Article

Automatic Detection of Discrimination Actions from Social Images

Zhihao Wu ¹ , Baopeng Zhang ^{1,*}, Tianchen Zhou ¹, Yan Li ¹ and Jianping Fan ²

¹ School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China; 19125255@bjtu.edu.cn (Z.W.); 18301121@bjtu.edu.cn (T.Z.); 17120395@bjtu.edu.cn (Y.L.)

² AI Lab, Lenovo Research, Beijing 100094, China; jfan1@lenovo.com

* Correspondence: bpzhang@bjtu.edu.cn

Abstract: In this paper, we developed a practical approach for automatic detection of discrimination actions from social images. Firstly, an image set is established, in which various discrimination actions and relations are manually labeled. To the best of our knowledge, this is the first work to create a dataset for discrimination action recognition and relationship identification. Secondly, a practical approach is developed to achieve automatic detection and identification of discrimination actions and relationships from social images. Thirdly, the task of relationship identification is seamlessly integrated with the task of discrimination action recognition into one single network called the Co-operative Visual Translation Embedding++ network (CVTransE++). We also compared our proposed method with numerous state-of-the-art methods, and our experimental results demonstrated that our proposed methods can significantly outperform state-of-the-art approaches.

Keywords: discrimination action recognition; relationship prediction; representation learning

Citation: Wu, Z.; Zhang, B.; Zhou, T.; Li, Y.; Fan, J. Automatic Detection of Discrimination Actions from Social Images. *Electronics* **2021**, *10*, 325. <https://doi.org/10.3390/electronics10030325>

Academic Editor: Byung Cheol Song
Received: 27 December 2020
Accepted: 26 January 2021
Published: 30 January 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In line with the popularity of smart devices, most social platforms, such as Facebook and Twitter, have incorporated image sharing into their important functions. Many people, particularly young users of social networks, may unintentionally share their discriminatory images (i.e., social images with discriminatory actions or attitudes) with the public, and may not be aware of the potential impact on their future lives and negative effects on others or individual groups. Unfortunately, social images and their tags, and text-based discussions, may frequently reveal and spread discriminatory attitudes (such as race discrimination, age discrimination, gender discrimination, etc.). Control over the sharing of these images and discussions has become one of the most significant challenges on social networks. Numerous application domains leverage artificial intelligence to improve application utility and automation level, such as communication systems [1], production systems [2,3] and human–machine interaction systems [4,5]. Many social network sites may potentially abuse the technologies of artificial intelligence and computer vision in automatically tagging human faces [6,7], recognizing users' activity and sharing their attitudes with others [8], identifying their associates, and, in particular, sorting people into various social categories. However, most social platforms have some images that contain discriminatory information, intentionally or unintentionally, and these discriminatory images can significantly affect people's normal life [9]. Therefore, to avoid the problem of users widely revealing their data to the public, discriminatory image recognition is an attractive approach for preventing direct image discrimination during social sharing. In particular, capturing various types of image discrimination and their most relevant discrimination-sensitive object classes and attributes, such as discriminative gestures, is an enabling technology for avoiding discrimination propagation. However, there are no relevant research works on discriminatory image recognition tasks, including necessary datasets and related algorithms.

The discrimination-sensitive object classes and attributes appearing in social images can be used to effectively characterize their sensitivity (discrimination). Thus, supporting the understanding of deep images plays an important role in achieving more accurate characterization and identification of discriminatory attitudes in social images (i.e., image discrimination). In addition to discrimination gestures, emotional interaction exists between persons, that is, the discriminatory relationship has a subject and an object. The semantics of the comparison of facial expressions between the subject and object may strengthen the judgement of the discriminatory relationship. We regard discriminatory detection as a classification problem, but a person's gestures and interactions are not suited to the use of the classical image processing method [10]. It can be seen from the above analysis that discriminatory image recognition can adopt both the relational model and the traditional action recognition model. Therefore, a recognition scheme combining the action and the relationship is proposed in this paper. We optimized the existing relationship model by exploiting semantic transformation and incorporated the action model to propose the Co-operative Visual Translation Embedding network++(CVTransE++).

According to our investigation, no dataset on discriminatory action has been compiled to date. We propose a data set named the Discriminative Action Relational Image dataset (DARIdata), which contains a total of 12,287 images, including six scenes. To make the dataset applicable to both the relational model and the action model, we defined the persons and the relationships according to the discriminatory actions, 11 person characteristic categories, and 17 relationship categories.

In this study, a discriminatory image dataset was constructed, and a network model was built on the dataset to achieve discriminatory image recognition (as shown in Figure 1). To summarize, the main contributions are three-fold.

Figure 1. With a discriminatory or friendly image as input, the model outputs the action category of persons in the image and the relationship category of the persons. – indicates be discriminated, + indicates discriminatory.

First, we propose a fully annotated dataset (DARIdata) to assist with the study of discriminatory action recognition and discriminatory relation prediction. To the best of our knowledge, this is the first benchmark dataset used for discriminatory emotional relationships.

Second, we propose an improved method of the traditional VTransE model (VTransE++) to improve semantic representation capability for discriminatory relationship recognition by integrating a multi-task module and feature enhancement units. Based on the experimental results, the proposed model can effectively reduce the misjudgment rate of relationship recognition.

Finally, we propose the CVTransE++ model based on feature fusion to realize the combination of the action and relationship models. Through comparison experiments, the effectiveness of this work was verified.

2. Related Work

2.1. Discrimination-Related Dataset

Discrimination is a kind of social emotion relation, including sentimental interaction and language interaction. Sentiment analysis is a popular research topic and has produced sentiment datasets in different fields. For example, JAFFE (Japanese Female Facial Expression Database) [11] and CK+ (Extended Cohn-Kanade Dataset) [12] propose datasets in the field of images; IMDB (Internet Movie Database) [13] and SemEval (Semantic Evaluation on Sentiment Analysis) [14] present datasets in the field of natural language; IEMOCAP (Interactive Emotional Dyadic Motion Capture Database) [15] produced a dataset in the field of audio; and DEAP (Database for Emotion Analysis Using Physiological Signals) [16] introduced a dataset in the field of brain waves. However, the sentiment categories of these datasets are relatively broad, and it is impossible to extract relevant descriptions of discriminatory emotions.

In the field of natural language, discriminatory emotion recognition has been previously researched. Ref. [17] successfully identified the discriminatory nature of short texts by constructing a discrimination vocabulary dictionary. Ref. [18] surveyed the methods used in the automatic detection of hate speech in text processing. However, in social image sharing, no relative method has been proposed to detect images' discriminatory emotion.

Discriminatory emotion can be considered as an emotional reaction and behavior occurring among different groups. This kind of emotional relationship is similar to the general relationship. VG (Visual Genome) [19] and VRD (Visual Relationship Detection) [20] are two generally accepted relational datasets, which include relational categories, such as predicate, action, spatial position, preposition, comparative, and verb. Although the relevant content cannot be extracted from the existing data sets, the method of making and labeling the data sets is noteworthy.

2.2. Action and Relation Recognition Method

Relationship-based or action-based methods can be employed in discriminatory and sentiment detection. We therefore first briefly discuss the advantages and disadvantages of these two methods.

Generally, the task of action recognition includes two steps: action representation and action classification. Most of the existing action recognition work focuses on feature extraction [21]. The features used for motion representation in static images generally include three categories: human key points [8,22], scenes or interactive objects [23–26], and human body local information [27,28]. Different action representation methods have different requirements for data set annotation, and different application scenarios require different types of annotation information. The action representation method based on key points of characters can accurately obtain their pose information, but has relatively high requirements for dataset annotation. Action representation based on scenes or objects interacting with people is suitable for action scenes that require fixed places or tools, such as playing violin or badminton. Most such studies pay attention to human–object interaction and ignore human–human interaction. The expression of discriminatory emotion does not require a specific medium and is not limited by the scene, thus action representations obtaining through scenes or interactive objects are not suitable for discriminatory images. Figure 2 shows images of discriminatory characters obtained from the Internet. It can

be seen that the expression of discriminatory actions is often accompanied by disdainful expressions or unfriendly gestures. The hand and face represent local information of the character. This information can provide an effective basis for recognizing discriminatory actions, and the labeling content is relatively simple compared with the key points of the human body. Therefore, the method based on local information is more suitable for discriminatory images.

Figure 2. Discriminatory images obtained from the Internet.

With the development of existing technologies, previous research is not limited to conventional tasks such as object recognition, segmentation, and tracking. Image understanding has attracted an increasing amount of attention, such as Visual Question Answer (VQA), image generation, and image captioning. Visual relation recognition is the foundation of the image comprehension task, and reasonable usage of visual relationships can also provide effective assistance for object detection [29], image retrieval [30], and other tasks [31]. The basic idea is that the relationship information in specific scenes can be used to play an effective auxiliary role. Discrimination is a directional relationship, thus the judgment of discriminatory emotional relationships might also assist in the recognition of discriminatory images. Visual relationship recognition methods can be roughly divided into three categories: the language prior-based method [20,32], graph neural network-based method [33–38], and representation learning-based method [39–41]. Prior knowledge of language plays an important role in the prediction of visual relationships. However, the prediction result tends to be based on the frequent content in the text and ignores the visual information. Graph neural networks can visually represent the entities and relationships in images, and can effectively use the context information between objects for iteration and reasoning. This method is mostly used in scenarios such as visual reasoning and question answering. Representation learning, also known as methods based on semantic transformation, pays more attention to relationship pairs and can obtain relationship categories directly through the visual features of two objects, which is more suitable for this work.

In summary, discrimination image recognition can leverage the action recognition model based on local information and the relationship model based on semantic transformation. The action recognition model based on local information can focus on the expression of local discriminatory actions of characters, but loses the information of the emotional receiver. The relationship model based on semantic transformation considers the information of emotion sender and receiver at the same time, but cannot focus on the

action details of characters. Therefore, combining the two kinds of schemes to compensate for each of the other's shortcomings is a feasible solution.

2.3. Relationship Model Based on Semantic Transformation

This paper employs the relational model based on semantic transformation as our basic network. The method of semantic transformation was first proposed in the field of natural language, which was induced by the translation invariance of word vectors [42], and triggered a wave of representation learning [43–47]. The method based on semantic transformation can transform the features of objects from visual space to relational space, and then obtain the relationship representation between objects and identify the relationship [38–41]. This method uses the triple $\langle \text{subject}, \text{relationship}, \text{object} \rangle$ to represent relationships, and $\text{subject} + \text{predicate} \approx \text{object}$. VTransE [39] is a method that successfully applies the idea of semantic transformation to the field of vision. It represents the subject and object in the triple as any object in the image, and the predicate as the relationship between two objects. Figure 3 shows the process of the semantic transformation of this model.

Figure 3. The process of the semantic transformation module of VTransE.

The semantic transformation module of VTransE consists of three parts: visual feature extraction, semantic transformation, and relationship classification. In the visual space, the convolution module is used to extract the visual features of the subject and object. In the semantic transformation, the features of the subject and object are mapped from visual space to relational space. The features of the subject and object in relational space are calculated to obtain the feature of relationships. Finally, image recognition is obtained through the relationship classification.

VTransE successfully applied the idea of semantic transformation to the field of vision. The model has the advantages of simple design, few parameters, and easy implementation. It also performs well on existing relational data sets. Therefore, we took VTransE as our basic network and further optimized the model according to the characteristics of discriminatory images.

3. Proposed Dataset: DARIdata

In this section, we elaborate on the detail of our proposed dataset: (1) definition and categories of discriminatory action; and (2) data collection and augmentation.

3.1. Definition and Categories of Discriminatory Action

From the perspective of sociology, discrimination is an emotional reaction and behavior between different interest groups. We find that discrimination is a kind of human emotion, and emotions play a crucial and often constitutive role in all of the important phases of action preparation and initiation [48]. That is, emotions affect the expression of actions. Therefore, people always have body movements that are affected by discrimination. The main work of this paper is to find and classify these representative actions.

Body language is a visual signal used by people in social interactions, including action, postures, and facial expressions. Body language can be divided into facial body language, hand body language, and posture body language according to the action of different parts of the human body [49]. Therefore, we describe the action discrimination in three aspects: facial body language, hand body language, and postural body language.

Among the three types of body language, the gesture is considered the main influencing factor. In the act of discrimination, facial expressions are mostly disgusted and angry, which is a commonality of discriminatory actions. Fixed discriminatory gestures generally have a relatively certain posture. Therefore, we classify the dataset based on the gesture.

Discriminatory actions are classified according to gestures. Symbolic body language has unique cultural and temporal attributes [50]. For example, making a circle with the thumb and forefinger and straightening the other three fingers means “OK” to Americans and Parisians, while it means zero or having nothing to people living in southern France, “I’m going to kill you” to Tunisians, money to Japanese, and blasphemy to Brazilians. Because the same body language has different meanings in different cultures, we select the non-friendly gestures summarized in Wikipedia [50], resulting in eight recognized discriminatory gestures.

We divide discriminatory gestures into the following eight categories: bras d’honneur (Bras d’honneur: A gesture of ridicule, an impolite and obscene way of showing disapproval that is most common in Romanic Europe); the middle finger (The middle finger: An obscenely obnoxious gesture, common in Western culture); fig sign (Fig sign: A mildly obscene gesture used at least since the Roman age in Western Europe that uses two fingers and a thumb. This gesture is most commonly used to deny a request); akanbe (Akanbe: An ironic Japanese facial expression, also used as a mockery, considered to be an immature mocking gesture); loser (Loser: A hand gesture generally given as a demeaning sign, interpreted as “loser”); talk to the hand (Talk to the hand: An English language slang phrase associated with the 1990s. It originated as a sarcastic way of saying one does not want to hear what the person who is speaking is saying); the little thumb (The little thumb: A gesture that indicates insignificance or clumsiness and will not be used in a formal situation); and thumb down (Thumb down: A scornful gesture indicates that the other party is weak). An example of each type of discriminatory action is provided in Figure 4a.

Figure 4. Discriminative Action Relational Image dataset (DARIdata). (a) Eight types of discriminatory actions summarized in DARIdata. The first line shows bras d’honneur, the middle finger, fig sign, akanbe; the second line shows the loser, talk to the hand, the little thumb, and thumb down. (b) Examples of interference images: akanbe—pushing glasses, bras d’honneur—encouragement. (c) Distribution of images per category. (d) The number of actions per category. (e) The number of relationships per category. – indicates be discriminated, + indicates discriminatory.

3.2. Data Collection and Augmentation

Before the data enhancement, our dataset had a total of 10,177 images, taken by 12 people, including six scenes. To express discriminatory action relationships, each image contained two people. We took 2582 friendly images as interference sets. These interference images contain actions similar to discriminating images, such as akanbe—pushing glasses and bras d’honneur—encouragement, as shown in Figure 4b. The remaining 7595 images

are discriminatory, and each type of discriminatory action is evenly distributed, as shown in Figure 4c.

In our dataset, we mainly labeled the relationship categories of every pair of people, the action categories of each person, the boxes of each person, and 211 of their hands and faces. There are 11 action categories (eight types of discriminatory, be discriminated, friendly, and normal) and 17 relationship categories (eight types of discriminatory, eight types of being discriminated and friendly). The number of discriminatory actions in each category can be obtained from Figure 4d. It can be seen that the number of discriminatory actions in each category is relatively balanced, and the sum of all discriminatory actions is relatively balanced with the non-discriminatory, friendly, and normal categories. The number of each type of discriminatory relationship can be obtained from Figure 4e.

Annotations contain categories and positions of subjects and objects, and the relationship categories. The position information is composed of the global and local (hand–face) bounding boxes of the character, and the boxes are labeled by weak supervision. Regarding weak labeling, we first selected 2000 images to train a Faster R-CNN [51] model that can detect faces and hands, and then we used this model to process all of the images. We obtained each person’s mask and box using Mask R-CNN [52], and finally, all global and local boxes were obtained using the human mask and the boxes of hands and faces.

We expanded the number of datasets and scenes by data augmentation. We used background replacement for data augmentation. Because deep semantic segmentation networks represented by Mask R-CNN [52] often lose edge information, we chose to use SSS (semantic soft segmentation) [53] in our work, which combines semantic features with traditional methods. About 2170 available images were selected from the images with background replacement, and the manual labels of the original images were used as the labeling information. Figure 5 shows a comparison of the two segmentation methods and the results of the data augmentation.

Figure 5. Comparison of the mask R-CNN and semantic soft segmentation (SSS). SSS can reduce the loss of local information.

4. Recognition Model of Discrimination Actions (CVTransE++)

To integrate the visual information of the emotional receiver, we used the relationship model as the basic network and optimized its semantic transformation module. To take into account the advantages of the action model and the relationship model, we proposed the Co-operative Visual Translation Embedding++ network (CVTransE++) model with a

recognition scheme based on feature fusion. Figure 6 presents the detailed working process of our proposed scheme. VTransE++ is the optimized relationship model, and CVTransE++ is the feature fusion-based model finally proposed in this paper. “VTransE++ semantic transformation” in CVTransE++ is the semantic transformation module in VTransE++. The motivation and design principles of the two models are described in detail below.

Figure 6. DVTransE++ is the optimized relationship model, and the Co-operative Visual Translation Embedding++ network (CVTransE++) is the feature fusion-based model finally proposed in this paper. “VTransE++ semantic transformation” in CVTransE++ is the semantic transformation module in VTransE++.

4.1. Optimized Relationship Model VTransE++

VTransE model uses the idea of semantic transformation to achieve the purpose of relationship detection in the visual field. The semantic transformation method transforms the features of the subject and the object from visual space to relational space to obtain the representation of relational predicates.

Existing semantic transformation methods can only deal with the relationship prediction task, and cannot obtain the category information of the subject and object characters. For the model to output the recognition results of the subject, relationship predicate, and object in the triplet at the same time, the network model often needs to add corresponding modules for the subject and object recognition task. This method not only increases the amount of code and parameters of the model, but also makes the network structure more complex. A similar approach is used in VTransE to handle multiple tasks simultaneously. VTransE contains two modules: target detection and relationship prediction. The identification of the subject and the object is completed in the target detection module, which is a pre-trained network model. VTransE fine-tunes the parameters of the target detection module by optimizing the relationship prediction model. Although this approach allows for an acceptable level of relational prediction, the model is always focused on a single task and ignores other relevant information that might help optimize the metric.

In the discriminatory image recognition task, the result of relationship prediction has a certain correlation with the result of character motion recognition. For example, if a person has been identified as a type of discrimination, a friendly relationship will not be formed between the two people, and if a friendly relationship has been identified, then neither of them is the category of discrimination. The main purpose of this paper is to recognize an image by predicting relationships, and the relationship prediction task is the main task. Therefore, based on a multi-task loss module, we developed the auxiliary role of the action recognition task on the relationship prediction task and proposed the VTransE++ model. The right half of Figure 6 shows the structure of VTransE++.

VTransE++ input is the visual information of any two characters in the image, which are respectively called the subject and object of the triples, and are represented as x_s and x_o . x_s and x_o are sent into the convolution module with the same structure to extract deeper visual semantic information, and the extracted information is expressed as x_s' and x_o' .

respectively. The convolution module here is the first two residual blocks in the ResNet [54]. The next stage is the semantic transformation stage, and the extracted visual features are fed to the mapping module to obtain the features of the subject and the object in the relationship space, which is expressed as $f(x'_s)$ and $f(x'_o)$. The preliminary representation of the relationship is obtained by logical operation $f(x'_o) - f(x'_s)$. Through the visual information, the action categories of the subject and the object can be directly obtained. If we transform the human features into the relationship space, it will increase the cost of feature extraction. Therefore, we use the convolution module on the visual features extracted by the subject and the object, and further extract advanced semantic features to obtain x''_s and x''_o . Then the category probability vectors y_s and y_o of the action prediction of the subject and object are obtained. In the feature reinforcement unit, the high-level semantic information of humans is integrated into the features of the relationship, and the feature representation of the relationship is obtained by $x_p = [f(x'_o) - f(x'_s), x''_s, x''_o]$. We can further attain the category probability vector y_p of relation prediction.

VTransE++ can simultaneously optimize three classification tasks, and the loss functions of the three tasks together constitute the loss function of VTransE++. The subject, object, relationship, and the final loss function of the model are described below, where t_s , t_o , and t_p are the truth vectors of the subject, object, and relationship in the triple, respectively.

$$\mathcal{L}_s = \sum_{(s,p,o) \in \mathcal{R}} -t_s^T \log(y_s) \quad (1)$$

$$\mathcal{L}_o = \sum_{(s,p,o) \in \mathcal{R}} -t_o^T \log(y_o) \quad (2)$$

$$\mathcal{L}_p = \sum_{(s,p,o) \in \mathcal{R}} -t_p^T \log(y_p) \quad (3)$$

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_o + \mathcal{L}_p \quad (4)$$

The loss functions of the VTransE+ and VTransE++ are the same. The difference is that VTransE+ has no feature enhancement unit, and the feature of the relationship is expressed as $x_p = f(x'_o) - f(x'_s)$.

4.2. Model Based on Feature Fusion CVTransE++

Optimizing the relationship model can use the image information of the characters as much as possible to improve the prediction accuracy of the relationship. However, the image information obtained by this method does not pay attention to the action details of the characters. Therefore, we proposed the idea of combining the action and relation models. Local information is critical in the image because it affects the discrimination judgment. Thus, the decision basis of the visual model is the process to find the key local information.

The interpretability of the model has attracted researchers' attention. In recent years, the method of interpreting image classification results through heat maps has become more popular. The Grad-CAM (Gradient-weighted Class Activation Mapping) [55] method expresses the category weight of image pixels in the form of a heat map. The larger the activation value of a part of the image by the neural network, the higher the weight, and the brighter the color of the part on the heat map. Figure 7 shows a heat map of an image with discriminatory characteristics. It is clear from the heat map of the discriminatory image that the key parts affecting the judgment are the person's hands and face. Our work exploits the local information, which can effectively reflect the actions of the characters, and the reasonable exploration of the local information can help improve the recognition rate of discriminatory images. In this study, we combined the relationship-based scheme with the action-based recognition scheme, exploited the local action information of the characters in the VTransE++ model, and proposed the discriminatory image recognition model CVTransE++.

Figure 7. Heat maps on an image with discriminatory characteristics.

Like VTransE++, CVTransE++ can accept both subject and object information. The difference is that the input of CVTransE++ includes the global and local information of the person, in which the global information indicates the entire person and the local information suggests the hand and face of the person. The framework of CVTransE++ is shown in the left part of Figure 6. We denote the subject's global and local information as x_{sg} and x_{ol} , respectively, and express the object's global and local information as x_{og} and x_{ol} , respectively. The semantic transformation module of VTransE++ is employed in our CVTransE++, and two semantic transformation modules are used to process global information and local information. Among these, after the local information of the characteristics (hands and faces) are converted to the same size, stitching is performed on the channel, and the first convolutional layer of the semantic conversion module for local information is changed accordingly. After the semantic transformation, the characteristics of the subject, relationship, and object under global and local information are obtained. We designate the feature of the subject, relationship, and the object under global information as x''_{sg} , x''_{pg} and x''_{og} , respectively, and denote the feature under local information as x''_{sl} , x''_{pl} and x''_{ol} , respectively. Then, through the feature fusion unit, the corresponding global and local features are fused to obtain the feature representation of the subject, relationship, and object:

$$x_s = [x''_{sg}, x''_{sl}] \quad (5)$$

$$x_p = [x''_{pg}, x''_{pl}] \quad (6)$$

$$x_o = [x''_{og}, x''_{ol}] \quad (7)$$

CVTransE++ also uses multi-task loss, and its loss function is the same as that of VTransE++.

5. Experiments

This paper focuses on discriminatory image recognition and our main work is the construction of a discriminatory image dataset and a discriminatory image recognition model. In order to verify the rationality of the dataset and the validity of the model, we designed several comparison experiments.

5.1. Experimental Configurations

Because no relevant prior research exists for discriminator image recognition, this article represents a pioneering study. The proposed CVTransE++ contains three contributions and generates two intermediate models, VTransE+ and VTransE++. In the model design stage of this work, the action recognition branch in the relational model can be regarded as a ResNet, which serves as the baseline of action recognition, and VTransE serves as the baseline of relationship prediction. The specific experimental settings are detailed as follows.

Our dataset has a total of 10,177 images. In our experiment, the annotation information of each sub-dataset was scrambled, and 4/5 of these were randomly selected as the training set, 1/2 as the validation set, and the remaining data as the test set to verify the effectiveness of data expansion in improving generalization ability. All experiments were evaluated on the test set.

We changed the size of the image to 224×224 and performed random cropping, flipping, and normalization. All of our experiments were modified based on ResNet18 [54] and VTransE. According to the structure of VTransE++ in Figure 6, we used the first two blocks to extract visual features of size $28 \times 28 \times 128$. Then, the features of the subject and the object were compressed into a 256-dimension vector and the relational representation was obtained through logical subtraction. The latter two blocks of ResNet continue feature extraction on the feature of size $28 \times 28 \times 128$, and perform pooling and linear operations to obtain a 128-dimension feature. We combined the 256-dimension relationship feature with the 128-dimension features of the subject and the object to obtain a new 512-dimension relationship feature. In CVTransE++, we extracted the characteristics of the 128-dimension subject and object, and the 512-dimension relationship characteristics through two VTransE++ semantic transformation modules. These were separately fused to obtain the final representation.

Our experiments verified seven models as follows, each of which was repeated three times. In the results, we take an average value. The learning rate was set to 0.01 and the epoch was 300.

- ResNet is a common classification network and was used as the baseline of action recognition. The input was the global information of the subject or object, and the output was its category.
- VTransE is a widely employed semantic transformation model in the visual field and was employed as the baseline of relationship prediction. The input was the global information of the subject and object, and the output was the category of their relationship.
- VTransE+ is an intermediate model, and is a multi-task loss module added to VTransE. The input was the global information of the subject and object, and the output was their category and their relationship.
- VTransE++ is an intermediate model of our work, and a feature enhancement unit is equipped in VTransE+. This model exploits the human's advanced semantic information to enhance the expression of relational features. The input was the global information of the subject and object, and the output was their category and their relationship.
- CVTransE++ is the final model proposed in this paper, and was realized by the combination of the action and relationship models based on VTransE++. The input was the global and local information of the subject and object, and the output was their category and their relationship.
- VTransE++ (4box) is a network structure that is the same as that of VTransE++. As a comparative experiment of CVTransE++, it verifies whether the improvement of CVTransE++ is only due to the increase in the amount of input information. The input of VTransE++ (4box) was the global and local information of the subject and object, and the output was their category and their relationship.

- DKRL (Description-embodied Knowledge Representation Learning) [56] is a semantic transformation model with a more complex network structure, which is applied in the field of natural language and contains four semantic transformation units. It not only processes the information of the same attribute (local or global) but also deals with the information of different attributes. In DKRL, we used the same semantic transformation module, feature fusion unit, and multi-task loss module in CVTransE++. Compared with CVTransE++, its network structure is more complex and feature representation is more abundant. In this paper, this experiment was used as a comparison experiment of CVTransE++ to verify the rationality of the feature fusion method proposed in this paper.

The above experiments were tested on the DARI dataset proposed in this paper. Due to the data expansion of DARI, DARI was divided into two parts: DARI-1 and DARI-2. Of these, DARI-1 contained only the captured image samples, and DARI-2 contained the captured samples and extended samples.

The main work of this paper is the recognition of discriminatory images, which contains three classification tasks: the action recognition task of the subject and the object, and the relationship prediction task. In this paper, the discrimination of image was judged by relationships. Therefore, relationship prediction was the main task among the three classification tasks. The sample distribution of this data set was balanced, so the performance of the model could be jointly evaluated by the classification accuracy of the three classification tasks. The accuracy rate is the ratio of the number of correctly classified samples to the total number of samples in the test sample. The formula is as follows:

$$accuracy = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(\hat{y}_i = y_i) \quad (8)$$

where $n_{samples}$ is the number of samples in the test set, y_i is the prediction result of the i^{th} sample, and $\hat{y}$ is the true value of the i^{th} sample.

5.2. Experiments Comparison

Figure 8 shows the results of different models on different tasks, and these experiments were all trained and tested on DARI-1. Table 1 records the experimental results of each model. The upper part shows the training result on DARI-1, the lower part shows the training result on DARI-2, and both parts were tested on DARI-1's test set.

Table 1. Experimental results.

Method	Subject	Object	Relationship
ResNet	62.85	64.32	-
VTransE	-	-	28.75
VTransE+	66.90	68.37	33.28
VTransE++	61.97	65.86	44.60
VTransE++(4box)	73.64	79.78	66.98
DKRL	86.26	85.33	56.61
CVTransE++	86.00	85.48	83.22
ResNet(DARI-2)	66.23	66.08	-
VTransE(DARI-2)	-	-	30.24
VTransE+(DARI-2)	67.24	68.83	35.00
VTransE+(DARI-2)	66.81	67.93	48.57
VTransE++(4box)(DARI-2)	75.11	83.32	70.27
DKRL(DARI-2)	90.67	91.18	63.79
CVTransE++(DARI-2)	90.58	91.79	88.59

Figure 8. Accuracies of DARI. (a) Class accuracies of different methods on a subset of action classes. (b) Class accuracies of different methods on a subset of relationship classes. (c) Action accuracy of the subject. (d) Action accuracy of the object. (e) Relationship accuracy.

5.2.1. Effectiveness of the Multi-Task Module

ResNet, VTransE, and VTransE+, ResNet, and VTransE are used as basic networks for action recognition and relationship prediction respectively, and VTransE+ is the network model that uses multi-task loss. The experimental effects of these models were compared.

Regarding relationship prediction, Table 1 shows that the relationship prediction accuracy of the VTransE+ model is higher than that of the VTransE model in both DARI-1 and DARI-2, and the multi-task module improves the relationship prediction accuracy by approx. 5%. In Figure 8b, the training stage of both two models is stable, and the recognition effect of VTransE+ is better than that of VTransE. In Figure 8c, the performances of VTransE and VTransE+ are not outstanding. The prediction results of VTransE tend to the “friendly” and “thumb down” categories, and we found that the number of these categories in the DARI-1 data set is higher than that of the other categories. Therefore, the VTransE model does not perform well for slightly uneven samples. Compared with the VTransE model, the VTransE+ model greatly alleviated this issue, which indicates that the multi-task loss module can improve the accuracy of relationship prediction.

Regarding action recognition, Table 1 shows that the accuracy of the action recognition of VTransE+ is higher than that of ResNet on both DARI-1 and DARI-2. From images in Figure 8a,d,e, the training effect of ResNet and VTransE+ model is stable, and the recognition effect of VTransE+ is slightly higher than that of the ResNet model, which also indicates that the multi-task loss module can improve the accuracy of action recognition. The multi-task module greatly improved the accuracy of the relationship prediction task and showed a smaller accuracy improvement for action recognition.

5.2.2. Effectiveness of the Feature Enhancement Unit

This experiment verifies whether the advanced semantic information of the person can assist the relationship prediction, and the comparative models include VTransE+ and VTransE++ models.

For the task of relationship prediction, as shown in Table 1, the recognition accuracy of VTransE++ is much higher than that of VTransE+. It can be seen from Figure 8b that both models have stable effects during the training process, and the recognition effect of VTransE++ is better than that of VTransE+. According to Figure 8c, VTransE++ can effectively identify the case of unbalanced samples, which relieves the problem that VTransE+

does not work in this case to a certain extent. Therefore, human's high-level semantic information can assist in relationship prediction.

Regarding the task of action recognition, as presented in Table 1, the recognition effect of VTransE++ is slightly lower than that of VTransE+. In Figure 8d,e, the training effect of VTransE+ and VTransE++ is stable, and the recognition accuracy is relatively similar. We believe that the reason for the slightly lower effect of VTransE++ is that the feature enhancement unit increases the influence of the relationship on the high-level semantic features of the person. As can be seen from Table 1, the accuracy of relationship prediction is generally lower than the accuracy of the action. Therefore, a higher influence of the relationship on human features reduces the accuracy of motion recognition.

The feature enhancement unit can ensure that the accuracy of action recognition is within an acceptable range, thus greatly improving the accuracy of relationship prediction. In this paper, image judgment is performed through relationships, and relationship prediction is the main task of our work.

5.2.3. Effectiveness of Feature Fusion

The experimental results of VTransE++, VTransE++(4box), and CVTransE++ were compared. Compared with VTransE++, CVTransE++ adds a feature fusion unit, which leads to different input information and model structure. The model structure of VTransE++ and VTransE++(4box) is similar, but the input information of VTransE++(4box) is the same as that of CVTransE++.

It can be seen from Table 1 and Figure 8b,c that although the recognition accuracy of VTransE++(4box) has been greatly improved compared to VTransE++, it is still much lower than that of CVTransE++. Therefore, the increase in the input information is not the key reason for the outstanding performance of CVTransE++, and feature fusion is the main reason. The three models performed stably during the training process, and CVTransE++ is outstanding in dealing with the problem of sample imbalance. In summary, feature fusion can improve the accuracy of relationship prediction. From Table 1 and Figure 8a,d,e, we can reach a similar conclusion regarding relationship prediction. This indicates that feature fusion can improve the accuracy of relationship prediction and action recognition at the same time, which demonstrates the effectiveness of feature fusion.

5.2.4. Effectiveness of Data Expansion

In the production stage of the data set, data expansion is performed by replacing the background of images to increase the sample size and scene size. We conducted the same operations on DARI-1 and DARI-2, and tested on the same test set. As can be seen from Table 1, among the three classification tasks, the recognition accuracy of each model trained on DARI-2 is generally higher than that of DARI-1. This suggests that data expansion can improve the generalization ability of the model.

5.2.5. Effectiveness of Feature Fusion Model

We also conducted an additional experiment to verify the rationality of the feature fusion method. The experimental results of CV-TransE++ and DKRL were compared. The difference between these two models is that DKRL contains four semantic transformation modules (subject global—object global, subject local—object local, subject global—object local, subject local—object global), which are more complex for the features fused by each classification task. From the perspective of relationship prediction, as shown in Table 1, the recognition accuracy of DKRL is much lower than that of CVTransE++, and the accuracy is reduced by 24.8% on both data sets on average. It can be seen from Figure 8b,c that the effect of the two models is stable in the training process, and the recognition effect of CVTransE++ is much better than that of DKRL. In the case of unbalanced samples, CVTransE++ also performs better than DKRL. To summarize, the combination of local and global information in a more complex way cannot effectively improve accuracy, but significantly decreases it.

Observing from Figure 8d,e, the training effects of CVTransE++ and DKRL are stable, and the recognition effect of DKRL is slightly higher than that of CVTransE++. When the samples are unbalanced, we can see that there is almost no difference between CVTransE++ and DKRL (see Figure 8a). Therefore, in the action recognition task, the combination of local and global information in a more complex way hardly changes the recognition effect.

To summarize, in the action recognition task, DKRL and CVTransE++ have similar effects, whereas in the relationship prediction task, the effect of DKRL is much lower than that of CVTransE++. We believe that, although the feature fusion method of DKRL is more complicated, the local information of the person cannot be obtained through feature mapping to obtain global information. That is, the global and local information cannot be replaced equivalently, the two pieces of information cannot be correctly obtained through logical operations, and the wrong relationship representation leads to a reduction in the accuracy of the relationship prediction task. Therefore, it is reasonable for us to employ the same attribute (local or global) for feature fusion.

Furthermore, to make our metric more convincing, we compared it with other metrics. As shown in Table 2, we collected the Precision, Recall, and F1 under different models. It can be seen from the table that the other three metrics are close to our metric. We also visualized the Confusion Matrix, as shown in Figure 9. Each row represents different models, and each column represents different tasks. For example, Figure 9a(1) represents the subject's recognition of actions in VTransE+ model, Figure 9a(2) represents the object's recognition of actions in VTransE+ model, and Figure 9a(3) represents the finally recognized relational predicate. It can be seen from Figure 9 that the darker the diagonal color on the graph, the better the classification effect of the model, and the sparser the matrix distribution, the worse the classification effect. With the improvement of the model, CVTransE++ and DKRL are better at the recognition of subject and object action, but CVTransE++ is better than DKRL in the recognition of relational predicate. Thus, the effectiveness of our model improvement is more obvious.

Table 2. Experimental results.

Method	Subject				Object				Relationship			
	Acc	Precision	Recall	F1	Acc	Precision	Recall	F1	Acc	Precision	Recall	F1
ResNet	62.85	-	-	-	64.32	-	-	-	-	-	-	-
VTransE	-	-	-	-	-	-	-	-	28.75	-	-	-
VTransE+	66.90	65.09	67.07	65.63	68.37	68.30	68.54	67.90	33.28	23.25	33.36	25.64
VTransE++	61.97	-	-	-	65.86	-	-	-	44.60	-	-	-
VTransE++(4box)	73.64	73.59	73.83	73.08	79.78	81.38	79.98	79.58	66.98	63.91	67.15	64.36
DKRL	86.26	87.26	86.48	86.62	85.33	88.93	88.56	88.57	56.61	56.78	56.76	55.01
CVTransE++	86.00	86.58	86.22	86.22	85.48	86.03	85.70	85.63	83.22	84.06	83.54	83.41

Figure 9. Confusion Matrix visualization. (a) VTransE+ model. (b) VTransE++ model (4box). (c) DKRL model. (d) CV-TransE++ model. (1) The first column represents the actions of the subject. (2) The second column represents the actions of the object. (3) The third column represents the relationship. It can be seen from the figure that, with the improvement of the model, the corresponding classification performance is improved.

6. Conclusions

In this paper, we propose using an action model and a relationship model to implement discrimination emotion detection of images. A discrimination image benchmark dataset was first established and a cooperative VTransE++ model was developed to precisely recognize the interaction between persons by fusing the action and relationship

models for discriminatory image detection. In contrast to the traditional action recognition problem, we explored the semantic relationship of human–human interaction in addition to specific action recognition. Results from experiments on the proposed benchmark verified the effectiveness of our optimized relationship model, and the fusion of the action and relationship models. We will release our benchmark dataset and continue to establish multi-person discrimination datasets in the future. We believe that our dataset and experiments can encourage more research on image discrimination detection.

Author Contributions: Conceptualization, B.Z.; Data curation, Z.W. and Y.L.; Investigation, Z.W. and T.Z.; Methodology, B.Z. and Y.L.; Project administration, B.Z.; Visualization, Y.L.; Writing—review and editing, J.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Science Foundation of China, grant number 61972027, 61872035.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jun-Ho, H.; Yeong-seok, S. Understanding Edge Computing: Engineering Evolution with Artificial Intelligence. *IEEE Access* **2019**, *7*, 164229–164245.
2. Ying, Y.; Jun-Ho, H. Customized CAD Modeling and Design of Production Process for One-person One-clothing Mass Production System. *Electronics* **2018**, *7*, 270.
3. Brian, D.; Martin, R. Automatic Detection and Repair of Errors in Data Structures. *ACM Sigplan Not.* **2003**, *38*, 78–95.
4. Nigel, B.; Sidney, D.; Ryan, B.; Jaclyn, O.; Valerie, S.; Matthew, V.; Lubin, W.; Weinan, Z. Automatic Detection of Learning-Centered Affective States in the Wild. In Proceedings of the 20th International Conference on Intelligent User Interfaces, Atlanta, GA, USA, 29 March–1 April 2015.
5. Zakia, H.; Jeffrey, F. Automatic Detection of Pain Intensity. In Proceedings of the 14th ACM International Conference on Multimodal Interaction, Santa Monica, CA, USA, 22–26 October 2012.
6. Arman, S.; Bulent, S.; Taha, M.B. Comparative Evaluation of 3D vs. 2D Modality for Automatic Detection of Facial Action Units. *Pattern Recognit.* **2012**, *45*, 767–782.
7. Lee, H.; Park, S.H.; Yoo, J.H.; Jung, S.H.; Huh, J.H. Face Recognition at a Distance for a Stand-alone Access Control System. *Sensors* **2020**, *20*, 785. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Neziha, J.; Francisco, J.P.; Jose, M.B.; Nouredine, B.; Med, S.B. Prediction of Human Activities Based on a New Structure of Skeleton Features and Deep Learning Model. *Sensors* **2020**, *20*, 4944.
9. Hoofnagle, C.; King, J.; Li, S. How Different are Young Adults from Older Adults When it Comes to Information Privacy Attitudes and Policies? *SSRN Electron. J.* **2010**. [\[CrossRef\]](#)
10. Manzo, M.; Pellino, S. Bucket of Deep Transfer Learning Features and Classification Models for Melanoma Detection. *J. Imaging* **2020**, *6*, 129. [\[CrossRef\]](#)
11. Lyons, M.; Akamatsu, S.; Kamachi, M. Coding Facial Expressions with Gabor Wavelets. In Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998.
12. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The Extended Cohnkanade Dataset (ck+): A Complete Dataset for Action Unit and Emotion-specified Expression. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, San Francisco, CA, USA, 13–18 June 2010.
13. Maas, A.L.; Daly, R.E.; Pham, P.T.; Huang, D.; Ng, A.Y.; Potts, C. Learning Word Vectors for Sentiment Analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, OR, USA, 19–24 June 2011.
14. Strapparava, C.; Mihalcea, R. Semeval-2007 task 14: Affective Text. In Proceedings of the Fourth International Workshop on Semantic Evaluations, Prague, Czech Republic, 23–24 June 2007.
15. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [\[CrossRef\]](#)
16. Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. Deap: A Database for Emotion Analysis Using Physiological Signals. *IEEE Trans. Affect. Comput.* **2011**, *3*, 18–31. [\[CrossRef\]](#)
17. Yuan, S.; Wu, X.; Xiang, Y. Task-specific Word Identification from Short Texts Using A Convolutional Neural Network. *Intell. Data Anal.* **2018**, *22*, 533–550. [\[CrossRef\]](#)
18. Paula, F.; Sergio, N. A Survey on Automatic Detection of Hate Speech in Text. *ACM Comput. Surv.* **2018**, *51*, 1–30.
19. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.J.; Shamma, D.A. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vis.* **2017**, *123*, 32–73. [\[CrossRef\]](#)
20. Lu, C.; Krishna, R.; Bernstein, M.; Fei-Fei, L. Visual Relationship Detection with Language Priors. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016.

21. Yu, X.; Zhang, Z.; Wu, L.; Pang, W.; Chen, H.; Yu, Z.; Li, B. Deep Ensemble Learning for Human Action Recognition in Still Images. *Complexity* **2020**. [[CrossRef](#)]
22. Qi, T.; Xu, Y.; Quan, Y.; Wang, Y.; Ling, H. Image-based Action Recognition Using Hint-enhanced Deep Neural Networks. *Neurocomputing* **2017**, *267*, 475–488. [[CrossRef](#)]
23. Yao, B.; Jiang, X.; Khosla, A.; Lin, A.L.; Guibas, L.; Li, F.-F. Human Action Recognition by Learning Bases of Action Attributes and Parts. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011.
24. Yao, B.; Li, F.-F. Grouplet: A Structured Image Representation for Recognizing Human and Object Interactions. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010.
25. Vu, T.H.; Olsson, C.; Laptev, I.; Oliva, A.; Sivic, J. Predicting Actions from Static Scenes. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
26. Ma, W.; Liang, S. Human-Object Relation Network for Action Recognition in Still Images. In Proceedings of the IEEE International Conference on Multimedia and Expo, London, UK, 6–10 July 2020.
27. Delaitre, V.; Laptev, I.; Sivic, J. Recognizing Human Actions in Still Images: A Study of Bag-of-Features and Part-based Representations. In Proceedings of the British Machine Vision Conference, Aberystwyth, Wales, UK, 30 August–2 September 2010.
28. Zhao, Z.; Ma, H.; You, S. Single Image Action Recognition Using Semantic Body Part Actions. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
29. Krishna, R.; Chami, I.; Bernstein, M.; Fei-Fei, L. Referring Relationships. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
30. Lu, P.; Ji, L.; Zhang, W.; Duan, N.; Zhou, M.; Wang, J. R-VQA: Learning Visual Relation Facts with Semantic Attention for Visual Question Answering. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018.
31. Johnson, J.; Gupta, A.; Fei-Fei, L. Image Generation from Scene Graphs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
32. Qi, M.; Li, W.; Yang, Z.; Wang, Y.; Luo, J. Attentive Relational Networks for Mapping Images to Scene Graphs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
33. Liu, X.; Liu, W.; Zhang, M.; Chen, J.; Gao, L.; Yan, C.; Mei, T. Social Relation Recognition from Videos via Multi-scale Spatial-Temporal Reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
34. Xu, B.; Wong, Y.; Li, J.; Zhao, Q.; Kankanhalli, M.S. Learning to Detect Human-Object Interactions with Knowledge. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
35. Goel, A.; Ma, K.T.; Tan, C. An End-to-End Network for Generating Social Relationship Graphs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
36. Shi, J.; Zhang, H.; Li, J. Explainable and Explicit Visual Reasoning over Scene Graphs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
37. Xu, D.; Zhu, Y.; Choy, C.B.; Fei-Fei, L. Scene Graph Generation by Iterative Message Passing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 22–25 July 2017.
38. Bin, Y.; Yang, Y.; Tao, C.; Huang, Z.; Li, J.; Shen, H.T. MR-NET: Exploiting Mutual Relation for Visual Relationship Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
39. Zhang, H.; Kyaw, Z.; Chang, S.F.; Chua, T.S. Visual Translation Embedding Network for Visual Relation Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 22–25 July 2017.
40. Wan, H.; Luo, Y.; Peng, B.; Zheng, W.S. Representation Learning for Scene Graph Completion via Jointly Structural and Visual Embedding. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018.
41. Hung, Z.S.; Mallya, A.; Lazebnik, S. Contextual Visual Translation Embedding for Visual Relationship Detection and Scene Graph Generation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**. [[CrossRef](#)] [[PubMed](#)]
42. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [[CrossRef](#)] [[PubMed](#)]
43. Wang, Z.; Zhang, J.; Feng, J.; Chen, Z. Knowledge Graph Embedding by Translating on Hyperplanes. In Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, Québec City, QC, Canada, 27–31 July 2014.
44. Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; Zhu, X. Learning Entity and Relation Embeddings for Knowledge Graph Completion. In Proceedings of the Twenty-ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.
45. Xiao, H.; Huang, M.; Hao, Y.; Zhu, X. TransA: An Adaptive Approach for Knowledge Graph Embedding. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.
46. Ji, G.; Liu, K.; He, S.; Zhao, J. Knowledge Graph Completion with Adaptive Sparse Transfer Matrix. In Proceedings of the Thirtieth AAAI conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
47. Xiao, H.; Huang, M.; Zhu, X. TransG: A Generative Model for Knowledge Graph Embedding. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016.
48. Nanay, B. Comment: Every Action is an Emotional Action. *Emot. Rev.* **2017**, *9*, 350–352. [[CrossRef](#)]

49. Yin, J. Body Language Classification and Communicative Context. In Proceedings of the International Conference on Education, Language, Art and Intercultural Communication, Zhengzhou, China, 5–7 May 2014.
50. Wikipedia. Gesture. Available online: <https://en.wikipedia.org/wiki/Gesture> (accessed on 10 December 2019).
51. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-cnn: Towards Real-time Object Detection with Region Proposal Networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QB, Canada, 7–12 December 2015.
52. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
53. Aksoy, Y.; Oh, T.H.; Paris, S.; Pollefeys, M.; Matusik, W. Semantic Soft Segmentation. *ACM Trans. Graph. (TOG)* **2018**, *37*, 1–13. [[CrossRef](#)]
54. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
55. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual Explanations from Deep Networks via Gradient-based Localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
56. Fan, M.; Zhou, Q.; Zheng, T.F.; Grishman, R. Distributed Representation Learning for Knowledge Graphs with Entity Descriptions. *Pattern Recognit. Lett.* **2017**, *93*, 31–37. [[CrossRef](#)]