This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

A Challenging Benchmark of Anime Style Recognition

Haotang Li, Shengtao Guo, Kailin Lyu, Xiao Yang, Tianchen Chen
Jianqing Zhu* and Huanqiang Zeng*
College of Engineering, Huaqiao University
{lei_houtong, shengtao_guo, Lyv_kailin, 1995131042, 2095121003}@stu.hqu.edu.cn
{jqzhu, zeng0043}@hqu.edu.cn

Abstract

Given two images of different anime roles, anime style recognition (ASR) aims to learn abstract painting style to determine whether the two images are from the same work, which is an interesting but challenging problem. Unlike biometric recognition, such as face recognition, iris recognition, and person re-identification, ASR suffers from a much larger semantic gap but receives less attention. In this paper, we propose a challenging ASR benchmark. Firstly, we collect a large-scale ASR dataset (LSASRD), which contains 20,937 images of 190 anime works and each work at least has ten different roles. In addition to the large-scale, LSASRD contains a list of challenging factors, such as complex illuminations, various poses, theatrical colors and exaggerated compositions. Secondly, we design a cross-role protocol to evaluate ASR performance, in which query and gallery images must come from different roles to validate an ASR model is to learn abstract painting style rather than learn discriminative features of roles. Finally, we apply two powerful person re-identification methods, namely, AGW and TransReID, to construct the baseline performance on LSASRD. Surprisingly, the recent transformer model (i.e., TransReID) only acquires a 42.24% mAP on LSASRD. Therefore, we believe that the ASR task of a huge semantic gap deserves deep and long-term research. We will open our dataset and code at https://github.com/nkjcqvcpi/ASR.

1. Introduction

Anime is a kind of animation products made in east Asia. A glance of anime images is shown in Figure 1. Anime was born in the last half-century and rapidly developed in the past 20 years. A large subculture group containing increasing audiences of anime worldwide is forming. Authors from various regions worldwide have different cultural backgrounds, which resulted in different painting styles in their anime works. With various styles and diverse themes, scenes of anime works could have more complex information than other forms of art or real-scene photos, which is an excellent carrier for research image understanding mechanisms of artificial intelligent.

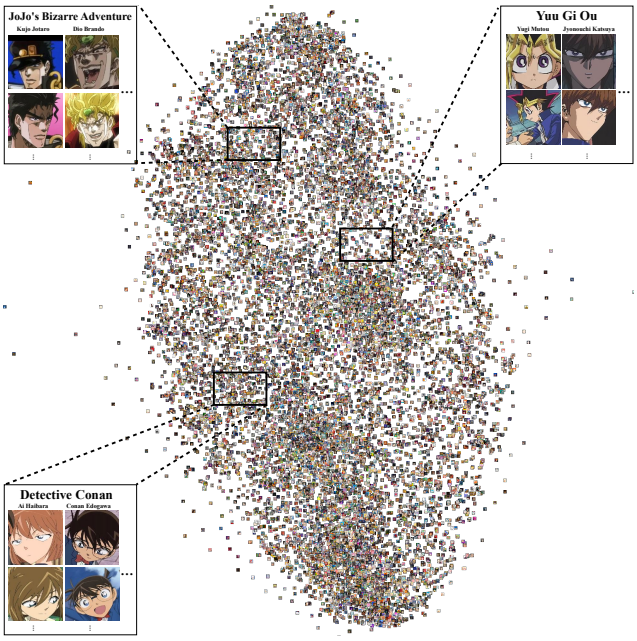


Figure 1. LSASRD Samples.

Human beings can recognize different faces by intuition [19]. With certain learning, they can conclude holistic style of different roles of one anime work [29]. When given images of new roles, one experience can tell those roles are from the same works or not [18]. We are interest in whether computers can do the same thing. For that, we define an anime style recognition (ASR) task for computers, which learns the painting style of anime works to predict whether the images of various given roles are from the same work or not.

*Corresponding authors

Table 1. Comparison of Datasets.

Dataset	iCartoonFace [41]	Danbooru2021 [2]	Manga109 [1]	COMICS [14]	nico-illst [13]	**LSASRD**
Images	389,678	400k	21,142	1.2m	400k	**20,937**
Roles	5,013	Not annotated	500k	Not annotated	Not annotated	**1,829**
Work	Not annotated	Not annotated	109	Not annotated	Not annotated	**190**
Applications	Face Detection and Recognition	Color Rendering	Text Detection	Plot Prediction	Color Rendering	**Style Recognition**

Distinctive from usual biometric recognition [5, 22, 23, 25, 37] tasks paying attention to learning identity context, ASR focuses on learning painting styles of anime works. Specifically, most common biometric recognition, e.g., face [25] recognition and iris [23] recognition, one category contains multiple images. But, in ASR, one category is work includes multiple roles and one role consists of multiple images. Thus, ASR is a more changeling computer vision task, which suffers form complex intra-class variations. In addition to biometric recognition, some researchers have made progress in understanding the arts. For example, Mao *et al*. [17] proposed a DeepArt method to learn joint representations that can simultaneously capture contents and style of visual arts. Garcia *et al*. [7] proposed a multi-modal retrieval method, namely, Text2Art, to retrieve arts by text queries. Shen *et al*. [28] designed a spatially-consistent feature learning method to discover visual patterns in art collections. Some research on images style focuses more on descriptions [8] or classification [3] on a single sample but a bunch of images with rich contextual information. However, anime could not be simply treated as a kind of fine art. Because anime is free in designing that contains complex illuminations, various poses, theatrical colors and exaggerated compositions.

Some anime image datasets [2, 13, 41] have been released recently. For example, Aizawa *et al*. [1] proposed the Manga109 dataset, which only contains scanned images of comic books and each image's text boxes and comic panels are labeled, so that Manga109 is especially for text detection and image splitting. Zheng *et al*. [41] proposed the iCartoonFace dataset for cartoon face detection and recognition. Although iCartoonFace has a great number of images, many images are consecutive frames a cartoon video, causing many images have similar content even miss faces.Ikuta *et al*. [13] proposed the Nico-illust dataset, which mainly is composed of painting images from Nicoico Seiga. Nico-illust is designed for anime color rendering. Figure 2 and Table 1 give a brief comparison. As as result, a ASR dataset richly annotated role and work information is still needed at present.

In this paper, an anime-style recognition dataset (LSASRD) is proposed. LSASRD contains 20937 face images of 1829 roles in 190 anime works between 1928 and 2021. Then, a cross-role protocol is designed by strictly dividing query set and gallery set with no same label of roles existing in both sets simultaneously, which enforces a deep learning model to holistically learn painting styles rather than identity contexts. Moreover, we apply mean inverse negative penalty (mINP) [37], mean average precision (mAP) [6] and cumulative matching characteristics (CMC) [26] curves to evaluate the ASR performance. Finally, we use two state-of-the-art person re-identification methods, namely, AGW [37] and TransReID [5] to construct baseline ASR performance. Surprisingly, we found that existing methods performed poorly on LSASRD, in which AGW performs 40.84% of mAP, and TransReID performs 42.76% of mAP, which shows that ASR is challenging and requires much research to be done. And with a few focus on the style of anime images at present, our research on anime-style images will benefit the development of information recommendations and image retrieval about the

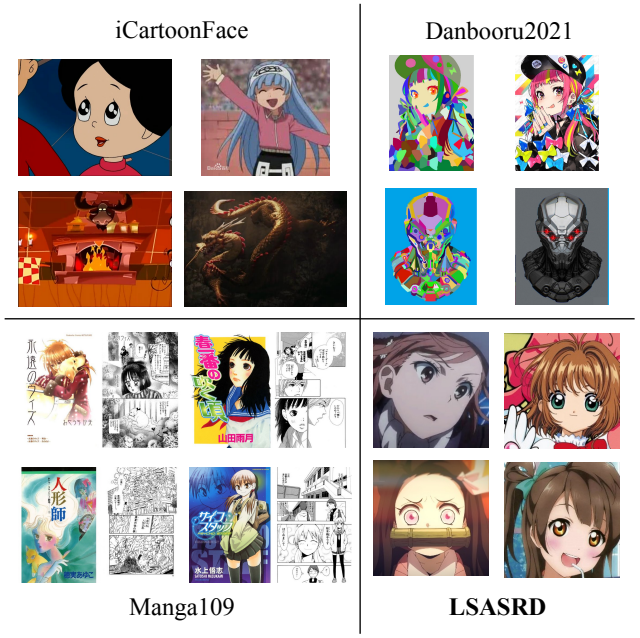


Figure 2. Examples of different datasets.

anime content.

This paper is organized as follows. Section 3 describes the detail of LSASRD. Section 4 designs a cross-role protocol to evaluate ASR. Section 5 introduces two baseline ASR methods. Section 6 reports experiment results and analysis. Section 7 concludes this paper.

2. Related Work

2.1. Anime Recognition

Recent research on anime is more likely to develop image generation [20], colorization [33], style transfer [24] or video interpolation [30] methods. Recognition [27] and detection [38] on anime roles has gradually received this year. Several datasets have been put out, *e.g*. Zhang *et al*. [41] present a new benchmark dataset named iCartoonFace, And provide two types of annotations, face recognition and face detection. Zhang *et al*. [39] make an Illustration Region dataset, this dataset named DanbooRegion., which contains a lot of artistic region compositions paired with corresponding cartoon illustrations. Aizawa *et al*. [1] built a dataset, which is named Manga109. This dataset provides numerous manga images and annotations. Zhang *et al*. [40] make a framework to generate manga from digital illustration and present a data-driven framework to Convert the illustration into three components. Iyyer *et al*. [14] design a dataset, which is named COMICS. Moreover, present three close-style tasks. For other research of anime, LSASRD focuses on the problem of identification. However, unlike conventional recognition problems, LSASRD does not just categorize the same role. It also needs to categorize different roles from the same anime.

2.2. Face Recognition

Face recognition is widely regarded as a sub-problem of image retrieval. Face recognition aims at extracting different features of a human face for recognition. With the development of face recognition, it is a very effective method to extract facial features for classification. Therefore, we choose to extract the facial features of roles in the experiment. The features of face recognition are also summarized. *e.g*., Liu *et al*. [16] introduce the CelebFaces Attributes (CelebA) dataset. This dataset has more than 200K images, each with 40 attribute annotations. Nowadays, face recognition is mainly collected in real life. For these datasets, the proposed methods can achieve very advanced results on the classification of these datasets. In anime, usual feature extraction methods and points may differ from real datasets. Human to abstract humanoid is a tremendous challenge for models.

2.3. Person Re-identification

Person Re-identification aims to use computer vision technology to determine a specific pedestrian in an image or video sequence, which needs to solve the difficulties *e.g*. complex illumination variations, viewpoint or pose. With many methods of Person Re-identification have been proposed which has significantly been developed as it is being more and more attention, some difficulties have been effectively resolved. Both CNN and Transformer neural networks are used for Person Re-identification. *e.g*. Ye *et al*. [37] make a strong AGW baseline. It can reach state-of-the-art or at least comparable performance on both single- and cross-modality Re-ID tasks. Besides, *e.g*. Ye *et al*. [37] propose an evaluation criterion mINP for Person Re-identification. He *et al*. [11] introduced a new transformer-based object ReID framework, And designed two modules, the jigsaw patch module and the side information embeddings. These two methods have achieved SOTA results.

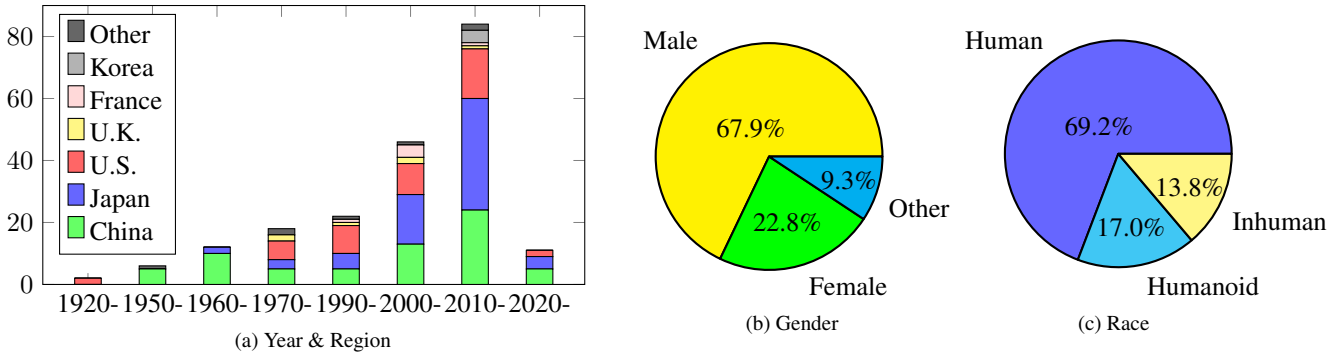
3. Large-Scale Anime Style Recognition Dataset

3.1. Overview

We present a well-labeled challenging dataset, Large Scale Anime Style Recognition Dataset (LSASRD), to facilitate the research on style recognition on anime images by collecting images from 190 anime and cartoon works covering 93 years from 13 countries and regions, 2D and 3D work into consideration concurrently. We choose at most ten roles for each work. All the images are obtained from the Internet. The images in the LSASRD dataset are mainly from existing anime and cartoons. Moreover, some are from comics or games of the same anime series. Unlike illustration or video datasets, we provide a moderate amount of contextual information in a wide variety of styles. LSASRD requires the ability of context understanding of image models. To embody our target of recognition of styles of anime works, we selected works from 13 countries and regions between 1928 and 2021, briefly shown in Figure 3a. Two-thirds of works are from China and Japan, and 39% were broadcast between 2010 and 2020. To differentiate from humanoid roles we are familiar with, we select several inhuman roles especially. In the statistics, 58.3% of roles in LSASRD are male, and 29.5% are female, as Figure 3b. 69.2% of roles are human, 16.9% is humanoid such as furry, and 13.8% is inhuman, such as animal, as Figure 3c.

3.2. Data Collection and Annotation

We fetch a list of anime works from Moegirlpedia and BiliBili. Sorting by popularity, regarding the distribution of eras and regions, we choose 190 works by weighted ran-



(a) Year & Region (b) Gender (c) Race

Figure 3. Distribution of time and regions in LSASRD.

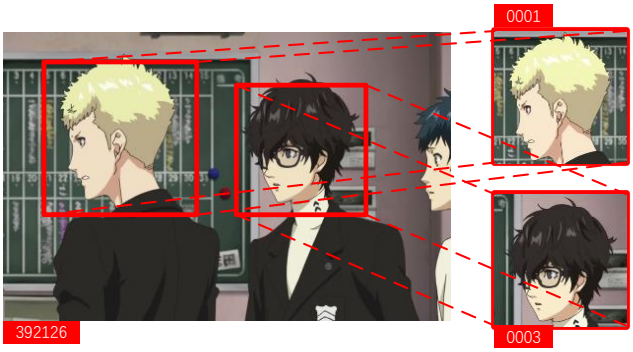


Figure 4. Annotations of LSASRD.

domly. We also consider it the same work for some serial spans an extended period or have many branches works. Then for the roles, we chose both the main characters and supporting roles of the work. While some work contains even less than ten roles, we have to reduce the number of labels. Finally, 1829 labels are presented in our dataset. While the images on the Internet of main roles significantly more than supporting ones, we thus construct a semantic environment that under different dataset partition conditions, the model is requested to learn the style from both lot-shot labels and few-shot labels, which simulate the tasks in the actual scene. We get images from picture searching engines and various online video websites in the public domain or under fair use. Unprocessed data contain 60k images of uneven quality, covering manga, comics, episodes, movies, even photos of peripheral products. Secondary creations are also under our consideration. We mark the bounding boxes and label the images manually using a self-development tool. Five collaborators clean and annotate the data for about two months; 20937 images are selected and well-labeled.

In order to control the uncertainties of the scenes and technical factors of drawing, we only use the images of the face part of each role. Put another way, images in LSASRD currently all are the portrait of roles. Moreover, the whole origin image can be used in more fields, i.e., detection and recognition. We plan to use them in further research of high-level semantics of entities beyond the roles. All images are cropped and only show head of each role, resized in 256px by 256px, as Figure 4 shows. We label the picture by the title of work and mark the face in the images using bounding boxes and labels of whom the faces belong. We also construct the metadata of each work, role, and image handcrafted, *e.g.* year, region, staff of works and the identity information *e.g.* gender, race of roles, to provide context information as rich as possible.

3.3. Challenge

Complex image content conditions. A part of pictures in lower clarity, too bright or too dark, large-area occlusion, exaggerated pose, and special composition distribute uniformly in the dataset. Various resolutions, colors, illumination, angles, or keep outs of objective factors build the typical difficulties of image data. The samples of typical challenges are shown in Figure 5.

Complex image style conditions. Focus on huge semantics gap recognition, images in LSASRD of a role selected in both the original work and secondary creation works, which means that the level of roles could contain different styles. The images are from multiple regions and eras with different styles and roles of human beings, humanoid and inhuman, which means that model can hardly learn the inherent patterns of facial features of a kind of role. Moreover, similar roles from different works and dissimilar roles from the same works are considered the difficulties of the semantics challenge, which is the core of our research.

More complicated than biometric recognition datasets. One subject in biometric recognition datasets [16, 32] commonly contains different images of the same person, such as different shots captured by different cameras for one person, different expressions on the same face. However, LSASRD is much more complex, in where one subject means an anime work of different

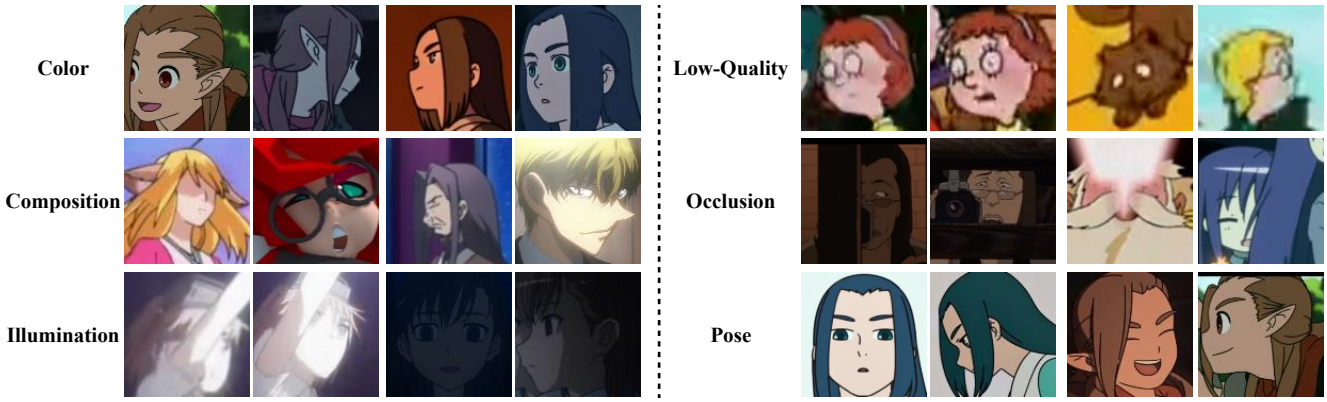


Figure 5. Examples of LSASRD.

roles and each role also of different images in different status. For models, they cannot just learn the identity of images because the identities are different in roles. Unlike the conception of cameras or views, roles can be considered different aspects of a machine to understand semantics between each works. They have to look for high-level common features.

4. Cross-Role Protocol

To evaluate the performance of model of understanding the style differentiation in LSASRD. The details of our protocol are described as follows.

4.1. Training Set and Testing Set Dividing

For our publishing LSASRD, images are randomly divided into two-part for training and testing, as shown in Table 2. Further, the test set is divided into query set and gallery set by the roles of each work and the ratio of 6:4. Notably, the roles of pictures in the query set are strictly do not exist in the gallery set. This cross-role data dividing allows for our protocol to be a cross-role protocol.

Table 2. Splitting of LSASRD.

Subset	Train	Gallery	Query
Work	114	76	76
Role	1097	439	293
Image	12562	5025	3350

In order to reduce the bias from data distribution and overfitting, we introduce 5-fold cross-validation to build our baseline dataset. It is a type of k-fold cross-validation when k = 5. A single 5-fold cross-validation is used with both a training and validation set. The total data set is split into five sets as Table 3 shows. One by one, a set is selected as the validation set. Then, the other four sets are used as training sets until all possible combinations have been evaluated. The training set is used for model fitting. The validation set is used for model evaluation for each hyper-parameter set. The mean value of each five evaluations calculates our baseline performance.

Table 3. 5-Fold Cross-Validation Setting.

Fold-k	1	2	3	4	5
Work	38	38	38	38	38
Role	364	367	367	366	365
Image	4187	4185	4188	4189	4188

4.2. Performance Metric

In order to evaluate the ability of model to learn semantics, we have considered several methods, such as simple classification or clustering. Finally, we choose Re-identification based on deep metric learning. In the Re-identification process, a candidate list sorted by the feature distances between the query and gallery images is returned from the dataset for each given query. The mean Inverse Negative Penalty (mINP) [37], mean Average Precision (mAP) [6] and Cumulative Matching Characteristics (CMC) [26] are used as performance metrics.

The mINP evaluates the ability to retrieve the hardest correct match, providing a supplement for measuring the Re-ID performance, which is formulated as Eq. (1).

$$mINP = \frac{1}{n}\sum_{i}(1 - NP_i) = \frac{1}{n}\sum_{i}\frac{|G_i|}{R_i^{hard}}, \quad (1)$$

where NP measures the penalty to find the hardest correct match, and calculated as $NP_i = \frac{R_i^{hard} - |G_i|}{R_i^{hard}}$, where R_i^{hard} indicates the rank position of the hardest match, and $|G_i|$ represents the total number of correct matches for query i.

The mAP calculate the mean value of average precision

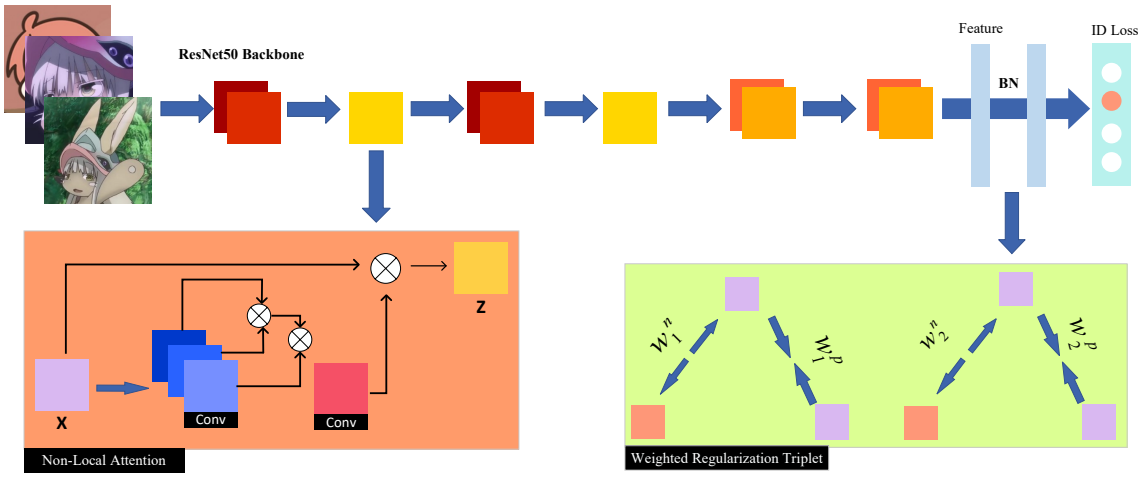


Figure 6. AGW framework [37]

scores for each batch, which is formulated as Eq. (2).

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{N_Q} \tag{2}$$

where N_Q is the number of batch, and AP is calculated as $AP = \frac{\sum_{k=1}^{n} P(k) \times rel(k)}{N_{Relevant}}$, where k is the rank of a size n recall list, N is the number of relevant works. $P(k)$ is the precision at cutoff, and $rel(k)$ indicates whether the k-th recall is correct or not. $N_{Relevant}$ is the number of total query images. In addition, the top-k match rate is also reported in experiments. The average value of the precision function of recall is the Average Precision instead of the single-value metrics precision and recall based on each query and gallery. mAP presents an overall performance for re-identification.

The CMC curves show the probability of a query identity appearing in different-sized candidate lists. The cumulative match characteristics at rank k can be calculated as Eq. (3).

$$CMC@Rank(k) = \frac{\sum_{q=1}^{Q} rel(q,k)(k)}{Q}, \tag{3}$$

where $rel(q,k)$ is an indicator function equaling one if the ground-truth of q image appears before rank k of gallery images, zero otherwise.

5. Baseline Methods

As shown in Figure 6, AGW includes backbone network structure, criterion and the hyper-parameter configuration for ResNet50 with non-local block (ResNet50 NL). For backbone, ResNet50 NL is the default, we also use pure ResNet with 50, 101 and 152 layers. Instance-batch normalization (IBN) leads to build the ResNet50 IBN A model. For a better comparison, we introduce squeeze and excitation (SE) modules in ResNet to construct SE ResNet with 50, 101 and 152 layers. Further, ResNext with SE module pulls in to build the SE ResNext50 and 101.

For criterion, label smoothing cross-entropy and weighted regularized triplet are used to calculate the loss. Label smoothing [21] encourages the differences between the logits of the correct class and the logits of the incorrect classes to be a constant. Weighted regularization inherits the advantage of relative distance optimization between negative and positive pairs but avoids introducing any additional margin parameters.

For the TransReID shown in Figure 7, the base and small size of Vision Transformer (ViT) [5] and data-efficient image transformers (DeiT) [31] are used as backbone, namely ViT-Base, ViT-Small and DeiT-Small. Jigsaw Patch Module (JPM) proposes rearranging the patch embeddings via shift and patch shuffle operations, generating robust features with improved discrimination ability and more diversified coverage. With JPM, we set an experiment of ViT-JPM. With default Transformer stride of 16, we set a group of stride of 12, namely ViT-Stride and DeiT-Stride. For the loss function, id loss and triplet loss are used in TransReID groups. It is worth noting that, for the loss of a batch of samples, each part of loss is the sum of half of the loss value of the first sample and half of the sum of the loss value of the remaining samples.

6. Experiment

To form baseline performance, we apply two recent re-identification methods, i.e., AGW [37] and TransReID [11], to train ASR models on our newly released dataset. In what follows, we firstly introduce our experiment setup and then

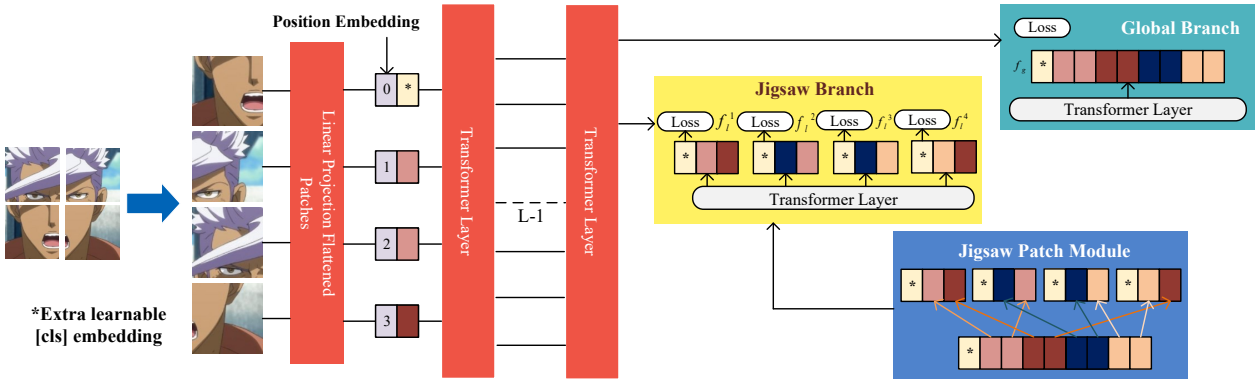


Figure 7. TransReID framework [11].

make an analysis of baseline performance.

6.1. Setup

We apply random horizontal flip, padding, random crop, normalization, and random erasing on training data. Mini-batch stochastic gradient descent are used as optimizer. During the training phase, each mini-batch contains 16 roles and each role has four images. The training epoch number is 100. Due to AGW and TransReID being two different models, their learning rates setting are different. The multi-step learning rate scheduler with warm-up is applied for AGW. The learning rate started at 1×10^{-4} and rise proportionally to 1×10^{-2} as setting in the first ten epochs of AGW. At the epoch of the 40^{th} and 80^{th}, the learning rate will drop to one-tenth of the present value. The cosine learning rate scheduler for TransReID with the margin sets to 0.5 and the scale sets to 30. The learning rate started at a third and rise proportionally to 8×10^{-3} as setting in the first five epochs of TransReID. For the rest of settings, we follow AGW [1] and TransReID [2].

6.2. Result and Analysis

Following the above-mentioned cross-role protocol, we evaluate the AGW model with different backbones (e.g., ResNet50 [10], ResNet NL [34], ViT [5] and DeiT [31]) on LSASRD. The results of AGW and TransReID are shown in Table 4 and in Table 5, respectively. The CMC cures of AGW and TransReID shown in Figure 8 and Figure 9, respectively. From Table 4, we can see that the ResNet50 gets the best mINP and mAP, while ResNet50 NL gets the best Rank1 and Rank5. Figure 8 more comprehensively that

[1]The code of original AGW can be found here https://github.com/mangye16/ReID-Survey.

[2]The code of original TransReID can be found here https://github.com/damo-cv/TransReID.

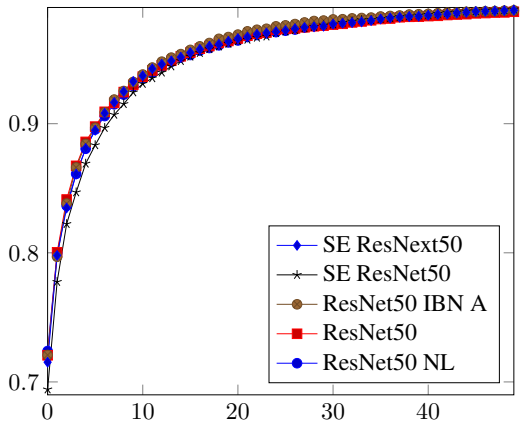


Figure 8. CMC curve of several experiments on AGW.

variations of ResNet50 resulted in the tiny performance difference. And with the increasing of the number of layers, the performance does not increase accompanying. In a possible consideration, with the same configuration of hyper-parameters, deeper models do not give full play to their due performance. According to Table 5, ViT-stride model get the best mINP, while ViT-small get the best mAP, Rank1 and Rank5. We can reasonably doubt that the ViT model pretrained on a huge amount of concrete images needs to be adjusted to fit our abstract images dataset.

We found no significant difference in experimental results of AGW based model and TransReID based model.It could be the reason that the configuration of models above focuses on images of natural persons. Current models have a good performance in extracting the texture feature of an image. However, they lack a mechanism for learning the semantics information in an image. We need a more specific configuration to explore the capacity limit of models.

Based on GradCam++ [4], we visualize the gradient and

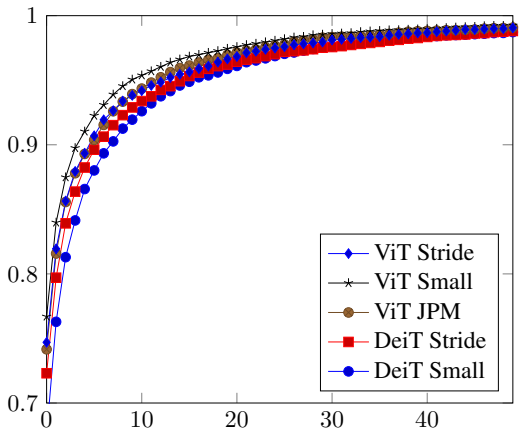


Figure 9. CMC curve of several experiments on TransReID.

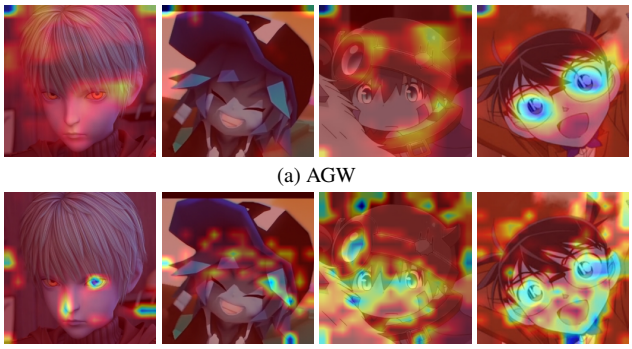
Table 4. Comparisons of performance on AGW with different backbones on LSASRD.

Metrics	mINP	mAP	R1	R5
ResNet50 [10]	**12.48**	**40.84**	72.06	88.60
ResNet50 NL [15]	12.40	40.80	**72.50**	**88.18**
ResNet101 [10]	12.30	40.18	70.78	87.28
ResNet152 [10]	12.26	40.34	71.82	88.04
SE ResNet50 [12]	10.76	38.10	69.44	86.90
SE ResNet101 [12]	10.86	38.42	70.10	86.86
SE ResNet152 [12]	10.28	36.80	67.26	86.10
SE ResNext50 [35]	10.90	39.44	71.52	88.10
SE ResNext101 [35]	9.32	37.56	71.54	88.52
ResNet50 IBN A [36]	10.90	40.74	71.52	88.10

Table 5. Comparisons of performance on TransReID with different Transformer settings on LSASRD.

Metrics	mINP	mAP	R1	R5
DeiT-Small [31]	10.58	36.58	67.54	86.56
DeiT-Stride [31]	11.34	39.66	72.32	88.22
ViT-Small [5]	12.48	**42.76**	**76.68**	**91.04**
ViT-Base [5]	11.34	36.70	65.70	82.98
ViT-JPM [11]	12.72	41.88	74.16	89.30
ViT-Stride [5]	**13.14**	42.24	74.72	89.34

activation of the last Bottleneck block of layer4 module of AGW using the ResNet50 NL backbone and the first LayerNorm layer of the first block in b1 module of TransReID using ViT. From Figure 10, we can observe that, for the first two 3D anime image samples, few features are noticed by AGW while TransReID notices the edges and shadows. For the last two samples, both models are attractive in the significant features of the roles, Kabuto of the third sample and the big and unique drawing style of eyes of the fourth sample. While TransReID mentioned more parts of the sample than AGW.



(a) AGW

(b) TransReID

Figure 10. Heat map [9] of samples. Samples are numbered as 1,2,3 and 4 of both AGW and TransReID from left to right.

7. Conclusion

This paper presents a challenging anime style recognition (ASR) benchmark, which explores the semantic understanding capability of the deep learning model. Concurrently, to promote research in the field of anime-style images, we construct a large-scale ASR dataset (LSASRD). With 20937 images of 1829 roles in 190 works, LSASRD contains rich metadata of every label and all images collected handcrafted. We design a cross-role protocol to evaluate ASR performance. Conducting different experiments on the LSASRD with two SOTA re-identification methods, we report the baseline performance on various models. Experiment results show that SOTA re-identification methods cannot achieve satisfactory performance on the ASR benchmark of all, and it shows that current methods are inadequate to extract abstract attributes, i.e., the features of between huge semantic gap. ASR provides a framework for research both the ability of semantic understanding of model and application of anime images.

Acknowledgements

This work was supported by National Natural Science Foundation of China under the Grants 61976098 and 61871434; National Key R&D Program of China under the Grant 2021YFE0205400; Natural Science Foundation for Outstanding Young Scholars of Fujian Province under the Grant 2019J06017; Collaborative Innovation Platform Project of Fuxiaquan National Independent Innovation Demonstration Zone under the Grant 2021FX03; National Training Program on Undergraduate Innovation and Entrepreneurship of China.

References

[1] Kiyoharu Aizawa, Azuma Fujimoto, Atsushi Otsubo, Toru Ogawa, Yusuke Matsui, Koki Tsubota, and Hikaru Ikuta. Building a manga dataset "manga109" with annotations for multimedia applications. *IEEE MultiMedia*, 27(2):8–18, 2020. 2, 3

[2] Anonymous, Danbooru community, and Gwern Branwen. Danbooru2021: A large-scale crowdsourced and tagged anime illustration dataset. https://www.gwern.net/Danbooru2021, January 2022. Accessed: DATE. 2

[3] Eva Cetinic, Tomislav Lipic, and Sonja Grgic. Fine-tuning convolutional neural networks for fine art classification. *Expert Systems with Applications*, 114:107–118, 2018. 2

[4] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *arXiv preprint arXiv:1710.11063*, 2017. 7

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2, 6, 7, 8

[6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 5

[7] Noa Garcia and George Vogiatzis. How to read paintings: Semantic art understanding with multi-modal retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018. 2

[8] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. *Advances in neural information processing systems*, 28:262–270, 2015. 2

[9] Jacob Gildenblat and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021. 8

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7, 8

[11] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15013–15022, October 2021. 3, 6, 7, 8

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 8

[13] Hikaru Ikuta, Keisuke Ogaki, and Yuri Odagiri. Blending texture features from multiple reference images for style transfer. In *SIGGRAPH Asia Technical Briefs*, 2016. 2

[14] Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daume, and Larry S Davis. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 7186–7195, 2017. 2, 3

[15] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 8

[16] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018. 3, 4

[17] Hui Mao, Ming Cheung, and James She. Deepart: Learning joint representations of visual arts. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1183–1191, 2017. 2

[18] Mark P Mattson. Superior pattern processing is the essence of the evolved human brain. *Frontiers in neuroscience*, page 265, 2014. 1

[19] Elinor McKone, Kate Crookes, Linda Jeffery, and Daniel D Dilks. A critical review of the development of face recognition: Experience is less important than previously believed. *Cognitive neuropsychology*, 29(1-2):174–212, 2012. 1

[20] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning gans. In *CVPR AI for Content Creation Workshop*, 2020. 3

[21] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. *When Does Label Smoothing Help?* 2019. 6

[22] João Neves, Fabio Narducci, Silvio Barra, and Hugo Proença. Biometric recognition in surveillance scenarios: a survey. *Artificial Intelligence Review*, 46(4):515–541, 2016. 2

[23] Kien Nguyen, Clinton Fookes, Raghavender Jillela, Sridha Sridharan, and Arun Ross. Long range iris recognition: A survey. *Pattern Recognition*, 72:123–143, 2017. 2

[24] Ori Nizan and Ayellet Tal. Breaking the cycle - colleagues are all you need. In *IEEE conference on computer vision and pattern recognition (CVPR)*, 2020. 3

[25] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015. 2

[26] P Jonathon Phillips, Hyeonjoon Moon, Syed A Rizvi, and Patrick J Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, 22(10):1090–1104, 2000. 2, 5

[27] Xiaoran Qin, Yafeng Zhou, Yonggang Li, Siwei Wang, Yongtao Wang, and Zhi Tang. Progressive deep feature learning for manga character recognition via unlabeled training data. In *Proceedings of the ACM Turing Celebration Conference - China*, ACM TURC '19, New York, NY, USA, 2019. Association for Computing Machinery. 3

[28] Xi Shen, Alexei A. Efros, and Mathieu Aubry. Discovering visual patterns in art collections with spatially-consistent feature learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[29] Wang Shugen. Framework of pattern recognition model based on the cognitive psychology. *Geo-spatial Information Science*, 5(2):74–78, 2002. 1

[30] Li Siyao, Shiyu Zhao, Weijiang Yu, Wenxiu Sun, Dimitris Metaxas, Chen Change Loy, and Ziwei Liu. Deep animation video interpolation in the wild. In *CVPR*, 2021. 3

[31] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers amp; distillation through attention. In *International Conference on Machine Learning*, volume 139, pages 10347–10357, July 2021. 6, 7, 8

[32] Mateusz Trokielewicz, Adam Czajka, and Piotr Maciejewicz. Assessment of iris recognition reliability for eyes affected by ocular pathologies. In *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–6. IEEE, 2015. 4

[33] Xinrui Wang and Jinze Yu. Learning to cartoonize using white-box cartoon representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[34] Bryan Ning Xia, Yuan Gong, Yizhe Zhang, and Christian Poellabauer. Second-order non-local attention networks for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3760–3769, 2019. 7

[35] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 8

[36] Jianping Shi Xingang Pan, Ping Luo and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *ECCV*, 2018. 8

[37] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 3, 5, 6

[38] Bin Zhang, Jian Li, Yabiao Wang, Zhipeng Cui, Yili Xia, Chengjie Wang, Jilin Li, and Feiyue Huang. Acfd: Asymmetric cartoon face detector, 2020. 3

[39] Lvmin Zhang, Yi Ji, and Chunping Liu. Danbooregion: An illustration region dataset. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 137–154. Springer, 2020. 3

[40] Lvmin Zhang, Xinrui Wang, Qingnan Fan, Yi Ji, and Chunping Liu. Generating manga from illustrations via mimicking manga creation workflow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5642–5651, 2021. 3

[41] Yi Zheng, Yifan Zhao, Mengyuan Ren, He Yan, Xiangju Lu, Junhui Liu, and Jia Li. Cartoon face recognition: A benchmark dataset. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2264–2272, 2020. 2, 3