60.6	34.3	14.2	78.6	24.1
DASCN (Ours)	59.3	68.0	63.4	42.4	38.5	40.3	45.9	59.0	51.6	39.7	59.5	47.6
Cycle-WGAN* [7]	56.4	63.5	59.7	48.3	33.1	39.2	46.0	60.3	52.2	-	-	-

223 limitation, here we take CUB as an example. Both the generators and discriminators are MLP with
 224 LeakyReLU activation. In the primal GAN, G_{SV} has a single hidden layer containing 4096 nodes and
 225 an output layer that has a ReLU activation with 2048 nodes, while the discriminator D_V contains a
 226 single hidden layer with 4096 nodes and an output layer without activation. G_{VS} and D_S in the dual
 227 GAN have similar architecture with G_{SV} and D_V respectively. We use $\lambda_1 = \lambda_4 = 10$ as suggested
 228 in [9]. For loss term contributions, we cross-validate and set $\lambda_2 = \lambda_3 = \lambda_6 = 0.01$, $\lambda_5 = 0.1$. We
 229 choose noise z with the same dimensionality as the class embedding. Our model is optimized by
 230 Adam [12] with a base learning rate of $1e^{-4}$.

231 4.3 Compared Methods and Experimental Results

232 We compare DASCN with state-of-the-art GZSL models. These approaches fall into two categories.
 233 (1) Embedding-based approaches: CMT [24], DEVISE [8], ESZSL [23], SJE [1], SAE [13], LESAE
 234 [18], SP-AEN [4], RN [25], KERNEL [32], PSR [2], DCN [17], TRIPLE [31]. This category suffers
 235 from the issue of the bias towards seen classes due to the lack of instances of the unseen classes. (2)
 236 Generative approaches: SE-GZSL [14], GAZSL [35], f-CLSWGAN [29], Cycle-WGAN [7]. This
 237 category synthesizes visual features of the seen and unseen classes and perform better for GZSL
 238 compared to the embedding-based methods.

239 Table 2 summarizes the performance of all the comparing methods under three evaluation metrics on
 240 the four benchmark datasets, which demonstrates that for all datasets our DASCN model significantly
 241 improves the ts measure and H measure over the state-of-the-arts. Note that Cycle-WGAN [7]
 242 employs per-class sentences as class semantic features on CUB dataset rather than per-class attributes
 243 that are commonly used by other comparison methods, so its results on CUB may not be directly
 244 comparable with others. On CUB, DASCN achieves 45.9% in ts and 51.6% in H, with improvements
 245 over the state-of-the-art 2.2% and 1.9% respectively. On SUN, it obtains 42.4% in ts measure and
 246 40.3% in H measure. On AWA1, our model outperforms the runner-up by a considerable gap in H
 247 measure: 1.9%. On aPY, DASCN significantly achieves improvements over the other best competitors
 248 25.5% in ts measure and 23.5% in H measure, which is very impressive. The performance boost is
 249 attributed to the effectiveness of DASCN that imitate discriminative visual features of the unseen
 250 classes. In conclusion, our model DASCN achieves a great balance between seen and unseen classes
 251 classification and consistently outperforms the current state-of-the-art methods for GZSL.

Table 3: Comparison between the reported results of Cycle-WGAN and our model. * indicates employing the same semantic features (per-class sentences (stc)) as Cycle-WGAN on CUB.

	FLO			CUB*			SUN			AWA1		
Method	ts	tr	H	ts	tr	H	ts	tr	H	ts	tr	H
Cycle-WGAN	59.1	71.1	64.5	46.0	60.3	52.2	48.3	33.1	39.2	56.4	63.5	59.7
DASCN	60.5	80.4	69.0	47.4	60.1	53.0	42.4	38.5	40.3	59.3	68.0	63.4

To further clarify the advantages of DASCN over Cycle-WGAN [7] in both methodology and empirical results, we conduct the following experiments: (1) we use the same semantic features (per-class sentences (stc)) as Cycle-WGAN uses for DASCN on the CUB dataset, (2) we add the FLO [6] dataset employed by Cycle-WGAN as a benchmark. As shown in Table 3, results on four benchmarks consistently demonstrate the superiority of DASCN. The main novelty of our work is the integration of dual structure mechanism and visual-semantic consistencies into GAN for bidirectional alignment and alleviating semantic loss. In contrast, Cycle-WGAN only consists of one GAN and a pre-trained regressor, which only minimizes L2 norm between the reconstructed and real semantics. As a result, Cycle-WGAN is rather weak and unreliable to preserve high-level semantics via the Euclidean distance. Compared to that, thanks to the dual-GAN structure and visual-semantic consistencies loss, DASCN explicitly supervises that the generated features have highly discriminative semantic nature on the high-level aspects and effectively preserve semantics via multi-adversarial learning in both form and content perspectives.

More specifically, we build two GANs for visual and semantic generation, and design two consistency regularizations accordingly: (1) semantic consistency to align the centroid of the synthetic semantics and real semantic, (2) visual consistency for not only matching the real visual features but also enforcing synthetic semantics to have highly discriminative nature to further generate effective visual features. Compared to the Cycle-WGAN that only minimizes L2 norm of reconstructed and real semantics, the novelty being introduced is the tailor-made semantic high-level consistency at a finer granularity.

Note that we not only generate synthetic semantic features from the synthetic visual features, but also further generate synthetic visual features again based on the synthetic semantic features, which is constrained by visual consistency loss to ensure the generated features have highly discriminative semantic nature. Such bidirectional synthesis procedures boost the quality of synthesized instances collaboratively via dual structure.

4.4 Ablation Study

We now conduct the ablation study to evaluate the effects of the dual structure, the semantic centroid regularization \mathcal{L}_{SC} , and the visual consistency constraint \mathcal{L}_{VC} . We take the single WGAN model f-CLSWGAN as baseline, and train three variants of our model by keeping the single dual structure or that adding the only semantic or visual constraint, denoted as Dual-WGAN, Dual-WGAN + \mathcal{L}_{SC} , Dual-WGAN + \mathcal{L}_{VC} , respectively. Table 4 shows the performance of each setting, the performance of the single Dual-WGAN on the H metric drops drastically by 4.9% on aPY, 1.4% on AWA1, 1.3% on CUB and 0.7% on SUN, respectively. This clearly highlights the importance of designed semantic and visual constraints to provide an explicit supervision to our model. In the case of lacking semantic or visual unidirectional constrains, on aPY, our model drops by 1.3% and 3.6% respectively, while on AWA1 the gap are 0.9% and 0.7%. In general, the three variants of our proposed model tend to offer more superior and balanced performance than the baseline. DASCN incorporates dual structure, semantic centroid regularization and visual consistency constraint into a unified framework and achieves the best improvement, which demonstrates that different components promote each other and work together to improve the performance of DASCN significantly.

4.5 Quality of Synthesized Samples

We perform an experiment to gain a further insight into the quality of the generated samples, which is one key issue of our approach, although the quantitative results reported for GZSL above demonstrate that the samples synthesized by our model are of significant effectiveness for GZSL task. Specifically, we randomly sample three unseen categories from aPY and visualize both true visual features and

Table 4: Effects of different components on four benchmark datasets with GZSL setting.

Methods	aPY			AWA1			CUB			SUN		
	ts	tr	H	ts	tr	H	ts	tr	H	ts	tr	H
WGAN-baseline	32.4	57.5	41.4	57.9	61.4	59.6	43.7	57.7	49.7	42.6	36.6	39.4
Dual-WGAN	34.1	57.0	42.7	57.5	67.4	62.0	44.5	57.9	50.3	42.7	36.9	39.6
Dual-WGAN + \mathcal{L}_{SC}	35.4	58.2	44.0	57.7	68.6	62.7	44.9	58.5	50.8	42.9	37.3	39.9
Dual-WGAN + \mathcal{L}_{VC}	36.7	62.0	46.3	58.3	67.3	62.5	45.2	59.1	51.2	43.5	36.5	39.7
DASCN	39.7	59.5	47.6	59.3	68.0	63.4	45.9	59.0	51.6	42.4	38.5	40.3

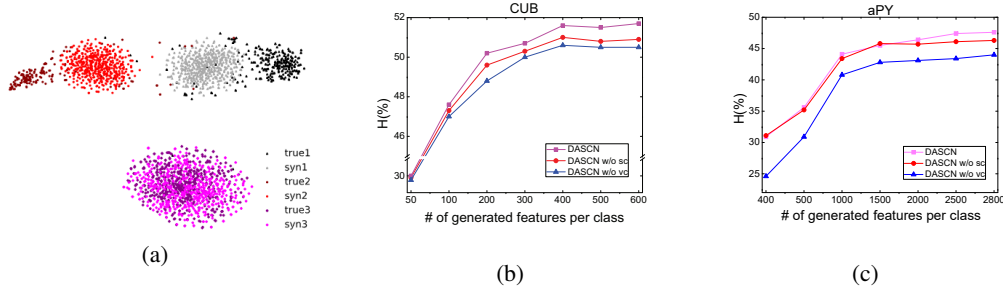


Figure 3: (a): t-SNE visualization of real visual feature distribution and synthesized feature distribution from randomly selected three unseen classes; (b, c): Increasing the number of samples generated by DASCN and its variants wrt harmonic mean H. DASCN w/o SC denotes DASCN without semantic consistency constraint and DASCN w/o VC stands for that without visual consistency constraint.

297 synthesized visual features using t-SNE [20]. Figure 3(a) depicts the empirical distributions of the
 298 true visual features and the synthesized visual features. We observe the clear patterns of intra-class
 299 diversity and inter-class separability in the figure. This intuitively demonstrates that not only the
 300 synthesized feature distributions well approximate the true distributions but also our model introduces
 301 a high discriminative power of the synthesized features to a large extent.

302 Finally, we evaluate how the number of the generated samples per class affects the performance of
 303 DASCN and its variants. Obviously, as shown in Figure 3(b) and Figure 3(c), we notice not only that
 304 H increases with an increasing number of synthesized samples and asymptotes gently, but also that
 305 DASCN with visual-semantic interactions achieves better performance in all circumstance, which
 306 further validates the superiority and rationality of different components of our model.

307 5 Conclusion

308 We propose DASCN, a novel generative model for GZSL, to address the challenging problem
 309 where existing GZSL approaches either suffer from the semantic loss or cannot guarantee the visual-
 310 semantic interactions. DASCN can synthesize inter-class discrimination and semantics-preserving
 311 visual features for both seen and unseen classes. The DASCN architecture is novel in that it
 312 consists of a primal GAN and a dual GAN to collaboratively promote each other, which captures the
 313 underlying data structures of both visual and semantic representations. Thus, our model can effectively
 314 enhance the knowledge transfer from the seen categories to the unseen ones, and can effectively
 315 alleviate the inherent semantic loss problem for GZSL. We conduct extensive experiments on four
 316 benchmark datasets and compare our model against the state-of-the-art models. The evaluation
 317 results consistently demonstrate the superiority of DASCN to state-of-the-art GZSL models.

318 Acknowledgments

319 This research is supported in part by the National Key Research and Development Project (Grant No.
 320 2017YFC0820503), the National Science and Technology Major Project for IND (investigational
 321 new drug) (Project No. 2018ZX09201014), and the CETC Joint Advanced Research Foundation
 322 (Grant No. 6141B08080101,6141B08010102).

References

- 323
- 324 [1] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output
325 embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on*
326 *Computer Vision and Pattern Recognition*, pages 2927–2936, 2015.
- 327 [2] Yashas Annadani and Soma Biswas. Preserving semantic relations for zero-shot learning.
328 In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages
329 7603–7612, 2018.
- 330 [3] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and
331 analysis of generalized zero-shot learning for object recognition in the wild. In *European*
332 *Conference on Computer Vision*, pages 52–68. Springer, 2016.
- 333 [4] Long Chen, Hanwang Zhang, Jun Xiao, Wei Liu, and Shih-Fu Chang. Zero-shot visual
334 recognition using semantics-preserving adversarial embedding networks. In *Proceedings of the*
335 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1043–1052, 2018.
- 336 [5] Xi Chen, Yan Duan, Rein Houthoofd, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan:
337 Interpretable representation learning by information maximizing generative adversarial nets. In
338 *Advances in neural information processing systems*, pages 2172–2180, 2016.
- 339 [6] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes.
340 In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1785.
341 IEEE, 2009.
- 342 [7] Rafael Felix, Vijay BG Kumar, Ian Reid, and Gustavo Carneiro. Multi-modal cycle-consistent
343 generalized zero-shot learning. In *Proceedings of the European Conference on Computer Vision*
344 *(ECCV)*, pages 21–37, 2018.
- 345 [8] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. De-
346 vise: A deep visual-semantic embedding model. In *Advances in neural information processing*
347 *systems*, pages 2121–2129, 2013.
- 348 [9] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville.
349 Improved training of wasserstein gans. In *Advances in neural information processing systems*,
350 pages 5767–5777, 2017.
- 351 [10] Yuchen Guo, Guiguang Ding, Jungong Han, and Yue Gao. Synthesizing samples fro zero-shot
352 learning. *IJCAI*, 2017.
- 353 [11] Taeksoo Kim, Moon-su Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to
354 discover cross-domain relations with generative adversarial networks. In *Proceedings of the*
355 *34th International Conference on Machine Learning-Volume 70*, pages 1857–1865. JMLR. org,
356 2017.
- 357 [12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint*
358 *arXiv:1412.6980*, 2014.
- 359 [13] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning.
360 In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages
361 3174–3183, 2017.
- 362 [14] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot
363 learning via synthesized examples. In *Proceedings of the IEEE conference on computer vision*
364 *and pattern recognition*, pages 4281–4289, 2018.
- 365 [15] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen
366 object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision*
367 *and Pattern Recognition*, pages 951–958. IEEE, 2009.
- 368 [16] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification
369 for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine*
370 *Intelligence*, 36(3):453–465, 2013.

- 371 [17] Shichen Liu, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Generalized zero-shot
372 learning with deep calibration network. In *Advances in Neural Information Processing Systems*,
373 pages 2005–2015, 2018.
- 374 [18] Yang Liu, Quanyue Gao, Jin Li, Jungong Han, and Ling Shao. Zero shot learning via low-rank
375 embedded semantic autoencoder. In *IJCAI*, pages 2490–2496, 2018.
- 376 [19] Yang Long, Li Liu, Ling Shao, Fumin Shen, Guiguang Ding, and Jungong Han. From zero-shot
377 learning to conventional supervised classification: Unseen visual data synthesis. In *Proceedings*
378 *of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1627–1636, 2017.
- 379 [20] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine*
380 *learning research*, 9(Nov):2579–2605, 2008.
- 381 [21] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and
382 recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern*
383 *Recognition*, pages 2751–2758. IEEE, 2012.
- 384 [22] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with
385 deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- 386 [23] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot
387 learning. In *International Conference on Machine Learning*, pages 2152–2161, 2015.
- 388 [24] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning
389 through cross-modal transfer. In *Advances in neural information processing systems*, pages
390 935–943, 2013.
- 391 [25] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales.
392 Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE*
393 *Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018.
- 394 [26] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings
395 and knowledge graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern*
396 *Recognition*, pages 6857–6866, 2018.
- 397 [27] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie,
398 and Pietro Perona. Caltech-ucsd birds 200. 2010.
- 399 [28] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a
400 comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern*
401 *analysis and machine intelligence*, 2018.
- 402 [29] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks
403 for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern*
404 *recognition*, pages 5542–5551, 2018.
- 405 [30] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for
406 image-to-image translation. In *Proceedings of the IEEE international conference on computer*
407 *vision*, pages 2849–2857, 2017.
- 408 [31] Haofeng Zhang, Yang Long, Yu Guan, and Ling Shao. Triple verification network for general-
409 ized zero-shot learning. *IEEE Transactions on Image Processing*, 28(1):506–517, 2018.
- 410 [32] Hongguang Zhang and Piotr Koniusz. Zero-shot kernel learning. In *Proceedings of the IEEE*
411 *Conference on Computer Vision and Pattern Recognition*, pages 7670–7679, 2018.
- 412 [33] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot
413 learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*,
414 pages 2021–2030, 2017.
- 415 [34] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image
416 translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international*
417 *conference on computer vision*, pages 2223–2232, 2017.

- 418 [35] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative
419 adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE*
420 *conference on computer vision and pattern recognition*, pages 1004–1013, 2018.