






Human-Computer Interactive Cross-Modal Zero Shot Fine-Grained Image Recognition Based on AI-powered CAD

Jingyuan He^{1,2*} , Bailong Yang²  and Yuan Tian³ 

^{1,3} School of Mathematics and Computer Science, Yan'an University, Yan'an 716000, Shaanxi, China, ¹18992118537@163.com, ³yuantian_tech@outlook.com

² Xi'an Research Institute of Hi-Tech, Xi'an 710025, Shaanxi, China, ²byang666@live.com

Corresponding Author: Jingyuan He, 18992118537@163.com

Abstract: Traditional methods may not fully integrate image and text description information when processing multimodal data, resulting in insufficient exploration of semantic relationships between modalities. This article combines image and text descriptions and applies them to zero-shot fine-grained image (FGI) recognition tasks to improve the zero-shot learning ability for new categories. This article selected the CUB-200-2011 dataset, Stanford-Dogs dataset, and Stanford-Cars dataset to provide text descriptions of images in the dataset and perform word segmentation on the text descriptions. It introduced CMSE-GAN (Cross-Modal Semantic Enhancement Generative Adversarial Network). It consisted of three parts: image feature filtering, cross-modal (CM) embedding, and CM generative adversarial networks (GAN). By introducing semantic and discriminative constraints to remove redundant information from visual features, this paper mapped semantic features. It filters image features to the same subspace and trains an adversarial network to generate features that are both visually realistic and contain rich semantic information. The experimental results showed that in traditional zero-shot learning, the recognition accuracy of CMSE-GAN on CUB-200-2011, Stanford-Dogs, and Stanford-Cars was 72.2%, 69.9%, and 67.8%, respectively. In generalized zero-shot learning, CMSE-GAN can still improve FGI recognition performance. The application of CMSE-GAN can effectively improve FGI recognition performance by combining image features and text descriptions.

Keywords: Fine-Grained Image, Image Recognition, Cross-Modal, Traditional Zero-Shot Learning, Generalized Zero-Shot Learning

DOI: <https://doi.org/10.14733/cadaps.2025.S6.264-277>

1 INTRODUCTION

As computer vision and natural language processing develop, multimodal data processing has become the hotspot in the field. Multimodal data usually consist of image and text descriptions,

complementing each other in expressing rich semantic information [1-2]. In image recognition (IR) tasks, to improve the performance of models, researchers have been exploring how to utilize the information between images and related text descriptions fully. However, in zero-shot FGI recognition tasks, i.e., recognizing those fine-grained categories that have not been seen in the training set, traditional methods face challenges [3-4]. Because the difference between fine-grained categories is often very small, and it may be difficult to achieve sufficient differentiation by the image alone [5-6]. Meanwhile, the traditional methods can not integrate textual information and image features organically well when textual descriptions are present, resulting in a limited ability of the model to learn new categories with zero samples.

Zero-shot fine-grained image recognition can overcome the dependence of traditional image recognition methods on a large amount of labeled data [7-8]. Zero-shot FGI recognition is aimed at solving the recognition difficulties on previously unseen fine-grained categories, improving the model's zero-shot learning (ZSL) ability for new categories, and promoting the model's better adaptation to multimodal scenarios in practical applications [9-10]. Yu Jun designed a hierarchical deep word embedding model by integrating sparse constraints and improved operators to solve the problem of predicting click features from visual features [11]. To solve the problem of fine-grained vehicle classification, Li Xiaoxu added a regularization term to the cross entropy loss. He proposed a new dual cross entropy loss and demonstrated good performance on three general image classification tasks [12]. In order to improve the application effect of FGI recognition in multimedia tasks, Rodriguez and Pau used attention in neural networks to select the most informative regions in the image, thereby improving classification [13]. FGI recognition can improve the ability to distinguish subtle differences between similar objects, achieve more accurate classification and recognition, and provide strong support for precise target localization, variety identification, and other fields [14-15]. By using FGI recognition to improve sensitivity to subtle differences in objects, computer systems can more accurately distinguish different subcategories within the same category.

CM IR can be utilized to improve the accuracy of search engines for smarter image retrieval and annotation [16-17]. CM IR fuses and comprehends information from different modalities to achieve more comprehensive and accurate IR and understanding. By combining visual, textual, sound, and other information, it can obtain deeper cognition, which helps better to understand scenes or objects [18-19]. CM IR helps improve recognition accuracy by simultaneously considering information from multiple modalities. When relying solely on image information may lead to ambiguity, combining text descriptions or audio information can provide stronger support and improve overall accuracy [20-21]. In order to improve the effectiveness of gait recognition, Li Guodong proposed a new multimodal gait recognition algorithm based on contour and posture features, which obtains set-level spatiotemporal features through time aggregation operations [22]. Daas Sara proposed a secure multimodal biometric recognition system based on deep learning (DL) methods, and the proposed fusion structure achieved an accuracy of 99.89% and an equal error rate of 0.05% [23]. These scholars, who combine information from multiple modalities, can provide a more comprehensive and diverse perspective and enhance understanding of image content, but they lack the application of CM analysis methods to zero-shot FGI recognition.

In order to improve the performance of FGI recognition, this paper fully integrates image and text descriptions and applies CMSE-GAN to improve the generalization performance of the model. This article selects three types of image datasets, including birds, dogs, and cars. By providing textual descriptions of the images, this article provides additional semantic information for each image. Image data can be standardized, and text descriptions can be segmented. This article integrates features from different modalities through CM feature fusion and transforms image features and semantic features into a unified feature space through CM embedding modules.

2 FGI RECOGNITION METHODS

2.1 Data Collection and Preprocessing

The task of FGI recognition for fine classification in object categories with similar appearances involves distinguishing different subcategories within the same category [24-25]. Compared to traditional image classification tasks, FGI recognition is more challenging. Because the differences between object categories are usually very subtle, the model needs to have high resolution and discrimination ability.

FGI recognition requires a large amount of annotated data, as well as handling the similarity and complexity between categories [26-27]. FGI recognition technology can help biologists classify and identify animals and plants more accurately, thereby promoting species conservation and ecological research. FGI recognition technology is of great significance for achieving the goals of intelligence and automation, providing more accurate and efficient solutions for various industries, and promoting technological progress and social development.

Multiple image datasets are used to analyze the performance of FGI recognition. The image datasets used include the CUB-200-2011 dataset, Stanford-Dogs dataset, and Stanford-Cars dataset [28-29], as displayed in Figure 1.

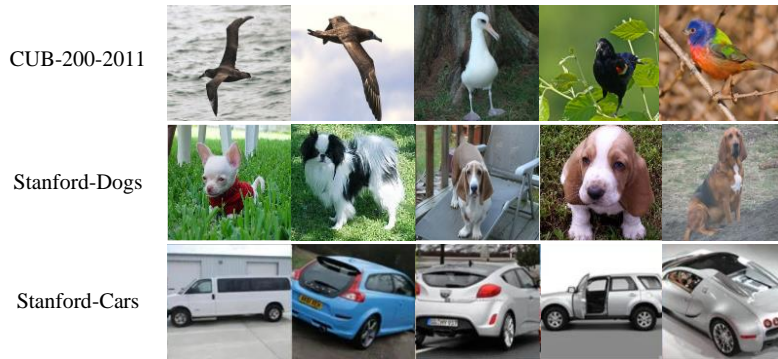




Figure 1: Collected image dataset.

To analyze the subtle differences, this text provides a textual description of the collected images. For animal text descriptions, there are types and postures, while for car text descriptions, there are brands and colors, as displayed in Table 1.

<i>Image</i>	<i>Text description</i>	<i>Kind</i>	<i>Posture</i>
	<i>A black-footed albatross is flying on the sea.</i>	<i>Black-footed albatross</i>	<i>Flying</i>
	<i>A Laysan albatross is flying on the sea.</i>	<i>Laysan albatross</i>	<i>Flying</i>



	<i>A Sooty albatross A Sooty albatross perches in the nest in the nest.</i>	<i>Sooty albatross</i>	<i>Perch</i>
	<i>A Bobolink stands.</i>	<i>Bobolink</i>	<i>Stands</i>

Table 1: Text datasets corresponding to CUB-200-2011.

The corresponding text dataset for Stanford-Dogs is shown in Table 2.





<i>Image</i>	<i>Text description</i>	<i>Kind</i>	<i>Posture</i>
	<i>A Japanese spaniel stands under a large tree.</i>	<i>Japanese spaniel</i>	<i>Stands</i>
	<i>A Maltese dog sits on the lawn.</i>	<i>Maltese dog</i>	<i>Sits</i>
	<i>A Pekinese lies on the ground.</i>	<i>Pekinese</i>	<i>lies</i>
	<i>A Papillon stands on the lawn.</i>	<i>Papillon</i>	<i>Stands</i>

Table 2: Text datasets corresponding to Stanford-Dogs.

By textually describing the images, additional semantic information can be provided for each image, thus increasing the richness and diversity of the dataset. The textual dataset corresponding to Stanford-Cars is shown in Table 3.

<i>Image</i>	<i>Text description</i>	<i>Brand</i>	<i>Colour</i>
--------------	-------------------------	--------------	---------------





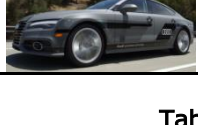
	<i>A white Audi car.</i>	<i>Audi</i>	<i>White</i>
	<i>A gray BMW car.</i>	<i>BMW</i>	<i>Grey</i>
	<i>A blue Audi car.</i>	<i>Audi</i>	<i>Blue</i>
	<i>A white Buick.</i>	<i>Buick</i>	<i>White</i>
	<i>A gray Audi car.</i>	<i>Audi</i>	<i>Grey</i>

Table 3: Textual datasets corresponding to Stanford-Cars.

Text description gives an opportunity for CM learning, i.e., combining textual information with image information and using them together to train a model. Through CM learning, the correlation between images and text can be better exploited to improve the ability of the image recognition task.

To ensure the quality and usability of the collected images, the images are subject to normalization and standardization. Image normalization is the scaling of the pixel values of an image to a fixed range, eliminating the brightness differences between different images and making it easier for the model to learn common features.

The formula for image normalization is:

$$G = \frac{i - \min i}{\max i - \min i} \quad (1)$$

$\min i$ and $\max i$ represent the minimum and maximum pixel values, respectively.

The standardized formula is:

$$S = \frac{i - \mu}{\sigma} \quad (2)$$

In formula 2, μ is the mean of the image pixels and σ represents the standard deviation.

For word segmentation processing of text descriptions, the text descriptions can be cut into words, and word segmentation based on spaces can be selected [30-31]. Word segmentation can cut the text into smaller units, so that the model can better understand and process the semantic information in the text. The result after "A Maltese dog sits on the lawn" participle is ["A", "Maltese", "dog", "sits", "on", "the", "lawn"].

By mapping each word to a high-dimensional vector space through Word2Vec, the semantic relationships between words can be preserved in the vector space. For each word w_i , the embedding vector corresponding to each word is represented as v_i , and the embedding representation of the description Text is:

$$Embedding\ Text = \frac{1}{N} \sum_{j=1}^N v_j \quad (3)$$

2.2 CM Feature Fusion

CM feature fusion can integrate features from different modalities, thereby enriching feature representations. CM feature fusion helps to improve the model's understanding and expression ability of data, making it more comprehensive in capturing the diversity and complexity of data.

The CM semantic enhancement GAN introduced in this article consists of three processes, namely, image feature filtering, CM embedding, and CM GAN. The process of image feature filtering is shown in Figure 2.

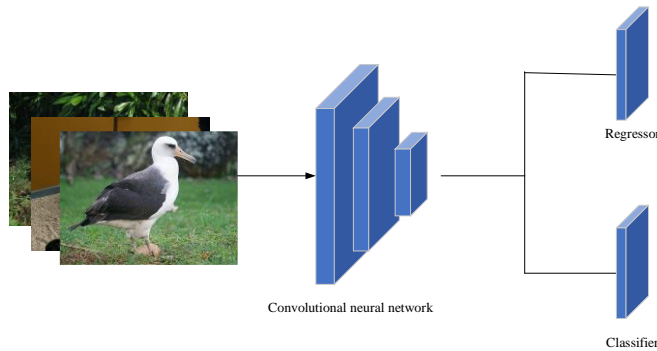


Figure 2: Process of image feature filtering.

The image features extracted through convolutional neural networks typically contain a large amount of label independent information. By introducing semantic and discriminative constraints, redundant information in visual features is removed while retaining key visual information. The classifier can be used as a discriminative constraint, and the regressor can be used as a semantic constraint.

The classifier is trained by learning cross-loss entropy, and the formula for cross-loss entropy is:

$$L(y, \hat{y}) = - \sum_i y_i \log \hat{y}_i \quad (4)$$

To obtain the optimal image features by simultaneously optimizing the loss of the classifier and regressor:

$$G_1 = G_2 + aG_3 \quad (5)$$

In formula 5, G_2 and G_3 represent the loss contributions of the classifier and regressor, respectively. a is a hyperparameter.

Semantic features can be programmed to be projected into a latent space of the same dimension as the filtered image features. The dimensionality-reduced image features and the

mapped semantic features are fed into a shared linear layer for generating semantically enhanced CM features. The process of CM embedding is shown in Figure 3.

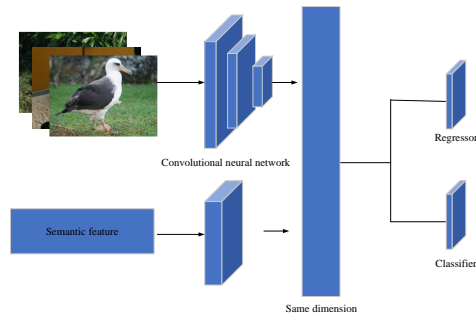


Figure 3: The process of CM embedding.

The CM embedding module is an important component designed to transform image features and semantic features into a unified feature space. The semantic features are projected into a latent space that has the same dimensions as the filtered visual features. This projection process is to ensure that the semantic and visual features are in the same dimension in preparation for subsequent feature fusion and processing. Finally, classifiers and regressors are applied to ensure that CM features do not lose their original discriminative information during embedding and to improve the stability and generalization of the model.

2.3 GAN Training

A Generative Adversarial Network is a deep learning model that consists of two networks: the generator and the discriminator. These two networks compete and collaborate with each other to continuously improve each other's performance by means of an adversarial approach to generate realistic data [32-33]. The role of the generator is to receive random noise or other types of inputs and convert them into data samples similar to the real data [34-35].

By training GAN, the generated visual features contain both semantic information and maintain visual authenticity, thereby increasing the recognition performance of the classifier. The GAN is shown in Figure 4.

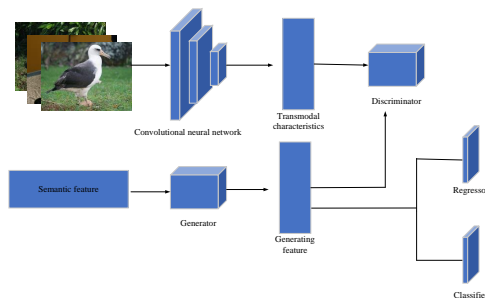


Figure 4: GAN section.

To maintain the stability of the GAN, generation and CM transformation were separated during the generation process. The generator is applied to generate features with the same dimensions, while the discriminator is responsible for distinguishing between generated and real CM features, driving the generator to generate more realistic features.

The generator of the GAN is responsible for generating image features from random noise. During the training process, the generator deceives the discriminator by generating realistic visual features, making it difficult for the discriminator to distinguish the differences between the generated features and the real data. Through this competitive and collaborative training approach, the generator gradually learns how to generate visually realistic features.

The calculation formula for the Softmax function is as follows:

$$\sigma z_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad (6)$$

The Softmax function converts each element into a probability value between 0 and 1, and the sum of all probability values is 1.

3 EXPERIMENTAL EVALUATION

FGI recognition can subdivide objects into more specific categories. This can provide a more accurate understanding and description of the objects in the image, thus meeting the requirements for detailed classification in specific application scenarios. Traditional methods are limited to known categories, while zero-shot FGI recognition allows the model to recognize fine-grained categories that have not been previously seen, thereby expanding the recognition range.

In order to improve the performance of zero shot FGI recognition, this paper adopts CM features to fuse image features with text descriptions. FGI recognition can be achieved by introducing CM semantic enhancement to generate adversarial networks.

Traditional ZSL refers to supervised learning tasks where the model is not exposed to certain categories of sample data during the training phase, and needs to classify or recognize these unseen categories during the testing phase.

The category accuracy with the highest confidence in the measurement average model output in the test set:

$$A = \frac{1}{\text{count } M} \sum_{m \in M} acc_m \quad (7)$$

In formula 7, acc_m represents the accuracy of the category with the highest confidence.

Generalized ZSL refers to supervised learning, where the model is not exposed to certain categories of sample data during the training phase. During the testing phase, it is necessary to classify or recognize these unseen categories, which can also be applied to other learning paradigms.

The accuracy of visible categories on the test set is expressed as:

$$B = \frac{1}{\text{count } H} \sum_{h \in H} acc_h \quad (8)$$

Based on a comprehensive evaluation of A and B, it is concluded that:

$$C = \frac{2 * A * B}{A + B} \quad (9)$$

The evaluation method of CM retrieval can be used to evaluate the performance of the model in terms of CM embedding. The accuracy of the model for IR is expressed as:

$$P = \frac{TP}{TP + FP} \quad (10)$$

The recall rate of IR is expressed as:

$$R = \frac{TP}{TP + FN} \quad (11)$$

To comprehensively analyze the performance of CM semantic enhancement GAN, multiple embedded and generative methods are set up for comparison.

The embedded methods set include DAP (Discriminative Attribute Prediction), CMT (Continuous Multimodal Transduction), SSE (Semantic Space Embedding), ALE (Attribute Label Embedding), EZSL (End-to-End Zero-shot Learning), SAE (Semantic Autoencoder), DEM (Discriminative Embedding Model).

The set generative methods include f-CLSWGAN (f-Conditional Least Squares Generative Adversarial), cycle-WGAN (Cycle-Consistent Wasserstein GAN), SE-GZSL (Semantic Embedding for Generalized Zero-Shot Learning), LisGAN (Language-based Image Synthesis GAN), SABR (Semantic-Aware Background Removal), f-VAEGAN (f-VAE GAN).

4 RESULTS AND DISCUSSION

4.1 Traditional ZSL

In traditional ZSL, the comparison results between the paper's model and embedded methods are shown in Figure 5.

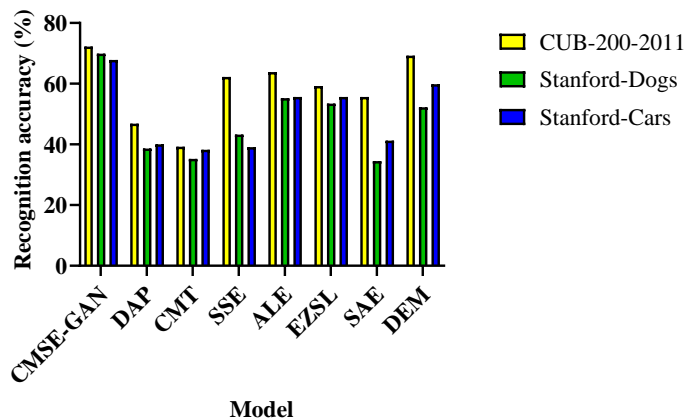


Figure 5: Comparison results with embedded methods (traditional ZSL).

CMSE-GAN adopts a CM semantic enhancement method, which can effectively fuse information from different modalities, providing a more comprehensive and rich perspective. In contrast, traditional embedded methods may not fully utilize the correlation information between different modalities. The results on three datasets indicate that CMSE-GAN has higher recognition accuracy for FGI recognition using traditional ZSL. The recognition accuracy of CMSE-GAN on CUB-200-2011, Stanford-Dogs, and Stanford-Cars is 72.2%, 69.9%, and 67.8%, respectively. CMSE-GAN can effectively improve its generalization ability for unseen categories during the learning process. By generating images corresponding to text descriptions, CMSE-GAN can provide more diverse training samples for new categories, thereby improving the recognition ability of unknown categories.

In traditional ZSL, the comparison between the paper's model and the generative method is shown in Figure 6.

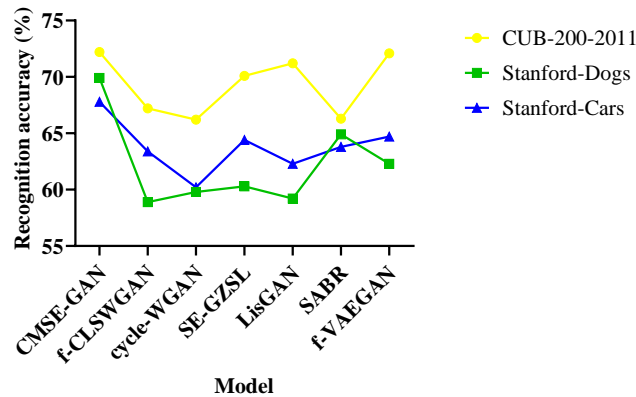


Figure 6: Comparison with generative methods (traditional ZSL).

The results in Figure 6 show that CMSE-GAN has higher recognition accuracy. CMSE-GAN adopts a CM semantic enhancement method, which can effectively fuse image and text description information, thereby more accurately capturing the semantic information of images.

4.2 Generalized ZSL

Generalized ZSL extends the scope of traditional ZSL. In generalized ZSL, the model can be exposed to not only some categories of sample data during the training phase but also some auxiliary information, such as attribute vectors or semantic embeddings. The comparison results between CMSE-GAN and embedded methods in generalized ZSL are shown in Table 4.

Model	CUB-200-2011			Stanford-Dogs			Stanford-Cars		
	A (%)	B (%)	C (%)	A (%)	B (%)	C (%)	A (%)	B (%)	C (%)
CMSE-GAN	54.2	80.4	64.8	56.8	57.8	57.3	44.9	43.3	44.1
DAP	0.0	86.6	0.0	1.5	66.8	2.9	4.8	24.2	8.0
CMT	0.4	85.6	0.8	2.4	67.9	4.6	6.6	40.2	11.3
SSE	10.2	82.2	18.1	3.8	72.3	7.2	16.8	30.2	21.6
ALE	13.6	84.5	23.4	23.4	64.3	34.3	22.2	34.6	27.0
EZSL	6.5	78.9	12.0	11.2	64.5	19.1	15.6	33.8	21.3
SAE	10.9	76.5	19.1	12.4	72.3	21.2	8.6	44.2	14.4
DEM	32.1	78.6	45.6	11.1	75.3	19.3	22.2	41.2	28.9

Table 4: Comparison results with embedded methods (generalized ZSL).

This article uses the indicators in formulas 7, 8, and 9 to evaluate and CMSE-GAN still has higher IR performance in the three datasets. CMSE-GAN can effectively integrate information from different modalities, improving its generalization ability for unseen categories by comprehensively integrating image and text descriptions. The comparison results between CMSE-GAN and generative methods are shown in Table 5.

Model	CUB-200-2011			Stanford-Dogs			Stanford-Cars		
	A (%)	B (%)	C (%)	A (%)	B (%)	C (%)	A (%)	B (%)	C (%)
CMSE-GAN	66.5	78.3	71.9	56.8	66.8	61.4	62.2	52.6	57.0

<i>f</i> - CLSWGAN	60.2	64.2	62.1	43.8	58.2	50.0	44.6	48.6	46.5
<i>cycle</i> - WGAN	34.6	56.8	43.0	52.4	63.2	57.3	42.4	38.9	40.6
SE-GZSL	48.9	53.1	50.9	42.2	54.4	47.5	45.2	50.8	47.8
LisGAN	46.8	52.9	49.7	54.6	64.2	59.0	46.2	44.1	45.1
SABR	59.2	54.3	56.6	53.5	63.3	58.0	50.1	42.1	45.8
<i>f</i> -VAEGAN	52.6	77.8	62.8	52.1	58.9	55.3	43.8	43.9	43.8

Table 5: Comparison results with generative methods (generalized ZSL).

CMSE-GAN can effectively integrate multimodal features of image and text descriptions, fully utilizing two different forms of information. Through this fusion, the model can have a more comprehensive understanding of the target object and improve its recognition ability for unseen categories.

4.3 CM Retrieval

The performance of CMSE-GAN in CM retrieval is shown in Figure 7.

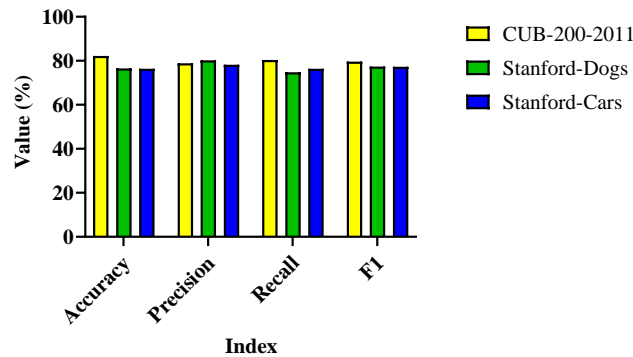


Figure 7: Performance of CM retrieval.

CMSE-GAN can effectively integrate information from different modalities, thereby enabling a more comprehensive understanding and representation of data. It provides richer and more comprehensive descriptions by combining multiple forms of information. In CM retrieval, the accuracy of CMSE-GAN on CUB-200-2011, Stanford-Dogs, and Stanford-Cars datasets was 82.2%, 76.5%, and 76.4%, respectively.

5 CONCLUSIONS

This paper cited a CM semantically augmented generative adversarial network for a fine-grained IR task. The ZSL capability for new categories is enhanced by effectively fusing image and text description information. Features with visual realism and rich semantic information are produced by adversarial training, thus improving fine-grained IR. Through experimental evaluation on the CUB-200-2011, Stanford-Dogs, and Stanford-Cars datasets, the results demonstrated that CMSE-GAN can effectively enhance fine-grained IR and is able to maintain high performance under both traditional ZSL and generalized ZSL. In some complex scenarios, the fusion of CM information may bring in a certain amount of noise and uncertainty, resulting in a degradation of model performance. The method of CM information fusion is further refined to improve the robustness of the model to noise and uncertainty in order to meet the challenges in complex scenarios.

Jingyuan He, <https://orcid.org/0009-0002-6192-0460>
 Bailong Yang, <https://orcid.org/0009-0005-3772-2634>
 Yuan Tian, <https://orcid.org/0009-0009-9834-4533>

REFERENCES

- [1] Ma, C.; Li, G.; Chen, S.; Mao, J.; Zhang, J.: Research on the usefulness recognition of travel online reviews based on the semantic fusion of multimodal data, *Journal of Intelligence*, 39(2), 2020, 199-207.
- [2] Liao, M.; Chen, L.; Wang, G.; Peng, S.: The intelligent identification and effectiveness of children with autism spectrum disorder by integrating multimodal data, *Scientific Bulletin*, 66(20), 2021, 2618-2628.
- [3] Wei, X.-S.; Song, Y.-Z.; Aodha, O. M.; Wu, J.: Fine-grained image analysis with deep learning: A survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12), 2021, 8927-8948, <https://doi.org/10.1109/TPAMI.2021.3126648>.
- [4] Fayou, S.; Ngo, H.; Sek, Y.: Combining multi-feature regions for fine-grained image recognition, *Int. J. Image Graph. Signal Process*, 14(1), 2022, 15-25, <https://doi.org/10.5815/ijigsp.2022.01.02>.
- [5] Koniusz, P.; Hongguang, Z.: Power normalizations in fine-grained image, few-shot image and graph classification, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2), 2021, 591-609, <https://doi.org/10.1109/TPAMI.2021.3107164>.
- [6] Chen, L.; Chao, X.; Pan, L.; Cao, J.; Zhang, R.: Pay attention to DenseNet's fine-grained model identification based on components, *Journal of Intelligent Systems*, 17(2), 2021, 402-410.
- [7] Jiang, L.; Liu, J.: Fine-grained image recognition using multi-component learning, *Journal of Computer-Aided Design and Graphics*, 35(7), 2023, 1032-1039.
- [8] Zhao, P.; Wang, C.; Zhang, S.; Liu, Z.: A zero-sample image classification method based on subspace learning of fusion reconstruction, *Journal of Computer Science*, 44(2), 2021, 409-421.
- [9] Dai, T.; Xie, Q.; Huang, J.; Sun, C.; Cong, S.: A weakly supervised fine-grained deep network method for wood classification, *Journal of Southwest University (Natural Science Edition)*, 44(10), 2022, 161-172.
- [10] Zhai, Y.; Zhang, Z.; Wang, Y.: A zero-sample image recognition method based on improved TransGAN, *Journal of Intelligent Systems*, 18(2), 2022, 352-359.
- [11] Yu, J.; Min, T.; Hongyuan, Z.; Yong, R.; Dacheng, T.: Hierarchical deep click feature prediction for fine-grained image recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2), 2019, 563-578, <https://doi.org/10.1109/TPAMI.2019.2932058>.
- [12] Li, X.; Liyun, Y.; Dongliang, C.; Zhanyu, M.; Jie, C.: Dual cross-entropy loss for small-sample fine-grained vehicle classification, *IEEE Transactions on Vehicular Technology*, 68(5), 2019, 4204-4212, <https://doi.org/10.1109/TVT.2019.2895651>.
- [13] Rodriguez, P.; Diego, V.; Guillem, C.; Josep, M. G.: Pay attention to the activations: A modular attention mechanism for fine-grained image recognition, *IEEE Transactions on Multimedia*, 22(2), 2019, 502-514, <https://doi.org/10.1109/TMM.2019.2928494>.
- [14] Sumbul, G.; Ramazan, G. C.; Selim, A.: Multisource region attention network for fine-grained object recognition in remote sensing imagery, *IEEE Transactions on Geoscience and Remote Sensing*, 57(7), 2019, 4929-4937.
- [15] Kuznetsova, A.; Hassan, R.; Neil, A.; Jasper, U.; Ivan, K.; Jordi, P.-T.; et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale, *International Journal of Computer Vision*, 128(7), 2020, 1956-1981.
- [16] Pan, J.; He, Z.; Li, Z.; Liang, Y.; Qiu, L.: Review of research on multimodal emotion recognition, *Journal of Intelligent Systems*, 15(4), 2020, 633-645.

- [17] Yang, X.; Lin, S.; Lu, X.; Wang, L.; Li, D.; Wang, B.; et al.: Multi-modal image fusion based on generative confrontation network, *Progress in Lasers and Optoelectronics*, 56(16), 2019, 40-49.
- [18] Liu, Z.; Duan, Q.; Shi, S.; Zhao, P.: RGB-D image significance detection based on multi-modal feature fusion supervision, *Journal of Electronics and Informatics*, 42(4), 2020, 997-1004.
- [19] Cao, X.; Zhang, Y.; Pan, M.; Zhu, S.; Yan, H.: Research on the recognition method of learning participation in the perspective of artificial intelligence--A deep learning experimental analysis based on multi-modal data fusion, *Distance Education Magazine*, 37(1), 2019, 32-44.
- [20] Abdullah, S. M.; Saleem, A.; Siddeeq, Y. A. A.; Mohammed, A. M. S.; Subhi, Z.: Multimodal emotion recognition using deep learning, *Journal of Applied Science and Technology Trends*, 2(02), 2021, 52-58, <https://doi.org/10.38094/jastt20291>
- [21] Wu, H.; Xin, M.; Yibin, L.: Spatiotemporal multimodal learning with 3D CNNs for video action recognition, *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3), 2021, 1250-1261, <https://doi.org/10.1109/TCSVT.2021.3077512>.
- [22] Li, G.; Lijun, G.; Rong, Z.; Jiangbo, Q.; Shangce, G.: TransGait: Multimodal-based gait recognition with set transformer, *Applied Intelligence*, 53(2), 2023, 1535-1547.
- [23] Daas, S.; Amira, Y.; Toufik, B.; Mouna, S.; Mohamed, B.; El-Bay, B.; et al.: Multimodal biometric recognition systems using deep learning based on the finger vein and finger knuckle print fusion, *IET Image Processing*, 14(15), 2020, 3859-3868, <https://doi.org/10.1049/iet-ipr.2020.0491>.
- [24] Wang, M.; Wu, Z.; Zhou, Z.: Research on the fine-grained identification of crop pests and diseases based on attention-improving CBAM, *Journal of Agricultural Machinery*, 52(4), 2021, 239-247.
- [25] Mao, L.; Xue, Y.; Wei, Y.; Zhu, T.: A glasses removal method for fine-grained face recognition, *Journal of Electronics and Informatics*, 43(5), 2021, 1448-1456.
- [26] Hu, X.; Peng, T.: Fine-grained image classification based on data-enhanced visual transformer, *Journal of Xihua University (Natural Science Edition)*, 41(6), 2022, 9-16.
- [27] Li, M.; He, L.; Lei, C.; Gong, Y.: SqueezeNet fine-grained image classification model based on attention feature fusion, *Journal of Yunnan University (Natural Science Edition)*, 43(5), 2021, 868-876.
- [28] Park, C.; Namme, M.: Dog-Species Classification through CycleGAN and Standard Data Augmentation, *Journal of Information Processing Systems*, 19(1), 2023, 67-79, <https://doi.org/10.3745/JIPS.02.0190>.
- [29] Wang, L.; Kai, H.; Xu, F.; Xitao, M.: Multilayer feature fusion with parallel convolutional block for fine-grained image classification, *Applied Intelligence*, 52(3), 2022, 2872-2883.
- [30] Shi, F.: The Chinese text corpus preprocessing module based on jieba Chinese word segmentation is implemented, *Computer Knowledge and Technology: Academic Edition*, 16(14), 2020, 248-251.
- [31] Tian, J.; Song, H.; Chen, L.; Sheng, G.; Jiang, X.: A method for identifying text entities of equipment failures for knowledge graph construction, *Power Grid Technology*, 46(10), 2022, 3913-3922.
- [32] Gui, J.; Zhenan, S.; Yonggang, W.; Dacheng, T.; Jieping, Y.: A review on generative adversarial networks: Algorithms, theory, and Applications, *IEEE Transactions on Knowledge and Data Engineering*, 35(4), 2021, 3313-3332, <https://doi.org/10.1109/TKDE.2021.3130191>.
- [33] Jabbar, A.; Xi, Li.; Bourahla, O.: A survey on generative adversarial networks: Variants, applications, and training, *ACM Computing Surveys (CSUR)*, 54(8), 2021, 1-49, <https://doi.org/10.1145/3463475>.
- [34] Saxena, D.; Jiannong, C.: Generative adversarial networks (GANs) challenges, solutions, and future directions, *ACM Computing Surveys (CSUR)* 54(3), 2021, 1-42, <https://doi.org/10.1145/3446374>.

- [35] Maeda, H.; Takehiro, K.; Yoshihide, S.; Toshikazu, S.; Hiroshi, O.: Generative adversarial network for road damage detection, *Computer-Aided Civil and Infrastructure Engineering* 36(1), 2021, 47-60, <https://doi.org/10.1111/mice.12561>.