Anime Character Identification and Tag Prediction by Multimodality Modeling: Dataset and Model

Fan Yi
School of Computer Science
Fudan University
Shanghai, China
fanyi20@fudan.edu.cn

Jiaxiang Wu
Youtu Lab
Tencent
Shanghai, China
wjxzju@zju.edu.cn

Minyi Zhao
School of Computer Science
Fudan University
Shanghai, China
zhaomy20@fudan.edu.cn

Shuigeng Zhou
School of Computer Science
Fudan University
Shanghai, China
sgzhou@fudan.edu.cn

2023 International Joint Conference on Neural Networks (IJCNN) | 978-1-6654-8867-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/IJCNN54540.2023.10191980

***Abstract*—In recent years, some advances have been achieved in classification and object detection related to animation. However, these works do not take full advantage of the tags and text description content attached to the anime data when they are created, which restricts both the related methods and data to unimodality, consequently leading to unsatisfactory performance. In this paper, we propose a novel multimodal deep learning network for Anime character identification and tag prediction by exploiting multimodal data. Considering that in many realistic scenarios, text annotations accompanying anime may be missing, we introduce the concept of curriculum learning in transformers to enable inference with only one modality. Another challenge lies in that the existing dataset does not meet our demand for large-scale multimodal deep learning. To train the proposed network, we construct a new anime dataset Dan: mul that contains over 1.6M images spread across more than 14K categories, with an average of 24 tags per image. To the best of our knowledge, this is the first dataset specifically designed for multimodal anime character identification. With the trained network, we can identify the anime characters in images and generate the related tags. Experiments show that our method achieves state-of-the-art performance on Dan: mul in animation identification.**

***Index Terms*—Anime character identification; Multimodal network; Dataset; Tag prediction: Curriculum learning.**

I. Introduction

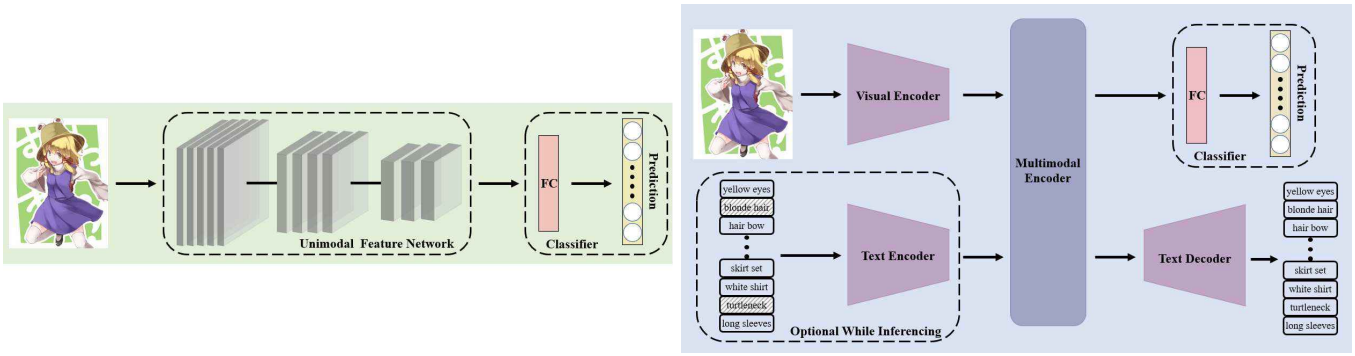
Thanks to the transformer architecture [1] and large image-text datasets [2]–[7], multimodal learning, especially vision-language network has become a paradigm in various computer vision (CV) fields. However, multimodal learning seems to be underexplored when it comes to the animation domain.

In the literature, there are several works of image understanding related to animation. The earliest works [8]–[10] focus on feature extraction of unimodal anime images. However, both the fusion of different poses of the same anime character or sketch-based methods to extract image features, the performance is relatively poor. Subsequent works [11], [12] study object detection in anime images based on facial features, which have obvious limitations. Later, Nguyen *et al.* [13] pioneered the multimodal approach, but its fusion method is inefficient. Recently, Rios *et al.* [14] used a unified structure to fuse modalities, but lacked modal alignment and achieved only modest performance. In summary, even though some attempts have been reported, there is no effective method of modality fusion in the field of animation.

From another perspective, one major reason for the lack of multimodal approaches in the field of animation is the shortage of data. Recently, several anime datasets have been proposed. However, there are still some problems with these datasets. Some of them contain only images and lack text description information, while others have difficulty in reaching the number of image-text pairs required for large-scale multimodal training. For example, Cartoon-11k [15] contains 11,120 images of 146 classes and each class having at least 35 images. WebCaricature [16] is a dataset of 252 subjects with a total of 6,042 caricatures and 5,974 photos. Manga109 [17] is a dataset containing 109 Japanese comic books, up to 21,142 pages in total. Besides the lack of text description information, these datasets are restricted to time-honored animation and collect a limited number of images. Some recent datasets, though may meet the demands of the number and quality of images, have different focuses and are not suitable for multimodal character identification. For example, iCartoonFace [18], consisting of 389,678 images of 5,013 identities, aims at cartoon face recognition, which favors the application of object detection methods and is not accompanied with appropriate text description information. To solve these problems, we try to construct a general and large-scale multimodal anime dataset.

In this paper, we build a new multimodal anime dataset called Dan:mul and propose a Multimodal Fusion and Decoupling Network called MFDN for anime character identification and tag prediction. As shown in Fig. 1, there are some critical differences between our model and the existing unimodal models, including the fusion of multiple modalities and the tag prediction task. We focus on anime character identification and our goal is to boost the accuracy using a large-scale multimodal dataset and to find ways to help the model decouple its dependence on text data, i.e., using multimodal data for training, while using only images as input for inference in line with the original unimodal task. In summary, our major contributions are as follows:

First, we propose a general multimodal character identification model MFDN by using anime data as raw vision and text input. As shown in Fig. 1(b), in the training phase, the model takes images and the corresponding tags as input. We use a transformer-based encoder for both vision and text modalities



(a) Existing unimodal learning framework.

(b) Our multimodal fusion network.

Fig. 1. Comparison between the existing unimodal learning framework and our multimodal fusion network.

and also utilize a transformer block as the multimodal encoder for modality fusion. The model is trained by recognizing the identity of the anime character in the image and reconstructing part of the tags. The introduction of Masked Language Model (MLM) loss and the concept of curriculum learning [19] stably increase the proportion of masked text information, making the model progressively decouple the dependence on text modality. As a result, the model achieves relatively high performance in the inference phase with only vision modality input, which is significant in real-life scenarios.

Second, we construct a large multimodal anime dataset Dan:mul, which contains 1,616,238 images of 14,413 categories. There are 25,404 tag categories in total and an average of 24 tags per image. In order to avoid long-tail distribution, we eliminate image categories with less than 10 images and filter the tags in terms of content and quantity to ensure the quality and generality of the dataset. Overall, MFDN achieves better performance by utilizing Dan:mul than DanbooruAnimeFaces:revamped (DAF:re) [20]. To the best of our knowledge, Dan:mul is the first dataset designed specifically for multimodal anime character identification with an unprecedented number of image categories.

Last but not least, we furthre use MFDN to predict anime character tags. As mentioned above, MFDN can use only images as input in the inference phase, which enables it to handle multi-label classification tasks. In addition, compared to other transformer-based multi-label classification methods, MFDN can predict much more tag categories to achieve a generation-like effect. For example, traditional transformer-based methods usually predict up to hundreds of tags due to memory limitation, while the number of tags our method can predict is determined by the tag categories of the dataset, which can be thousands.

II. Related Work

A. Multimodal Learning

Original multimodal learning mainly considers representation learning and modality fusion to achieve some performance improvements. Most of them are discriminative models [21]–[26] belonging to supervised learning. While discriminative models perform well in the task of classification or regression, they cannot work when there are missing data or patterns. Discriminative models also require a large amount of labeled data, which can be costly to obtain in some applications.

Subsequently, with the popularity of Transformer [1], the emergence of Vision Transformer (ViT) [27] and the success of pre-training in NLP, e.g., BERT [28], multimodal learning also shifted to self-supervised learning and stepped into the pre-training model. Whether it is the single-stream models UNIMO [29], ViLT [30], the dual-stream models ViLBERT [31], LXMERT [31], or eventually, the most popular dual-encoder models CLIP [32], ALBEF [33], BLIP [34], VLmo [35], they all have in common the need for large-scale datasets such as COCO [4], VG [5], SBU [3], CC [2], CC12M [6], LAION [7]. Although pre-training decouples the dependence on a single modality and enables downstream tasks such as image-text retrieval, such models are more demanding on the language modality during training, requiring the input of coherent sentences with semantic meaning.

However, if the text of the dataset has only tags but no syntactic structure and semantic relations, pre-training is difficult to work. Our approach combines the advantages of the above two types of methods, re-adopts supervised multimodal learning to reduce the textual requirements of the dataset and improves the model performance by borrowing the modality fusion method from multimodal pre-training models.

B. Anime Character Identification

Anime character identification is a more challenging image classification task, due to its multiple categories and similar image features. The earliest work focused on feature extraction on unimodal anime images, e.g. [8] fused different pose features of unified anime characters for retrieval, [9] and [10] both used sketch-based approach to extract image features. Later work [11], [12] focused on extracting local representative features, i.e. object detection for face features, for subsequent tasks. Recent work [13] started to use textual information as an aid for the initial fusion of vision-language modality, but the fusion was rather simple and inefficient due to the lack of suitable image encoders like ViT at that time. After that, [14]

used Transformer as an encoder for both modalities, achieving structural unification, but lacking the task of modal alignment and with average performance. Inspired by this model, we add MLM task and the concept of curriculum learning to propose the Multimodal Fusion and Decoupling Network.

C. Multi-Attribute Recognition

Multi-Attribute Recognition also has a wide range of applications in different domains, such as fashion search [36] and product retrieval [37]. Due to the properties of our approach, MFDN can also be derived for the task of tag prediction, which also belongs to multi-attribute recognition. To our best knowledge, this is the first implementation of multi-attribute recognition in the field of animation. In addition, the traditional transformer-based method [38] occupies one transformer embedding length for each attribute to be predicted, which leads to limited scalability and can only cope with dozens of attributes, compared to our method, which has the same number of scalability as the dataset tag categories and can achieve the effect of pseudo-generation.

D. Curriculum Learning

When curriculum learning [19] was first proposed, it was an optimization strategy that centered on training from simple to difficult. Subsequent work [39] has defined it as increasing the diversity and information of the data as training advances, i.e., adding more new training samples and adjusting the weights accordingly. However, curriculum learning is adopted only as a concept in our approach, and the actual operation is somewhat contrary to the definition, because the proportion of masked text information is gradually increased, so the amount of input information obtained by the model is actually gradually reduced, but the training difficulty still maintains a gradual increase. Whatever, with the help of the concept of curriculum learning, MFDN has a more generalized application.

III. Dataset

The key to multimodal learning is the richness of the images and their corresponding text. Thus, a large-scale image-text anime dataset with a large number of character categories and tag categories is necessary for training a general model that can identify anime characters. However, the available datasets are limited in terms of image categories and numbers or lack rich text labels. For these reasons, we constructed a new multimodal anime dataset called Dan:mul.

A. Dataset construction

We build Dan:mul from an existing large online anime database Danbooru [40]. To ensure the generality and quality of the dataset, we collect the latest version (2021) of the database and use only images under the 512px subset. To simplify character identification into a classification task, we keep those images in which only one anime character appears. In addition, since our method is based on supervised multimodal learning, image classes with fewer than 10 images are removed to avoid the long-tail distribution problem. Based on these processes, we construct the image part of the dataset, containing 1,616,238 images with 14,413 categories.

TABLE I
Dan:mul Summary of Train, Test, Val.

	Train	**Test**	**Val**
No.Image	1,125,100	398,412	92,726
Min No. Image / Class	7	2	1
Max No. Image / Class	29,807	10,645	2,130
Avg No. Image / Class	78.06	27.64	6.43
Min No. Tags / Image	0	0	0
Max No. Tags / Image	327	265	194
Avg No. Tags / Image	24.067	24.074	23.904

As an important part of the multimodal dataset, text data are collected and filtered strictly by the following methods. Firstly, we get the tags associated with the image features from the database, and filter out some of the facial expressions that have no real meaning using regular expressions, such as '>_<', '=_=', '^3^', 's.m.s.'. The next essential step is to measure the importance of the tags, which is composed of two steps. a) Count all tags that appear only once in the dataset and remove them. b) Use the TF-IDF metric to measure the importance and value of tags. Since a word or phrase is bound to appear only once as a tag in the text of a single sample, the TF of all tags in a single sample is consistent. In this case, the fewer the total number of a tag occurring in the entire dataset, the larger the IDF, and the larger the TF-IDF, the more representative it is in this sample. However, this cannot account for the importance of this tag in the whole dataset, so we count the TF-IDF of each sample and accumulate the same tags as:

$$TF{-}IDF_{sum-i} = \sum^{c_i} TF \times IDF_i = (\sum^{c_i} \frac{1}{k_i}) \times \log \frac{C}{c_i}, \quad (1)$$

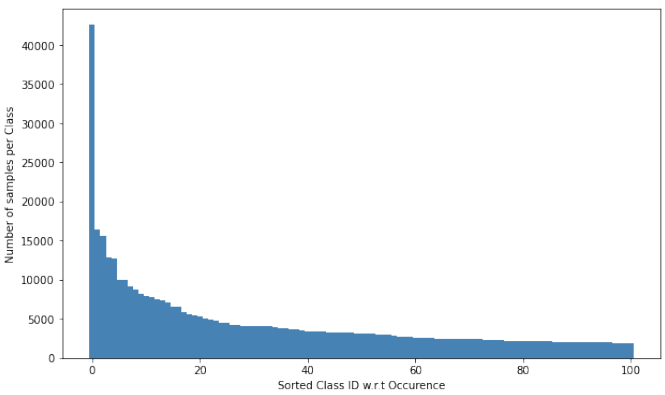
where c_i is the number of samples in which tag i appears, k_i is the total number of tags for samples in which tag i appears, and C is the total sample size.

We set a threshold value of 0.2 as a benchmark to remove the less representative tags and finally obtained the textual part of the dataset, containing 25,404 categories.

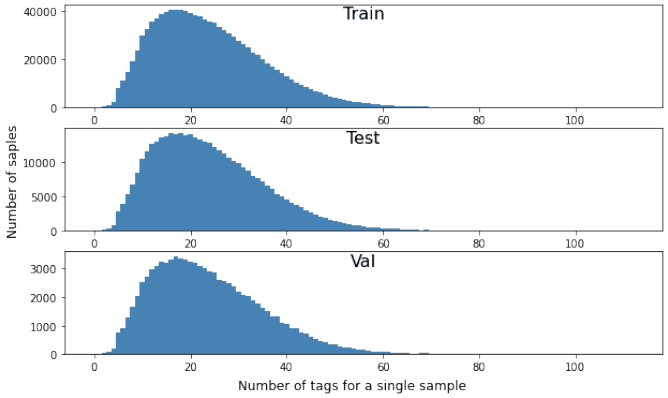
Then we divide the whole dataset into training set, testing set, and validation set on the scale of 0.7, 0.25, and 0.05. In all, we construct a large-scale multimodal anime dataset.

B. Dataset Analysis

As shown in Fig. 2(a), it is clear that the distribution of samples corresponding to each image category is long-tailed, with one class having more than 40,000 images, about twenty classes having more than 10,000 images, and the rest of the classes having less than 5,000 images. In Table I, we present the statistics under three divisions of the dataset. Since the minimum number of samples for a single category is 10, even for the smallest category, at least 1 image is kept in the validation set, which ensures reliability in verifying the model capability. In the training set, the average number of images for the categories is close to 80, ensuring the generality of



(a) Distribution of the number of images per category.



(b) Distribution of the number of tags in the sample on Train, Test, Val.

Fig. 2. The dataset statistics of Dan:mul. (a) The number of images per category is distributed long-tailed. The top 20 categories have a large number of images, and the subsequent categories are more evenly spread. (b) The distribution of the number of tags in the sample under all three sets is relatively approximate, close to the normal distribution of $\mu = 20$.

TABLE II
COMPARISON OF DATASET STATISTICS.

Dataset	Total	Train	Test	Val
DAF:re [20]	463,437	322,947	94,440	46,050
Dan:mul	1,616,238	1,125,100	398,412	92,726

Dataset	Size	Class	Tags
DAF:re [20]	128px	3,263	13,505
Dan:mul	512px	14,413	25,404

TABLE III
SHARED DATASET SUMMARY.

Dataset	Test	Class
Dan:share	22,629	2,623

the dataset. Of course, due to the existence of the long-tailed distribution, the classification and recognition of categories with fewer samples is still challenging.

Regarding the tag analysis in Table I and Fig. 2(b), we can find that the minimum number of tags corresponding to each image is 0 in either division, which means that even if Dan:mul is large-scale, the problem of missing modalities is still unavoidable in some samples. This also motivates us to improve our model framework with the goal of decoupling language modality from vision modality and even going further to achieve tag prediction, which allows us to complement this missing textual information with this model. In addition, each sample has 24 tags on average, and we can also see the distribution of the number of tags in the samples under each division in Fig. 2(b). The training set and testing set are basically the same, and the validation set is relatively less. The number of tags in most samples is concentrated between 10 and 30, which also gives us a reference for the length of text sequences used in our subsequent multimodal models.

Dan:mul and DAF:re [20] are both constructed on the basis of the Danbooru database [40], and their statistics are shown in Table II. There are several main changes:

- Dan:mul is much more expanded in the order of dataset size, close to 4 times.
- Our image resolution is 4 times higher than DAF:re and the number of categories has doubled over 4 times, making it more general and challenging in comparison.
- Our tags, even after comprehensive importance filtering, are still 2 times larger than DAF:re, meaning we can combine more textual information for multimodal learning.

Dan:mul and DAF:re come from different versions of the Danbooru database, the former is the 2021 version and the latter is the 2018 version, and this database is expanded year by year, so these images are homologous. Disregarding the images and text that are filtered out in the process of constructing the dataset, DAF:re can be regarded as a subset of Dan:mul. Based on this, we filter out the shared part of the testing set of these two datasets as a common testing set Dan:share, as shown in Table III, so that it is easy to compare the differences in training effects using different datasets.

IV. METHOD

A. Problem Definition

Given an image v_i with several tags t_i that describe some features of the image, the task is to identify the anime character appearing in the image, assuming that the number of tags is k_i. In addition, this problem is further extended to use multimodal inputs in the training phase but only images as inputs in the inference phase, achieving good performance as well.

B. Multimodal Fusion and Decoupling Network

We propose Multimodal Fusion and Decoupling Network (MFDN), an end-to-end vision-language transformer model that accepts visual and textual inputs, as shown in Fig. 3.

1) Vision Embedding: We adopt ViT-L [27]-style architecture to extract vision features, using only a 12-layer encoder with pre-trained parameters. Each input sample is resized to 128 $\times$ 128 and normalized after data augmentation with

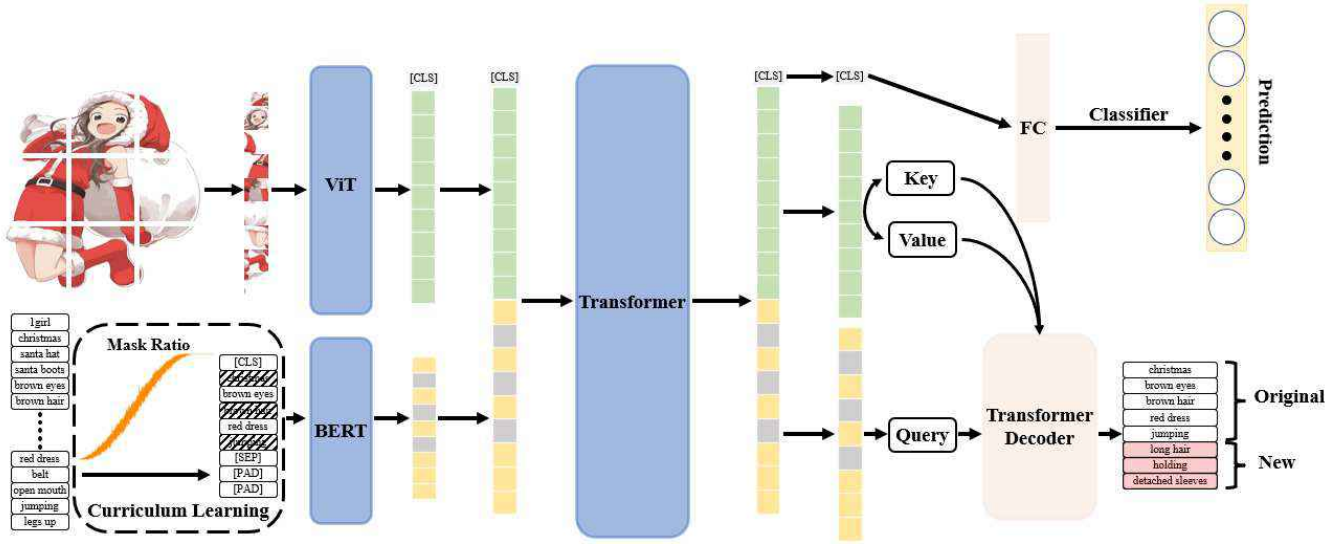


Fig. 3. The network architecture of MFDN. Images and text are encoded by ViT [27] and BERT [28] respectively. The concatenation of the two modality embeddings is fused by Transformer and then split. The CLS is responsible for image classification. The visual modality is used as Key and Value, and the language modality is used as Query of the decoder, and the decoded result is passed through the multi-classification head to get the prediction tags. MLM with curriculum learning strategy is added before the encoding of tags.

random cropping, flipping, attribute changes, etc. The sample is then divided into a list of 16 $\times$ 16-sized patches. A linear projection layer is used to project these patches into the required embedding dimension of ViT, which for ViT-L is a 1024-dimensional embedding. Taking a resized image as an example, an input tensor of shape 128 $\times$ 128 $\times$ 3 (height $\times$ width $\times$ channel) will result in an embedding of 8 $\times$ 8 $\times$ 1024. Of course, we need to flatten and then add CLS, which means we get a 65 $\times$ 1024 vision embedding $V_i \in \mathbb{R}^{l_v \times d}$.

2) Language Embedding: A tag (word or phrase) differs from a sentence in that it does not need to consider the relationship with other texts and has more complete semantic information as a whole. Since our text data are tags, the pre-trained BERT [28] WordPiece (WP) tokenizer is too fine-grained and inappropriate, so we make our own word tokenizer to tokenize tokens in terms of tags, i.e., a tag as a token for input. Since Dan:mul has an average of 24 tags for a sample, we choose $l_t = 32$ as the default sequence length, and the embedding is 1024-dimensional consistent with ViT-L. If k_i exceeds the default length, random sampling will be performed among t_i, and the rest of the structure is the same as BERT-base, which means given the input t_i we will get $T_i \in \mathbb{R}^{l_t \times d}$.

3) Multimodal Encoder: Our Multimodal Encoder is referenced from ViLT [30] and consists of a 12-layer (hidden size 1024) transformer, with an overall structure similar to giving the first half of ViT-L to Vision Encoder and the second half to Multimodal Encoder. The input of the encoder is $M \in \mathbb{R}^{l_m \times d}$, which is the concatenation of vision embedding and language embedding, l_m is the sequence length, obviously $l_m = l_v + l_t$, and d is the dimension of the hidden space embedding. Let $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$ be the projection matrices to project M to the key space, query space, and value space, respectively:

$$Q = MW_Q, K = MW_K, V = MW_V. \tag{2}$$

The embedding matrix M is updated as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V. \tag{3}$$

Consistent with ViT-L, we set d_k to 1024, and we also use a multi-head self-attention (MSA) layer with 16 heads, an MLP layer with the intermediate size 4096, and add layer normalization (LN) to the input of each layer. This multimodal encoder implements the fusion of vision and language modalities in a single-stream manner and captures the fine-grained relationship between each image patch and text token.

4) Multi-category Classification: We use a Multi-category classification objective to learn global multimodal representations. Following the vision-language transformer described above, a linear layer with softmax is used as a classification head and applied to the [CLS] of encoder output to obtain the character identification result. Then we compute a cross-entropy loss as:

$$loss^{cls} = -\sum_{i}^{N} p_i \log q_i, \tag{4}$$

where N is the number of categories, p_i is the target value of the i-th category, and q_i is the predicted probability value of the i-th category.

5) Masked Language Model: In addition to the classification objective, we also use the MLM objective to enhance the language modality encoder and further align the vision and language modalities. Specifically, we randomly discard a portion of the tags and feed the remaining token embeddings

into the BERT encoder. We replace the discarded tags with trainable vectors [MASK] added to the same locations as the original input, as the gray boxes in Fig. 3, in such a way as to create the complete input for the Transformer Decoder. In the 4-layer decoder, we use the language embedding as Query, the vision embedding as Key and Value for cross attention computation, and finally get embedding $T' \in \mathbb{R}^{l_t \times d}$. T' is processed through l_t classification headers to get the tag classification corresponding to each token, and the loss is calculated using cross-entropy:

$$loss^{mlm} = -\frac{\sum_{t_p \in S_M} \sum_{i}^{K} y_{pi} \log x_{pi}}{|S_M|}. \tag{5}$$

S_M is a dynamic set containing the tags that are masked in this sample, but excluding $[CLS_{text}]$, $[SEP]$, and $[PAD]$ that fills the sequence when $t_i < l_t - 2$ (minus fixed occurrences of $[CLS_{text}]$ and $[SEP]$). In addition, t_p is the tag on the p-th token, K is the number of tag categories, y_{pi} is the target value of the i-th category on the p-th token, x_{pi} is the predicted probability value of the i-th category on the p-th token.

We can find that during the execution of the MLM task, there are actually many language modality tokens that are not involved in the calculation of the loss, but these tokens still provide predictions for some tags that may be outside the ground truth, which can be defined as new. It is interesting that some of these new tags also describe some features of the image very well. In all, the final loss is shown below, where $\lambda^{cls} = 1$, $\lambda^{mlm} = 1$:

$$loss = \lambda^{cls} loss^{cls} + \lambda^{mlm} loss^{mlm}. \tag{6}$$

6) Curriculum Learning: We use the concept of curriculum learning to gradually decouple the dependence of our model on language modalities in order to meet the problem of the lack of language modalities in practical applications. As shown in Fig. 3, we do not fix the proportion of masking during the execution of the MLM, but gradually increase it from zero according to the progress of training steps. After that, we mask and hold all the language modality inputs when the training process reaches the halfway point. The masking proportion referenced to the cosine decay, s_{cur} is the current number of training steps, s_{all} is the total number of training steps:

$$ratio = 1 - max(0, \frac{1}{2}(1 + \cos\frac{2 \times s_{cur}}{s_{all}}\pi)). \tag{7}$$

V. Experiments

A. Training Details

MFDN is trained end-to-end for 50 epochs with a batchsize of 64. For the optimization, we use an SGD optimizer with a base learning rate of 0.001 and a learning schedule with warmup and then two-stage cosine decay, which sets a learning rate threshold of 6e-4, and lets the model do a cosine decay from 1e-3 to 6e-4 in the first 30 epochs and from 6e-4 to 1e-12 in the last 20 epochs. The model can use a larger learning rate for a better local search in the early stage and use a smaller learning rate to converge to a better optimization result. This greatly mitigates the consequence of large batchsize.

TABLE IV
DATASET COMPARISON ON DAN:SHARE.

Dataset	Single Modality	
	Top-1 Accuracy	**Top-5 Accuracy**
DAF:re [20]	82.68%	91.93%
Dan:mul	**84.16%**	**92.13%**
Dataset	**Multi Modality**	
	Top-1 Accuracy	**Top-5 Accuracy**
DAF:re [20]	87.93%	95.09%
Dan:mul	**90.89%**	**96.31%**

B. Comparison of datasets

In Table IV, we compare the dataset Dan:mul with the similar dataset DAF:re [20] by training the model using these two datasets separately. To be fair and to avoid the distribution gap between training and testing, we use the corresponding versions of the images in the DAF:re and Dan:mul testing sets for the experiments conducted on Dan:share, respectively, i.e., the images are from the same source, but the preprocessing follows the original dataset's pipeline, respectively.

We use a unimodal version of the model, equivalent to ViT-L, and a multimodal version of the model, without MLM on both datasets for training and testing. As shown in Table IV, the model trained on Dan:mul achieves a 1.5% improvement in unimodality and even a 3.0% improvement in multimodality. These results provide evidence that Dan:mul has improved in image quality, generality, and especially in tag selection.

C. Character Identification

In Table V, we compare our approach with ViT-L [27] on the testing set of Dan:mul and Dan:share. Dan:share must satisfy both Dan:mul and DAF:re filtering conditions, and DAF:re only retains images that appear at least 20 times, which means that the samples in Dan:share see more samples of the same category during training, so the difficulty of this testing set is relatively low. The Dan:mul testing set is more difficult because there are more samples and more categories, but fewer average samples of each category used in training. We would like to test the difference in the performance of the different methods in these two settings.

MFDN without MLM has a very large improvement of 10.7% and 6.7% compared to ViT-L on the two datasets, indicating that multimodal fusion can be very helpful for the character identification task. However, this model performs poorly with only image input and is far inferior to the unimodal trained ViT-L, which confirms that simple modality fusion makes the model overly dependent on language modality, and the addition of our MLM task and the concept of curriculum learning is essential to solve this problem.

As can be seen, with the full configuration, MFDN achieves a certain improvement of 2.0% and 1.7% on the two datasets compared to simple multimodal fusion, indicating that the

TABLE V
COMPARISON OF MFDN AND ITS VARIANTS ON ANIME CHARACTER IDENTIFICATION.

Method	Train Input Modality	Test Input Modality	MLM	Dan:mul		Dan:share	
				Top-1 Accuracy	Top-5 Accuracy	Top-1 Accuracy	Top-5 Accuracy
ViT-L [27]	Single	Single	-	77.16%	86.92%	84.16%	92.13%
MFDN	Multi	Multi	-	87.86%	94.35%	90.89%	96.31%
MFDN	Multi	Multi	✓	**89.87%**	**95.51%**	**92.67%**	**97.19%**
MFDN	Multi	Single	-	41.22%	56.96%	50.11%	67.02%
MFDN	Multi	Single	✓	*78.33%*	*87.42%*	*85.21%*	*92.63%*

TABLE VI
COMPARISON OF PERFORMANCE, OVERHEAD, AND SCALABILITY OF TAG PREDICTION METHODS.

Method		Error	Epoch Time	Prediction Category
L2L [38]	**Top-25 Attritube**	24.64%	75min	25
	Top-200 Attribute	24.94%	240min	200
MFDN	**Top-25 Subset**	26.42%	180min	25K
	Top-200 Subset	28.73%	180min	25K

MLM task is helpful in aligning the two modalities and promoting fusion, while the concept of curriculum learning allows our method to rely only on image input to achieve outstanding performance. The 1.2% and 1.1% improvements are still obtained on both datasets compared to the unimodal training ViT-L. In addition, the performance improvement of MFDN on the Dan:mul test set is slightly more significant than that of Dan:share, implying that our method may be more suitable for handling some more difficult image categories, which is meaningful in practical usage scenarios.

D. Tag Prediction

Given that MFDN uses a curriculum-learning MLM task, we can compare the tag prediction task with L2L [38]. As shown in Table VI, since L2L is based on Transformer structure and does binary classification for each attribute, it cannot support the prediction for too many tags, so we only select the 25 and 200 tags with the most occurrences in the dataset as two tasks for comparison. It can be found that even though our method is slightly inferior to the current state-of-the-art method in terms of error rate, the difference is not great. Our model can support prediction for any subset of 25K tag types, while what L2L can predict is based on the tag categories specified during training. In addition, L2L is difficult to expand due to the memory limitation, and the training time overhead is already 1.5 times larger than our method when the number of tag categories is 200. From this point of view, MDFN has strong scalability, and due to its wider classification head and millions of training samples, it can predict some tags that do not conform to the ground truth but can describe the image features well, which achieves the effect of pseudo-generation.

E. Generation Application

We apply MFDN to several randomly selected images to demonstrate the significance of our tag generation application.

(a) Case of Wakasagihime.

(b) Case of Kitakami.

Fig. 4. Our application results on Dan:mul testing set. Tags that appear in both the ground truth and the prediction are bolded, tags that are valuable in the ground truth but not generated are marked blue, and meaningful tags outside the ground truth are marked red.

In Fig.4(a), this is an anthropomorphic little mermaid, and the overall dress is blue tones. Some important parts of the labeled features such as “blue eyes”, “blue hair”, and “kimono” are generated, but several obvious features such as “head fins”, “smile”, and “solo” are missed. Last but not the least, MFDN generates tags such as “bangs”, “eyebrows visible through hair”, and “long sleeves” that precisely describe the image features. The generation of these tags will raise the error rate of tag prediction, but is practical and can help us better recognize images, and even as high-quality annotations for some images. Similarly, for the girl in Fig.4(b), MFDN identifies the color and style of the skirt, generating “green skirt” and “pleated skirt”. It also identifies the style of the clothing, generating “serafuku”, “neckerchief”, and “long sleeves”. In addition, some features of the face, such as “blush” and “closed mouth”, are finely identified, but unfortunately, the recognition of the two facial features “purple eyes” and “smile” fails, indicating that the prediction of facial features may need to be improved.

VI. Conclusion

In this paper, we propose a novel anime character identification network MDFN using multimodal fusion, MLM, and curriculum learning strategies. We also contribute a new multimodal anime dataset, Dan:mul, which contains over 1.6M images of 14,413 classes, and 25,404 tag categories. MDFN trained on Dan:mul is shown to have good performance in both anime character identification and tag prediction. Our approach demonstrates that language modality can be decoupled from the multimodal fusion task and MDFN can be used as a tag generator, both of which are useful in real-world applications. We expect that our work will inspire more research on effective multimodal modeling in the field of animation.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, vol. 30, 2017.

[2] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *ACL*, 2018, pp. 2556–2565.

[3] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," in *NeurIPS*, vol. 24, 2011.

[4] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.

[5] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, pp. 32–73, 2017.

[6] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *CVPR*, 2021, pp. 3558–3568.

[7] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv preprint arXiv:2111.02114*, 2021.

[8] J. Yu, D. Liu, D. Tao, and H. S. Seah, "On combining multiple features for cartoon character retrieval and clip synthesis," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 5, pp. 1413–1427, 2012.

[9] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa, "Sketch-based manga retrieval using manga109 dataset," *Multimedia Tools and Applications*, vol. 76, pp. 21 811–21 838, 2017.

[10] R. Narita, K. Tsubota, T. Yamasaki, and K. Aizawa, "Sketch-based manga retrieval using deep features," in *ICDAR*, vol. 3. IEEE, 2017, pp. 49–53.

[11] X. Qin, Y. Zhou, Y. Li, S. Wang, Y. Wang, and Z. Tang, "Progressive deep feature learning for manga character recognition via unlabeled training data," in *ACM TURC*, 2019, pp. 1–6.

[12] X. Qin, Y. Zhou, Z. He, Y. Wang, and Z. Tang, "A faster r-cnn based method for comic characters face detection," in *ICDAR*, vol. 1. IEEE, 2017, pp. 1074–1080.

[13] N.-V. Nguyen, C. Rigaud, A. Revel, and J.-C. Burie, "Manga-mmtl: Multimodal multitask transfer learning for manga character analysis," in *ICDAR*. Springer, 2021, pp. 410–425.

[14] E. A. Rios, M.-C. Hu, and B.-C. Lai, "Anime character recognition using intermediate features aggregation," in *ISCAS*. IEEE, 2022, pp. 424–428.

[15] A. Talib, M. Mahmuddin, H. Husni, and L. E. George, "A weighted dominant color descriptor for content-based image retrieval," *Journal of Visual Communication and Image Representation*, vol. 24, no. 3, pp. 345–360, 2013.

[16] J. Huo, W. Li, Y. Shi, Y. Gao, and H. Yin, "Webcaricature: a benchmark for caricature recognition," *arXiv preprint arXiv:1703.03230*, 2017.

[17] A. Fujimoto, T. Ogawa, K. Yamamoto, Y. Matsui, T. Yamasaki, and K. Aizawa, "Manga109 dataset and creation of metadata," in *Proceedings of the 1st international workshop on comics analysis, processing and understanding*, 2016, pp. 1–5.

[18] Y. Zheng, Y. Zhao, M. Ren, H. Yan, X. Lu, J. Liu, and J. Li, "Cartoon face recognition: A benchmark dataset," in *ACM MM*, 2020, pp. 2264–2272.

[19] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *ICML*, 2009, pp. 41–48.

[20] E. A. Rios, W.-H. Cheng, and B.-C. Lai, "Daf: Re: A challenging, crowd-sourced, large-scale, long-tailed dataset for anime character recognition," *arXiv preprint arXiv:2101.08674*, 2021.

[21] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, 2015, pp. 2625–2634.

[22] A. Karpathy, A. Joulin, and L. F. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *NeurIPS*, vol. 27, 2014.

[23] S. Liu, S. Liu, W. Cai, H. Che, S. Pujol, R. Kikinis, D. Feng, M. J. Fulham *et al.*, "Multimodal neuroimaging feature learning for multiclass diagnosis of alzheimer's disease," *IEEE transactions on biomedical engineering*, vol. 62, no. 4, pp. 1132–1140, 2014.

[24] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *EMNLP*, 2015, pp. 2539–2544.

[25] F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[26] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, and J.-M. Odobez, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1583–1597, 2016.

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[29] W. Li, C. Gao, G. Niu, X. Xiao, H. Liu, J. Liu, H. Wu, and H. Wang, "Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning," *arXiv preprint arXiv:2012.15409*, 2020.

[30] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *ICML*. PMLR, 2021, pp. 5583–5594.

[31] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *NeurIPS*, vol. 32, 2019.

[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*. PMLR, 2021, pp. 8748–8763.

[33] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *NeurIPS*, vol. 34, 2021, pp. 9694–9705.

[34] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*. PMLR, 2022, pp. 12 888–12 900.

[35] H. Bao, W. Wang, L. Dong, Q. Liu, O. K. Mohammed, K. Aggarwal, S. Som, and F. Wei, "Vlmo: Unified vision-language pre-training with mixture-of-modality-experts," *arXiv preprint arXiv:2111.02358*, 2021.

[36] K. E. Ak, A. A. Kassim, J. H. Lim, and J. Y. Tham, "Learning attribute representations with localization for flexible fashion search," in *CVPR*, 2018, pp. 7708–7717.

[37] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *CVPR*, 2016, pp. 1096–1104.

[38] W. Li, Z. Cao, J. Feng, J. Zhou, and J. Lu, "Label2label: A language modeling framework for multi-attribute learning," in *ECCV*. Springer, 2022, pp. 562–579.

[39] X. Wang, Y. Chen, and W. Zhu, "A survey on curriculum learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4555–4576, 2021.

[40] Anonymous, D. community, and G. Branwen, "Danbooru2021: A large-scale crowdsourced and tagged anime illustration dataset." [Online]. Available: https://gwern.net/Danbooru2021