GCN-BASED MULTI-MODAL MULTI-LABEL ATTRIBUTE CLASSIFICATION IN ANIME ILLUSTRATION USING DOMAIN-SPECIFIC SEMANTIC FEATURES

Ziwen Lan[†]*, Keisuke Maeda*[††]*, Takahiro Ogawa*[††] *and Miki Haseyama*[††]

[†] Graduate School of Information Science and Technology, Hokkaido University
[††] Faculty of Information Science and Technology, Hokkaido University
E-mail: {lan, maeda, ogawa}@lmd.ist.hokudai.ac.jp, miki@ist.hokudai.ac.jp

2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9898071

ABSTRACT

This paper presents a multi-modal multi-label attribute classification model in anime illustration based on Graph Convolutional Networks (GCN) using domain-specific semantic features. In animation production, since creators often intentionally highlight the subtle characteristics of the characters and objects when creating anime illustrations, we focus on the task of multi-label attribute classification. To capture the relationship between attributes, we construct a multi-modal GCN model that can adopt semantic features specific to anime illustration. To generate the domain-specific semantic features that represent the semantic contents of anime illustrations, we construct a new captioning framework for anime illustration by combining real images and their style transformation. The contributions of the proposed method are two-folds. 1) More comprehensive relationships between attributes are captured by introducing GCN with semantic features into the multi-label attribute classification task of anime illustrations. 2) More accurate image captioning of anime illustrations can be generated by a trainable model by using only real-world images. To our best knowledge, this is the first work dealing with multi-label attribute classification in anime illustration. The experimental results show the effectiveness of the proposed method by comparing it with some existing methods including the state-of-the-art methods.

***Index Terms*—** Anime illustration, graph convolutional networks, semantic feature, multi-modal classification, image captioning.

1. INTRODUCTION

With the development of the animation industry in recent years, several studies related to anime illustration such as illustration editing [1,2], line-art colorization [3,4] and cartoon face generation [5,6] have been conducted. Due to the increasing number of anime illustrations, there is a growing need for techniques to classify them, and these techniques have a potential to enhance the above various studies. In order to implement the classification techniques for anime illustration, it is necessary to know the contents of the anime illustrations.

It is well known that image classification is a fundamental task in the computer vision community and plays an important role in a wide range of applications. Since images contain multiple objects in the great majority of cases, multi-label image classification has aroused extensive attention in recent years. In real-life problems such as medical image classification [7] and recommendation systems [8], multi-label image classification can help us get better solutions. Therefore, it can be expected that the multi-label classification technique is also effective for anime illustrations.

Many multi-label classification methods for anime illustration have been proposed [9, 10]. For example, a method [9] based on convolutional neural network (CNN) achieved good classification results. Since objects always co-occur in anime illustrations, it is necessary to take the relationships between labels into consideration so that we can improve the accuracy of classification. In recent years, great progress has been made in the research of graph convolutional network (GCN) [11]. GCN shows a great ability for modeling the relationships between the labels on a graph structure. For example, Feng et al. proposed a GCN-based multi-label anime illustration classification method [10], which modeled the correlation between different labels by constructing a complete graph.

However, the classification task of anime illustration is slightly different from that of real-world image. Anime illustrations are artificially created images, and the creators of anime illustrations often intentionally highlight the fine characteristics of the characters and objects which are called attributes. For this reason, to construct the classification method for anime illustrations, we need to consider not only simple objects in the image but also their attributes. Therefore, it is necessary to solve the multi-label attribute classification task [12], and there is no work focusing on this task for anime illustration.

In generic image classification, many methods for multi-label attribute classification have been proposed [13–15], and they focused on the capture of relationship between attributes from the visual information of the illustration. As mentioned above, artificially created anime illustrations contain more detailed information than real-world images, so it is necessary to take more semantic information into account. However, it is difficult to grasp the deeper and more detailed connections between attributes only from visual information. Therefore, to grasp semantic features specific to anime illustration, we introduce the image captioning [16] into the the classification task of anime illustration. The image captioning is a more detailed factor consisted of the semantic information, and is expected to be useful for representing the attributes. However, the available models [16,17] of image captioning are learned from only real-world images. There are several differences between anime illustrations and real-world images which are called domain-shift, and the domain-shift will lead to a decrease in the accuracy of image captioning results. In other words, it is not feasible to apply this technique for real-world images directly to anime illustrations.

This work was partly supported by JSPS KAKENHI Grant Number JP21H03456.

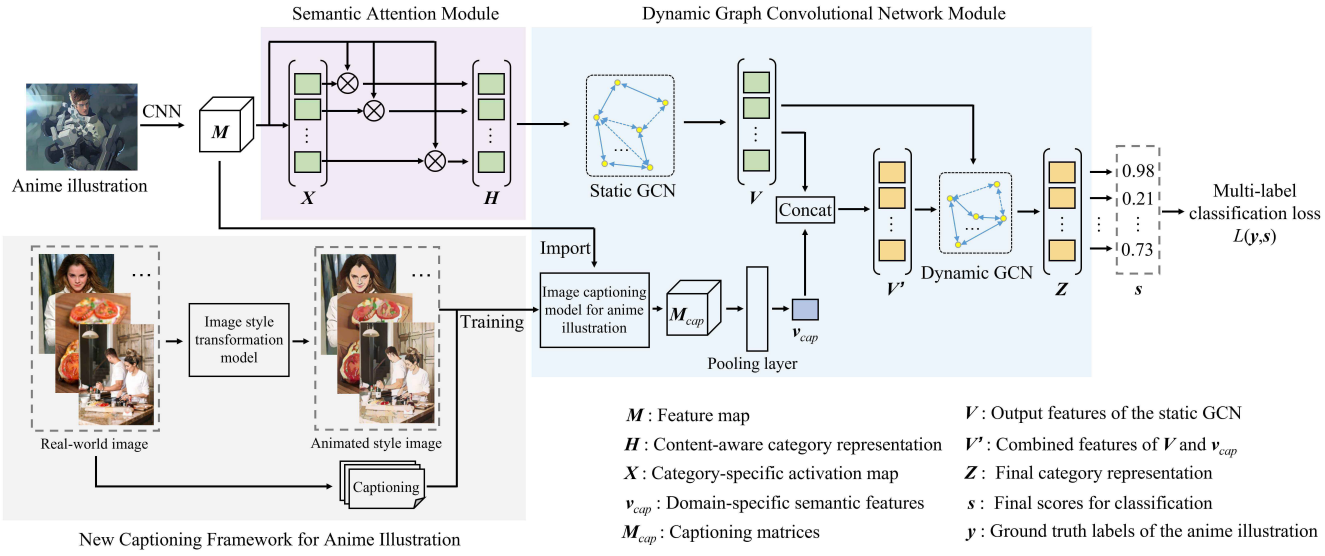

Fig. 1. Overview of the GCN-based multi-modal multi-label attribute classification model for anime illustrations using domain-specific semantic features. We train a captioning model for anime illustration by using the captioning of real-world images and corresponding animated style images generated by the image style transformation model. We employ this captioning model to obtain the semantic feature $\boldsymbol{v}_{cap}$ of anime illustration. The semantic feature $\boldsymbol{v}_{cap}$ is used in the GCN-based multi-label classification model to improve the classification accuracy.

In this paper, we propose a multi-modal multi-label attribute classification model for anime illustrations based on GCN using domain-specific semantic features. This is the first attempt to introduce the semantic features to the task of multi-label attribute classification in anime illustrations. The technical contributions of this paper are as follows.

Contribution (i): To deal with the task, multi-label attribute classification in anime illustration, we construct a GCN-based model using not only visual features from images but also semantic features suitable for anime illustration.

Contribution (ii): We propose a trainable model for the image captioning of anime illustrations by using only real-world images to extract highly expressive image captioning of anime illustrations.

In contribution (i), we can capture the relationship between different attributes in the images by introducing semantic features for anime illustration based on image captioning into GCN, and we further improve the classification accuracy. In contribution (ii), we convert real-world images into animated style images and import them in pairs into the image captioning model pre-trained with real-world images, and train the model to make each feature space closer by considering the differences between the two captioning results. In this way, we can generate captioning of anime illustrations by using only real-world images. These two contributions show that successful multi-label attribute classification for anime illustration becomes feasible.

2. OUR MULTI-LABEL ATTRIBUTE CLASSIFICATION MODEL

In this section, we show the details of the proposed model. First, we explain how we construct a new captioning framework for anime illustrations (2.1). Then we describe the general flow of the GCN-based multi-modal multi-label attribute classification using domain-specific semantic features (2.2). Finally, we explain the final classification and the loss used for the training of the proposed model (2.3). Figure 1 shows the overview of the proposed model.

2.1. New Captioning Framework for Anime Illustration

This subsection shows how we construct the new captioning framework for anime illustrations by combining real images and their corresponding art style transformations to generate features that represent the semantic contents of anime illustrations.

First, we put real-world images into the image style transformation model and convert them into animated style images. In our method, we use white-box cartoonization model [18] as the image style transformation model. The white-box cartoonization model can generate high-quality cartoonized images from real-world images, and with this model, we can get pairs of real-world images and animated style images.

Next, we input real-world images into a pre-trained captioning model [17] to obtain the corresponding captioning results. Under normal circumstances, since simply converting a real image into an animated style image does not change rough contents of the image, we can regard the captioning of the real-world image as the ground truth of that of the corresponding animated style image. We use the animated style image and the captioning of the corresponding real-world image to train the captioning model so that the distance between the captioning result of the animated style image and that of the real-world image becomes close. The training process will be mentioned in Section 3. In this way, the problem of the domain-shift can be solved, and we can obtain domain-specific semantic features for anime illustrations.

2.2. GCN-based Multi-modal Multi-label Attribute Classification Using Domain-specific Semantic Features

This subsection describes the general flow of the GCN-based multi-label attribute classification using domain-specific semantic features. Inspired by the uni-modal version of previous GCN [19], our multi-label attribute classification model mainly consists of the following two parts: Semantic Attention Module (SAM) and Dynamic Graph Convolutional Network Module (DGCNM) as shown in Fig. 1.

The SAM aims to estimate the content-aware category representations after extracting feature maps from the images. For an input image, we first employ a CNN backbone to extract the convolutional feature maps $\boldsymbol{M} \in \mathbb{R}^{H\times W\times D}$. H means the height of $\boldsymbol{M}$, W means the width of $\boldsymbol{M}$, and D means the number of channels of $\boldsymbol{M}$. From the feature maps $\boldsymbol{M}$, a set of category-specific activation maps $\boldsymbol{X} \in \mathbb{R}^{H\times W\times C}$ are generated by Class Activation Mapping [20]. Then we use the activation maps $\boldsymbol{X}$ to convert $\boldsymbol{M}$ into a content-aware category representation $\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_C]^\top \in \mathbb{R}^{C\times D}$ as follows:

$$\boldsymbol{h}_c = \sum_{i=1}^{H}\sum_{j=1}^{W} x_{i,j}^{c}\boldsymbol{m}_{i,j}, \tag{1}$$

where C is the number of categories, $x_{i,j}^{c}$ is the (i,j)-th weight of c-th activation map $\boldsymbol{X}_c \in \mathbb{R}^{H\times W}$, and $\boldsymbol{m}_{i,j} \in \mathbb{R}^{D}$ is the (i,j)-th feature vector of the feature map $\boldsymbol{M}$.

The DGCNM aims to obtain the final category representation by GCN using the contents in each specific input image. We take the content-aware category representation $\boldsymbol{H}$ as the input node features, and feed it into a static GCN. We define the single-layer static GCN as $\boldsymbol{V} = LReLU(\boldsymbol{A}_s\boldsymbol{H}\boldsymbol{W}_s) \in \mathbb{R}^{C\times D}$. $LReLU(\cdot)$ is the activation function LeakyReLU [21]. $\boldsymbol{A}_s$ and $\boldsymbol{W}_s$ are respectively the correlation matrix and the state update weights. During the training process, $\boldsymbol{A}_s$ and $\boldsymbol{W}_s$ are randomly initialized by the gradient descent.

To introduce the semantic features into the GCN model, we employ the image captioning model that is trained for anime illustration in the previous subsection to obtain the captioning matrices $\boldsymbol{M}_{cap} \in \mathbb{R}^{H\times W\times D}$. Then we transform the captioning matrices into semantic features $\boldsymbol{v}_{cap} \in \mathbb{R}^{D}$ that can be imported into the multi-label classification model by a pooling layer. Note that $\boldsymbol{v}_{cap}$ is defined as follows:

$$\boldsymbol{v}_{cap} = p(\boldsymbol{M}_{cap}), \quad \text{s.t. } \boldsymbol{M}_{cap} = cap(\boldsymbol{M}), \tag{2}$$

where $p(\cdot)$ means the pooling layer, and $cap(\cdot)$ means the captioning model mentioned in subsection 2.1. After that, we integrate $\boldsymbol{v}_{cap}$ and the output features $\boldsymbol{V}$ of the static GCN. Specifically, we simply concatenate $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_C]^\top \in \mathbb{R}^{C\times D}$ with $\boldsymbol{V}_{cap} = [\boldsymbol{v}_{cap}, \boldsymbol{v}_{cap}, \ldots, \boldsymbol{v}_{cap}]^\top \in \mathbb{R}^{C\times D}$ to get the combined feature $\boldsymbol{V}' = [\boldsymbol{V}, \boldsymbol{V}_{cap}] \in \mathbb{R}^{C\times 2D}$. Then we import $\boldsymbol{V}'$ into the dynamic GCN. We define the output $\boldsymbol{Z} \in \mathbb{R}^{C\times D}$ of the dynamic GCN as follows:

$$\boldsymbol{Z} = LReLU(\boldsymbol{A}_d\boldsymbol{V}\boldsymbol{W}_d), \text{ where } \boldsymbol{A}_d = \sigma(\boldsymbol{W}_A\boldsymbol{V}'). \tag{3}$$

$\boldsymbol{W}_A$ and $\boldsymbol{W}_d$ mean the the weights of a $conv$ layer that formulates the dynamic correlation matrix $\boldsymbol{A}_d$ and the state-update weights, respectively. $\sigma(\cdot)$ represents the sigmoid activation function. As a result, we can capture the relationship between different attributes in the images by GCN and import the information of textual descriptions of the anime illustration into the GCN to further improve the classification accuracy.

2.3. Final Classification and Loss

This subsection explains the final classification and loss. For the final classification, we use the output $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_C]^\top$ of the dynamic GCN and predict scores of each category. Particularly, we put each vector $\boldsymbol{z}_c$ of the final category representation $\boldsymbol{Z}$ into a binary classifier and get the predict scores s^c of category c. We concatenate the scores s^c to obtain the final scores $\boldsymbol{s} = [s^1, s^2, \ldots, s^C]^\top$. According to previous studies [19, 22, 23], the loss function $L(\boldsymbol{y}, \boldsymbol{s})$ is calculated as follows:

$$L(\boldsymbol{y}, \boldsymbol{s}) = \sum_{c=1}^{C} y^c \log(\sigma(s^c)) + (1-y^c)\log(1-\sigma(s^c)), \tag{4}$$

where $\boldsymbol{y} \in \mathbb{R}^C$ means the ground truth labels of an image, and $y^c = \{0, 1\}$ indicates the presence or absence of label c in the image.

3. EXPERIMENTS

This section shows the details of our experiments. First, we explain the training of our captioning framework (3.1). Then we show the details of the training of the whole multi-label classification model (3.2). Next, we introduce the comparison methods and the evaluation metrics used in this experiment (3.3) and finally show the results of the experiment (3.4).

3.1. Training of Captioning Model for Anime Illustrations

In order to perform the transfer learning of the captioning model, we first need to produce a dataset of animated style images from real-world images. We used 8,000 real-world images from Flickr8k dataset [24] and imported them into the image style transformation model [18] shown in subsection 2.1 to generate their corresponding animated style images. Then we used the animated style images as input and the captioning of the corresponding real-world images as ground truth to train the captioning model [17].

3.2. Training of Whole Multi-label Classification Model

In our experiments, we used Danbooru2020 [9] dataset for training the whole multi-label attribute classification model. Danbooru2020 dataset is a large anime illustration dataset with over 4.2 million images and over 130 million tags. From the dataset, we extracted about 25,000 anime illustrations, which include 100 common attribute classes, and each illustration contains an average of 6.3 attribute labels. We used 75% of the 25,000 images as the training set and the remaining 25% as the validation set.

We employed ResNet-101 [25] as the backbone of the GCN-based attribute classification model. The negative slope of LeakyReLU utilized in the DGCNM was set to 0.2. The input images were randomly cropped, resized to 448×448 pixels and flipped horizontally for data augmentation. We chose Stochastic Gradient Descent as our optimizer. Its momentum and weight decay were respectively set to 0.9 and 1.0×10^{-4} . We set the initial learning rate to 0.5 for SAM/DGCNM and 0.05 for the backbone CNN.

Table 1. Performance comparison between our model and other image classification models. We mark the best results as bold.

Method	OP	OR	OF1	CP	CR	CF1	mAP
ResNet-101 [25]	60.8	55.4	58.3	62.0	58.0	61.3	63.4
SSGRL [26]	64.0	56.6	61.1	71.2	57.2	60.2	69.4
DAN [12]	64.7	51.8	57.5	67.1	56.6	61.3	64.3
ML-GCN [23]	61.2	53.6	59.1	68.8	61.4	65.9	66.3
ADD-GCN [19]	63.4	**59.2**	60.3	71.6	63.1	68.4	70.1
P-GCN [22]	65.3	57.8	60.7	**73.9**	58.5	69.8	70.4
Ours	**67.6**	58.1	**61.4**	73.0	**63.8**	**71.0**	**71.2**

Table 2. The ablation experimental results. We can see the influence of the accuracy of the captioning model (BLEU) on the final classification (mAP). In the ablation experiment, we randomly selected 100 anime illustrations with ground truth captionings from Danbooru2020 dataset and calculated BLEU score and mAP.

Method	BLEU [27]	mAP
ADD-GCN + captioning (by model trained for real image)	44.5	64.1
Ours (using captioning obtained by model trained for anime illustration)	**54.2**	**67.4**

3.3. Comparison Methods and Evaluation Metrics

To verify the effectiveness of the proposed method, we employed a series of comparison methods as follows: ResNet-101 [25], Semantic Specific Graph Representation Learning (SSGRL) framework [26], Deep Attribute Network (DAN) [12], Multi-Label Image Recognition with GCN (ML-GCN) [23], Attention-Driven Dynamic GCN (ADD-GCN) [19] and Prediction Learning GCN (P-GCN) [22]. Comparison method ResNet-101 is the backbone of our method. SSGRL and DAN are the classification models based on CNN, and ML-GCN, ADD-GCN and P-GCN are the classification models based on GCN. Among these comparison methods, ADD-GCN and P-GCN are the state-of-the-art methods. Each of these models was trained with the same hyper-parameters as ours. To confirm the performance improvement after employing the image captioning in multi-label classification model, all the comparison methods were trained with visual features of the anime illustrations only.

To compare our method with other methods, according to previous studies [19, 22, 23], we adopted the average of overall precision, recall, F1 score (expressed respectively as OP, OR and OF1) and the average of per-class precision, recall, F1 score (expressed respectively as CP, CR and CF1) to evaluate the performance of the methods. When measuring the precision, recall and F1 score, for each image, the label c is predicted as positive if the score s^c calculated in subsection 2.4 is greater than 0.5. Also, we adopted the average precision (AP) and the mean average precision (mAP) that were often used in multi-label classification tasks [19].

Moreover, we conducted an ablation experiment. Specifically, we compared our proposed model with the image captioning model trained by real-world images. To investigate the contribution of using image captioning, we evaluated the effectiveness of the captioning model by BLEU score [27] that is commonly used in the image captioning task. In general, higher BLEU score means higher accuracy of the captioning results. To calculate BLEU score, we randomly chose 100 anime illustrations from Danbooru2020 dataset and respectively imported them into the captioning model trained by real-world images and that trained by the method in subsection 2.1.

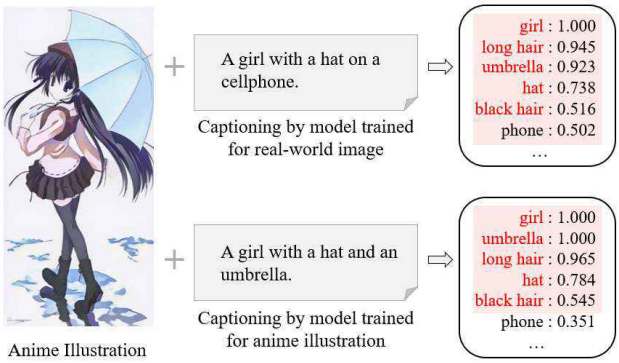

Fig. 2. An example of the ablation experiment. All the scores s are sorted in descending order. We mark the ground truth labels in red font, and the final classification results of each method in the red box.

3.4. Experimental Results

We show the comparison between the proposed method and other image classification methods in Table 1. By comparing with the baseline networks like ResNet-101 [25] which constructed the same latent space as ours, we can see the improvement of the performance after the employment of GCN by the experimental results. Moreover, by comparing with the state-of-the-art methods based on GCN like ADD-GCN [19], we also confirm that introducing the domain-specific semantic features into the GCN model is effective.

We also show the ablation experimental results in Table 2. From this table, it can be confirmed that the higher the accuracy of the image captioning is, the higher the final classification performance is. Also, Fig. 2 shows the example of our method and the comparison method in the ablation experiments. As shown in this figure, the label *phone*, which is incorrectly classified as positive by the method using the captioning model trained for real-world image, is correctly classified as negative after using the captioning model trained for anime illustration. Therefore, the effectiveness of our method is verified.

4. CONCLUSIONS

In this paper, we have proposed a GCN-based multi-modal multi-label attribute classification model in anime illustration using domain-specific semantic features. We capture the relationship between different attributes by introducing semantic features specific to anime illustration into GCN, and we further improve the classification accuracy. In addition, to generate more accurate image captioning for constructing our GCN model, we propose a trainable framework for the image captioning of anime illustration. The experiments show the effectiveness and the rationality of the proposed method.

5. REFERENCES

[1] Qingyuan Zheng, Zhuoru Li, and Adam Bargteil, "Learning to shadow hand-drawn sketches," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7436–7445.

[2] Keita Awane, Koki Tsubota, Hikaru Ikuta, Yusuke Matsui, Kiyoharu Aizawa, and Naohiro Yanase, "Improving the quality of illustrations: Transforming amateur illustrations to a professional standard," in *Proc. IEEE International Conference on Image Processing*, 2021, pp. 584–588.

[3] Felipe Coelho Silva, Paulo André Lima de Castro, Hélio Ricardo Júnior, and Ernesto Cordeiro Marujo, "Mangan: Assisting colorization of manga characters concept art using conditional gan," in *Proc. IEEE International Conference on Image Processing*, 2019, pp. 3257–3261.

[4] Tzu-Ting Fang, Duc Minh Vo, Akihiro Sugimoto, and Shang-Hong Lai, "Stylized-colorization for line arts," in *Proc. IEEE International Conference on Pattern Recognition*, 2021, pp. 2033–2040.

[5] Yanghua Jin, Jiakai Zhang, Minjun Li, Yingtao Tian, Huachun Zhu, and Zhihao Fang, "Towards the high-quality anime characters generation with generative adversarial networks," in *Proc. NIPS Workshop on Machine Learning for Creativity and Design*, 2017.

[6] Bing Li, Yuanlue Zhu, Yitong Wang, Chia-Wen Lin, Bernard Ghanem, and Linlin Shen, "Anigan: Style-guided generative adversarial networks for unsupervised anime face generation," *arXiv preprint arXiv:2102.12593*, 2021.

[7] Zongyuan Ge, Dwarikanath Mahapatra, Suman Sedai, Rahil Garnavi, and Rajib Chakravorty, "Chest X-rays classification: A multi-label and fine-grained problem," *arXiv preprint arXiv:1807.07247*, 2018.

[8] Himanshu Jain, Yashoteja Prabhu, and Manik Varma, "Extreme multi-label loss functions for recommendation, tagging, ranking and other missing label applications," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 935–944.

[9] The Danbooru Community and Gwern Branwen, "Danbooru2020: A large-scale crowdsourced and tagged anime illustration dataset," `https://www.gwern.net/Danbooru2020`, 2021.

[10] Pengfei Deng, Jingkai Ren, Shengbo Lv, Jiadong Feng, and Hongyuan Kang, "Multi-label image recognition in anime illustration with graph convolutional networks," in *Proc. AAAI Conference on Artificial Intelligence*, 2020.

[11] Thomas N. Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. International Conference on Learning Representations*, 2017.

[12] Soubarna Banik, Mikko Lauri, and Simone Frintrop, "Multi-label object attribute classification using a convolutional neural network," *arXiv preprint arXiv:1811.04309*, 2018.

[13] Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, and Lubomir Bourdev, "Panda: Pose aligned networks for deep attribute modeling," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1637–1644.

[14] Olga Russakovsky and Fei-Fei Li, "Attribute learning in large-scale datasets," in *Proc. European Conference on Computer Vision*, 2010, pp. 1–14.

[15] Sheng Huang, Mohamed Elhoseiny, Ahmed Elgammal, and Dan Yang, "Learning hypergraph-regularized attribute predictors," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2015, pp. 409–417.

[16] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.

[17] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. IEEE International Conference on Machine Learning*, 2015, pp. 2048–2057.

[18] Xinrui Wang and Jinze Yu, "Learning to cartoonize using white-box cartoon representations," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8090–8099.

[19] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao, "Attention-driven dynamic graph convolutional network for multi-label image recognition," in *Proc. European Conference on Computer Vision*, 2020, pp. 649–665.

[20] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

[21] Andrew L. Maas, Awni, Hannun, and Andrew Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. International Conference on Machine Learning*, 2013.

[22] Zhaomin Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo, "Learning graph convolutional networks for multi-label recognition and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[23] Zhaomin Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo, "Multi-label image recognition with graphconvolutional network," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5177–5186.

[24] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier, "Collecting image annotations using Amazon's mechanical turk," in *Proc. NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010, pp. 139–147.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[26] Tianshui Chen, Xu, Muxin, Xiaolu Hui, Hefeng Wu, and Liang Lin, "Learning semantic-specific graph representation for multi-label image recognition," in *Proc. IEEE International Conference on Computer Vision*, 2019, pp. 522–531.

[27] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.