Exploiting Multi-domain Visual Information for Fake News Detection

Peng Qi[1,2], Juan Cao[1,2], Tianyun Yang[1,2], Junbo Guo[1] and Jintao Li[1]

[1]Key Laboratory of Intelligent Information Processing & Center for Advanced Computing Research,
Institute of Computing Technology, CAS
[2]University of Chinese Academy of Sciences
Beijing, China
{qipeng,caojuan,yangtianyun,guojunbo,jtli}@ict.ac.cn

Abstract—The increasing popularity of social media promotes the proliferation of fake news. With the development of multimedia technology, fake news attempts to utilize multimedia content with images or videos to attract and mislead readers for rapid dissemination, which makes visual content an important part of fake news. Fake-news images, images attached to fake news posts, include not only fake images that are maliciously tampered but also real images that are wrongly used to represent irrelevant events. Hence, how to fully exploit the inherent characteristics of fake-news images is an important but challenging problem for fake news detection. In the real world, fake-news images may have significantly different characteristics from real-news images at both physical and semantic levels, which can be clearly reflected in the frequency and pixel domain, respectively. Therefore, we propose a novel framework <u>M</u>ulti-domain <u>V</u>isual <u>N</u>eural <u>N</u>etwork (MVNN) to fuse the visual information of frequency and pixel domains for detecting fake news. Specifically, we design a CNN-based network to automatically capture the complex patterns of fake-news images in the frequency domain; and utilize a multi-branch CNN-RNN model to extract visual features from different semantic levels in the pixel domain. An attention mechanism is utilized to fuse the feature representations of frequency and pixel domains dynamically. Extensive experiments conducted on a real-world dataset demonstrate that MVNN outperforms existing methods with at least 9.2% in accuracy, and can help improve the performance of multi-modal fake news detection by over 5.2%.

Index Terms—fake news detection, fake-news images, multi-domain, social media

I. Introduction

Social media has become a major information platform, where people acquire the latest news and express their opinions freely. However, the convenience and openness of social media have also promoted the proliferation of *fake news*, which caused many detrimental societal effects. For instance, during the month before the 2016 U.S. presidential election campaign, the Americans encountered between one and three fake stories on average from known publishers [1], which inevitably misled the voters and influenced the election results. Hence, the automatic detection of fake news has become an urgent problem of great concern in recent years [2]–[5].

The development of multimedia technology promotes the evolution of self-media news from text-based posts to multimedia posts with images or videos, which provide better storytelling and attract more attention from readers. It is

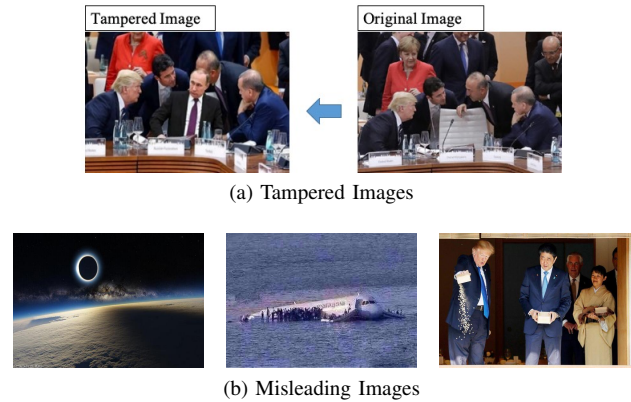(a) Tampered Images

(b) Misleading Images

Fig. 1: Examples of fake-news images: (a) A tampered image where Putin is spliced on the middle seat at G-20 to show that he is in the middle of an intense discussion between other world leaders. (b) Three misleading images: the left one is an artwork which is reposted as a photo of solar eclipse (March 2015); the middle one is a real image captured in 2009 New York plane crash, but it is claimed to be the wrecked Malaysia Airlines MH370 in 2014; the right one is a real image used to misinterpret Trump's behavior while he is just imitating Abe.

estimated that the average number of reposts for posts with images is 11 times larger than those without images [6]. Unfortunately, this advantage is also taken by fake news which usually contains misrepresented or even tampered images to attract and mislead readers for rapid dissemination. As a result, visual content has become an important part of fake news that cannot be neglected.

According to existing research [7], ***fake-news images***, images attached to fake news posts, include not only fake images that are maliciously tampered but also real images that are wrongly used to represent irrelevant events. In detail, we broadly classify fake-news images into two categories: tampered images and misleading images. ***Tampered images*** mean fake-news images which have been digitally modified as Fig. 1a shows, equal to *fake images* in our common sense. ***Misleading images*** refer to fake-news images that haven't experienced any manipulation, but the content is misleading as described in Fig. 1b. These misleading images usually derive
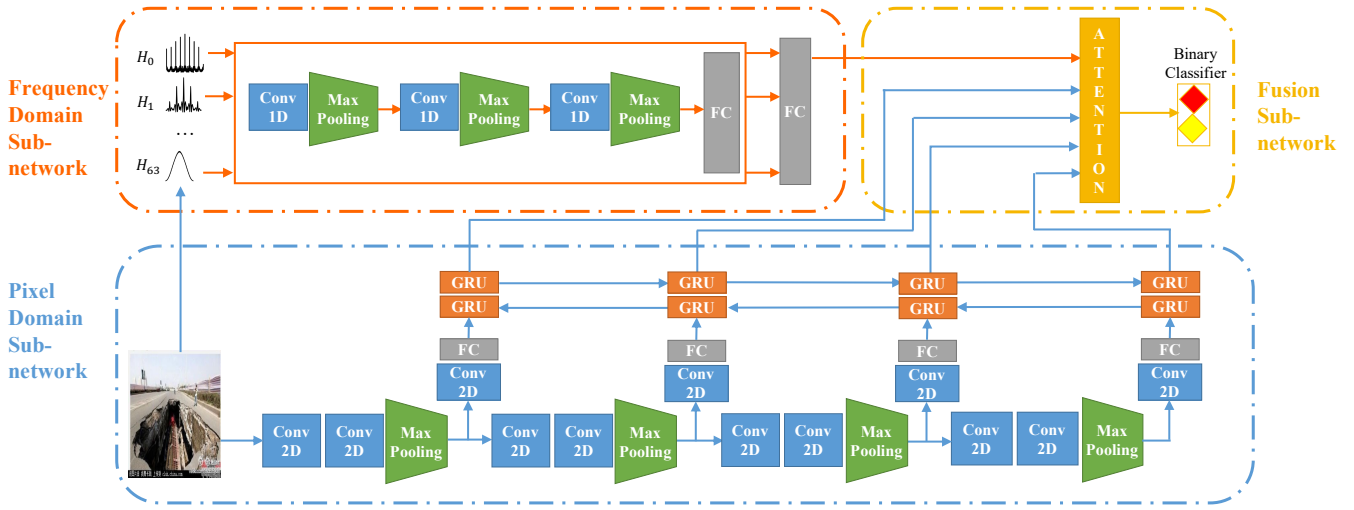
Fig. 2: An illustration of the proposed Multi-domain Visual Neural Network (MVNN). It mainly consists of three components: a frequency domain sub-network, a pixel domain sub-network, and a fusion sub-network. The frequency domain sub-network first transforms the input image from the pixel domain to the frequency domain and utilizes a CNN-based model to capture the physical characteristics of this image. The pixel domain sub-network employs a multi-branch CNN-RNN network to extract the features of different semantic levels of the input image. The fusion sub-network dynamically fuses feature vectors obtained from the frequency and pixel domain sub-network through an attention mechanism to classify the input image as a fake-news or real-news image.

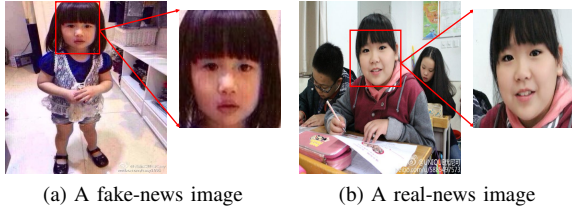(a) A fake-news image   (b) A real-news image

Fig. 3: Comparison of fake-news images and real-news images at the physical level. We observe that the fake-news image (a) has an obvious block effect while the real-news image (b) is more clear. We enlarge the faces of the two images for better comparison.

(a)   (b)   (c)

Fig. 4: Comparison of fake-news images and real-news images at the semantic level. We can find that fake-news images are more visually striking and emotional provocative than real-news images, even though they describe the same type of event such as fire (a), earthquake (b) and road collapse (c).

from artworks or outdated images which are published on an early event [8]. Existing researches on fake news detection mainly focus on text content [9]–[12] and social context [13], [14], while a few works utilize visual information to detect fake news. Some works [7], [15] extract forensics features to evaluate the authority of the attached images, but these forensics features are mostly crafted manually for detecting specific manipulated traces, which are not applicable for misleading images. Other works [16]–[18] utilize a pre-trained convolutional neural network like VGG19 [19] to obtain general visual representations, which can hardly capture the semantic commonalities of fake-news images due to the lack of task-relevant information. Therefore, how to fully exploit the inherent characteristics of fake-news images is an important but challenging problem for distinguishing fake news from real news by utilizing visual contents.

In the real world, fake-news images may have significantly different characteristics from real-news images at both physical and semantic levels:

At the physical level, fake-news images might be of low quality, which can be clearly reflected in the **frequency domain**. For example, after being uploaded to and downloaded from social platforms multiple times, misleading images usually have heavier re-compression artifacts than real-news images such as a block effect as shown in Fig. 3. Besides, there are manipulation traces in tampered images inevitably. Considering that re-compressed and tampered images often present periodicity in the frequency domain, which can be

easily characterized by CNN which has the ability to capture spatial structure features, we design a CNN-based network to automatically capture the characteristics of fake-news images in the frequency domain (see the top part of Fig. 2).

At the semantic level, fake-news images also exhibit some distinct characteristics in the **pixel domain** (also known as spatial domain). Fake news publishers tend to utilize images to attract and mislead readers for rapid dissemination; thus fake-news images often show visual impacts [20] and emotional provocations [2], [21] just as Fig. 4 shows. These characteristics have been proven related to many visual factors from low-level to high-level [22]; thus we build a multi-branch CNN-RNN network to extract features of different semantic levels (see the bottom part of Fig. 2), for fully capturing the characteristics of fake-news images in the pixel domain.

We have shown that the visual features at both physical and semantic levels are important for detecting fake-news images; thus fusing the visual information of the frequency and pixel domains has the potential to improve the performance of fake news detection. Intuitively, not all features contribute equally to the task of fake news detection, which means some visual features play more important roles than others in evaluating whether a given image is a fake-news or real-news image. Therefore, we employ an attention mechanism [23] to fuse these visual features from different domains dynamically.

To sum up, we propose a <u>M</u>ulti-domain <u>V</u>isual <u>N</u>eural <u>N</u>etwork (MVNN) framework (see Fig. 2), which can learn effective visual representations by combining the information of frequency and pixel domains for fake news detection. The proposed model consists of three main components: a frequency domain sub-network and a pixel domain sub-network used to capture the characteristics of fake-news images at physical and semantic levels respectively, and a fusion sub-network to fuse these features dynamically. In summary, the contributions of this paper are three folds:

1) To the best of our knowledge, we are the first to utilize multi-domain visual information for fake news detection, which captures the inherent characteristics of fake-news images at both physical and semantic levels.
2) We propose a novel framework MVNN, which exploits an end-to-end neural network to learn representations of frequency and pixel domains simultaneously and effectively fuse them.
3) We conduct extensive experiments on a real-world dataset to validate the effectiveness of the proposed model. The results demonstrate that our model is much better than existing methods, and the visual representations learned by our model can help improve the performance of multi-modal fake news detection by a large margin. Besides, we show that the information of frequency and pixel domains are complementary for detecting fake news.

II. Related Work

In the task of fake news detection, the main challenge is how to utilize information from different modalities to distinguish fake news posts and real news posts. Most of existing approaches focus on text content [9]–[12] and social context [13], [14] which refers to information generated during the news propagation process on social networks. Recently, visual information has been shown to be an important indicator for fake news detection [2], [6]. With the popularity of multimedia content, researchers begin to incorporate visual information to detect fake news.

Some early works use basic statistical features about the attached images to help detect fake news posts, such as the number of attached images [10], [24], image popularity and image type [6]. However, these statistical features are so basic that they can hardly represent complex distributions of visual contents in fake news.

Visual forensics features are generally used for image manipulation detection. To evaluate the authority of attached images, some works extract visual forensics features, such as block artifact grids (BAG), to assist fake news detection. For example, the Verifying Multimedia Use task on 2015 [25] and 2016 MediaEval benchmark [7] provide seven types of image forensics features to help detect manipulated and misleading use of web multimedia content. Based on these provided forensics features, [15] extract advanced forensics features and combine them with post-based and user-based features in a semi-supervised learning scheme to tackle the news verification problem. However, most of these forensics features are crafted manually for detecting specific manipulated traces, which are not applicable to the real images attached to fake news. Besides, these hand-crafted features are labor-expensive and limited to learn complicated patterns, causing the poor generalization performance on the task of fake news detection.

Inspired by the power of CNN, most existing works based on multimedia contents use a pre-trained deep CNN like VGG19 [19] to obtain general visual representations and fuse them with text information. Specifically, [16] first incorporates multi-modal contents on social networks via deep neural networks to solve fake news detection problem; [17] proposes an end-to-end event adversarial neural network to detect newly-emerged fake news events based on multi-modal features; [18] presents a novel approach to learn a shared representation of multi-modal information for fake news detection. However, these works focus on how to fuse the information of different modalities, ignoring effectively modeling visual contents. These visual features they adopted are too general to reflect the intrinsic characteristics of fake-news images due to the lack of task-relevant information, which degrades the performance of visual contents in fake news detection.

To overcome these limitations of existing works, we propose a novel deep learning network to model visual contents at both physical and semantic levels for the task of fake news detection.

III. Problem Formulation

Fake news is defined as news articles that are intentionally and verifiably false and could mislead readers [26], which has been widely adopted in recent studies [2]. Different from the

traditional definition, in the context of microblog, fake news aims at news posts that are published on social media by users which are usually less than 140 characters instead of news articles. Formally, we state this definition as follows,

*Definition 1: **Fake news:** In the context of microblog, a piece of fake news is a news post that is intentionally and verifiably false.*

*Definition 2: **Fake-news images:** A fake-news image is an image attached to fake news.*

The problem addressed in this paper is how to utilize the visual content to identify a news post as real or fake, which equals to classify a given image as a fake-news image or real-news image. Consequently, we formally define the studied problem as follows:

***Problem 1:** Given a set of news posts $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$, corresponding images $\mathcal{I} = \{i_1, i_2, \ldots, i_m\}$, and labels $\mathcal{Y} = \{y_1, y_2, \ldots, y_m\}$, learn a classifier f that can utilize the corresponding image to classify whether a given post is fake news ($y_t = 1$) or real news ($y_t = 0$) , i.e., $\hat{y}_t = f(i_t)$.*

IV. METHODOLOGY

In this section, we first make an overview of the proposed Multi-domain Visual Neural Network (MVNN) and then introduce the three major components in detail.

A. Model Overview

The goal of the proposed MVNN model is utilizing the visual information in frequency and pixel domains to evaluate whether the given image is a fake-news or real-news image. As shown in Fig. 2, MVNN includes three major components: a frequency domain sub-network, a pixel domain sub-network and a fusion sub-network. For an input image, we feed it into the frequency and pixel domain sub-network to obtain features at physical and semantic levels respectively. The fusion sub-network takes these learned features as input and learns a final visual representation of this image to predict whether this image is a fake-news or real-news image.

B. Frequency Domain Sub-network

The physical features of images can reflect their originality to a certain extent which is helpful for detecting fake news; thus we design the frequency domain sub-network to extract features of the input image at the physical level. According to existing studies about image forensics [27]–[30], discrete cosine transform (DCT) has been widely utilized for capturing the tampered and re-compressed architects, so we employ DCT on the input image to transform it from pixel domain to frequency domain. Considering that tampered images and re-compressed images often present periodicity in the frequency domain, which can be easily characterized by CNN with the ability to capture spatial structure characteristics; thus we design a CNN-based network to capture the characteristics of fake-news images in the frequency domain. The detailed architecture of this part is shown in Fig. 5.

For an input image, we first employ block DCT on it to obtain 64 histograms of DCT coefficients corresponding to 64
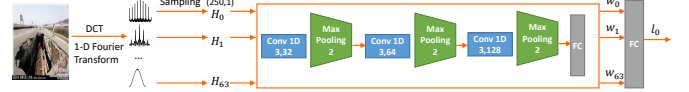

Fig. 5: The detailed architecture of the frequency domain sub-network. For an input image, we first transform it from the pixel domain to the frequency domain, and design a CNN-based network to obtain its feature representation in the frequency domain.

frequencies. Following the process of [28], we then carry 1-D Fourier transform on these DCT coefficient histograms to enhance the effect of CNN. Considering that CNN needs an input of a fixed size, we sample these histograms and obtain 64 250-dimensional vectors, which can be represented as $\{H_0, H_1, \ldots H_{63}\}$. After being pre-processed, each input vector is fed to the shared CNN network to obtain corresponding feature representation $\{w_0, w_1, \ldots w_{63}\}$. The CNN network consists of three convolutional blocks and a fully connected layer, and each convolutional block is composed of a one-dimensional convolutional layer followed by a max-pooling layer. To accelerate the convergence of the model, we set the number of filters in the convolution layers to be incremental. Existing works about image forensics usually only consider coefficients of a part of frequencies. However, we find that all frequencies contribute to the task of fake news detection, hence we fuse the feature vectors of all frequencies by concatenating and feed the final feature representation l_0 to the fusion sub-network. Specifically, we experiment with different fusing methods, and results show that concatenating performs best in this task.

C. Pixel Domain Sub-network

We have shown that fake-news images have different characteristics from real-news images at the semantic level, indicating that the semantic features are important in detecting fake news. Therefore, we design the pixel domain sub-network for extracting visual features of the input image at the semantic level, of which the detailed architecture is shown in Fig. 6.

CNN learns high-level semantic representations through layer-by-layer abstraction from local to the global view, of which the earlier layers prefer low-level features such as
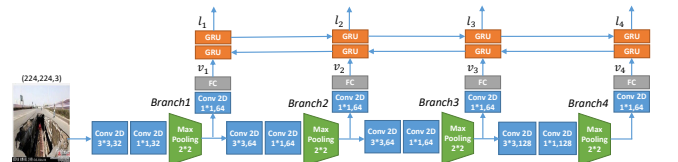

Fig. 6: The detailed architecture of the pixel domain sub-network. For an input image, we utilize a multi-branch CNN-RNN network to extract its features of different semantic levels in the pixel domain.

color, line and shape, while the later layers tend to focus on the objects. In the process of abstraction, low-level features will inevitably suffer some losses, which further suggests that the bottom and intermediate layers in CNN can provide complementary information for the top layers. Many works have proven integrating features from different layers can help achieve better performance than only using high-level features for some tasks such as salient object detection and image emotion classification [31]–[33]. We have shown that fake-news images usually show some visual impacts and emotion provocations, which have been proven related to many visual factors from low-level to high-level [22]. Therefore, to fully capture the semantic characteristics of fake-news images, we build a multi-branch CNN network to extract multiple levels of features and utilize a bidirectional GRU (Bi-GRU) network to model the sequential dependencies between these features.

As Fig. 6 described, the CNN network mainly consists of four blocks composed of a 3×3 and 1×1 convolution layer and a max-pooling layer. The input image is fed to the CNN, and features extracted from four branches will pass through a 1×1 convolutional layer and a fully connected layer, obtaining the corresponding feature vectors $v_t, t \in [1, 4]$. These features represent different parts of images, such as line, color, texture, and object, which characterize different levels of features from local to global view. Motivated by the Inception module used in GoogLeNet [34], we use 1×1 convolution layer to reduce the dimension and increase the model's representative power because it increases the non-linear activation and promotes the information fusion of different channels.

Intuitively, there are strong dependencies between different levels of features. For example, middle-level features such as textures, consist of low-level features such as lines, and meanwhile compose high-level features such as objects. Therefore, we utilize GRU to model these dependencies between low-level and high-level features. Specifically, we model these features from different levels as a sequence $V = \{v_t\}, t \in [1, 4]$, where v_t represents the visual features extracted from t-th branch in the CNN network described in Fig. 6. Then the entire pipeline in GRU at time step t can be presented as follows:

$$r_t = \sigma(W_r[v_t, h_{t-1}] + b_r) \tag{1}$$

$$z_t = \sigma(W_z[v_t, h_{t-1}] + b_z) \tag{2}$$

$$\tilde{h}_t = \tanh(W_{\tilde{h}}[v_t, r_t \odot h_{t-1}] + b_{\tilde{h}}) \tag{3}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{4}$$

where $r_t, z_t, \tilde{h}_t, h_t$ are the reset gate, update gate, hidden candidate, and hidden state respectively. W are the weight matrices and b are the bias terms. Besides, σ stands for the sigmoid function and $\odot$ represents the element-wise multiplication.

Considering that the dependencies among different levels of features can be estimated from both local to global view and global to local view; we utilize the bidirectional GRU to model the relations from two different views. The Bi-GRU

contains the forward GRU $\overrightarrow{f}$ which reads from v_1 to v_4, and backward GRU $\overleftarrow{f}$ which reads from v_4 to v_1:

$$\overrightarrow{l_t} = \overrightarrow{GRU}(v_t), t \in [1, 4] \tag{5}$$

$$\overleftarrow{l_t} = \overleftarrow{GRU}(v_t), t \in [1, 4] \tag{6}$$

For each timestep t, we obtain the hidden representations by concatenating the forward hidden state $\overrightarrow{l_t}$ and the backward hidden state $\overleftarrow{l_t}$, i.e., $l_t = [\overrightarrow{l_t}, \overleftarrow{l_t}]$, which compose the final semantic feature representation $L = \{l_t\}, t \in [1, 4]$.

D. Fusion Sub-network

We assume that the physical and semantic features of images are complementary in detecting fake news; thus we propose the fusion sub-network to fuse these features. Specifically, we utilize the output vector of the frequency domain sub-network l_0 and pixel domain sub-network $\{l_1, l_2, l_3, l_4\}$ to make predictions. $\{l_1, l_2, l_3, l_4\}$ represent visual features from different semantic levels while l_0 represents visual feature from the physical level. Intuitively, not all features contribute equally to the task of fake news detection, which means some visual features play a more important role than others in evaluating whether a given image is a fake-news or real-news image. For instance, for some tampered images which present obvious tampering traces, the physical features perform better than semantic features; for some misleading images which have not experienced severe re-compressions, the semantic features are more effective. Therefore, we highlight those valuable features via attention mechanism, and the enhanced image representation is computed as follows:

$$\mathcal{F}(l_i) = v^T \tanh(W_f l_i + b_f), i \in [0, 4] \tag{7}$$

$$\alpha_i = \frac{\exp(\mathcal{F}(l_i))}{\sum_i \exp(\mathcal{F}(l_i))} \tag{8}$$

$$u = \sum_i \alpha_i l_i \tag{9}$$

where W_f denotes the weight matrix and b_f is the bias term, v^T denotes a transposed weight vector and $\mathcal{F}$ is the score function which measures the significance of each feature vector. Afterward, we obtain the normalized weight of i-th feature vector α_i via a softmax function and compute the advanced representation of the input image as a weighted sum of different feature vectors. The vector v is randomly initialized and jointly learned during the training process.

Till now, we have obtained a high-level representation u of the input image, which models the characteristics of this image at both physical and semantic levels. We use a fully connected layer with softmax activation to project this vector into the target space of two classes: fake-news image and real-news image, and gain the probability distributions:

$$p = \text{softmax}(W_c u + b_c) \tag{10}$$

In the proposed model MVNN, we define the loss function as cross-entropy error between the predicted probability distribution and the ground truth:

$$L = -\sum \left[y \log p + (1-y) \log (1-p) \right] \quad (11)$$

where y is the ground truth with 1 representing fake-news image and 0 representing real-news image and p denotes the predicted probability of being fake-news image.

V. EXPERIMENTS

In this section, we conduct experiments on a real-world dataset to evaluate the effectiveness of the proposed model MVNN. Specifically, we aim to answer the following evaluation questions:

- **EQ1**: Is MVNN able to improve the performance of fake news detection based on visual modality?
- **EQ2**: How effective are different domains and other network components: attention, Bi-GRU and branches in the pixel domain sub-network, in improving the performance of MVNN?
- **EQ3**: Can MVNN help improve the performance of multi-modal fake news detection?

We first introduce the dataset we used and some representative methods which model the visual contents for fake news detection as baselines; next, we describe the implementation details of the proposed MVNN model. And then we compare MVNN with those baselines to answer **EQ1**, and we investigate the ablation study to answer **EQ2**, which includes quantitative and qualitative analyses. To answer **EQ3**, we further compare MVNN with baselines on multi-modal fake news detection. Last, we study some typical cases to vividly illustrate the importance of multiple domains for fake news detection.

A. Dataset

Considering that fake news detection based on multimedia content is a pretty new task, there are a few standard multimedia fake news datasets available. The two most widely used datasets are the Twitter dataset presented in the MediaEval Verifying Multimedia Use benchmark [7], and Weibo dataset built in [16]. However, there are a lot of duplicate images in the Twitter dataset, causing the number of distinctive images less than 500, which makes the Twitter dataset too small to support the training of the proposed model. Therefore, in this paper, we conduct our experiments on the Weibo dataset alone to evaluate the effectiveness of the proposed model.

In Weibo dataset, the fake news posts are crawled from May 2012 to January 2016 and verified by the official rumor debunking system of Weibo[1], which actually serves as a reputable source to collect fake news posts in literature. The real news posts are collected during the same period as the fake news from Weibo and are verified by Xinhua News Agency, an authoritative news agency in China. According to [16], duplicated and very small images have been removed for ensuring the quality of the entire dataset. To ensure that each

[1]https://service.account.weibo.com/

post corresponds to an image, text-only posts are removed and only one image is saved for posts with multiple illustrations. In total, this dataset includes 4749 fake news posts and 4779 real news posts with corresponding images. In our experiment, we first use the K-means algorithm to cluster all news posts into 200 clusters, and split the whole dataset into the training set, validation set and testing set to ensure that there is no event overlap between these sets, which further prevents the model from overfitting on event topics. The training, validation and testing sets contain a number of posts approximately with a ratio of 7:1:2 as in [17].

B. Baselines

To validate the effectiveness of MVNN, we choose several representative methods which are used to model the visual contents for fake news detection as baselines:

- **Forensics Features (FF)+LR**: Reference [15] proposes new image forensics features and evaluates the effectiveness of these features in detecting fake news. We adopt the logistic regression (LR) algorithm to make use of these features to make classifications.
- **Pre-trained VGG**: Pre-trained VGG is widely used as a feature extractor in existing studies about multi-modal fake news detection [16]–[18]. Specifically, these works use the output of the last layer of a 19-layer VGGNet pre-trained on ImageNet, of which the feature dimension is 4096. Similarly, we use LR to classify these features.
- **Fine-tuned VGG**: It has become a common practice that fine-tuning a pre-trained model on a task-relevant dataset for achieving a promising performance on specific tasks. Hence, in addition to Pre-trained VGG, we also use Fine-tuned VGG as a compared experiment. In particular, we use the training data to fine-tune pre-trained VGG19, and use the same method as Pre-trained VGG to extract features and make classifications.
- **ConvAE**: AutoEncoder (AE) is a kind of artificial neural network used to learn efficient data codings in an unsupervised manner. Convolutional AutoEncoder (ConvAE) [35] extends the AE framework through utilizing the convolutional layer to compose the encoder and decoder for obtaining a better understanding of images than plain fully connected layers. Considering that only a few methods are used to model the visual contents for fake news detection, we use ConvAE as an additional comparison method to show the full advantage of the proposed model. To avoid any bias, we design the encoder to be the same structure as the pixel domain sub-network without branches. After pre-training the ConvAE, we use the encoder to extract features and use LR to make classifications.

C. Implementation Details

In this section, we introduce the implementation details of the proposed model for reproducibility, including the parameter setting and the training strategy. In the frequency domain sub-network, the two fully connected layers contain 16 and

TABLE I: The performance comparison of fake news detection in single visual modality.

Method	Accuracy	Precision	Recall	F1
FF+LR	0.650	0.612	0.579	0.595
Pre-trained VGG	0.721	0.669	0.738	0.702
Fine-tuned VGG	0.754	0.74	0.689	0.714
ConvAE	0.734	0.685	0.744	0.713
MVNN	**0.846**	**0.809**	**0.857**	**0.832**

TABLE II: Ablation study of MVNN.

Method	Accuracy	Precision	Recall	F1
MVNN	**0.846**	**0.809**	**0.857**	**0.832**
w/o frequency domain	0.794	0.792	0.728	0.758
w/o pixel domain	0.737	0.698	0.717	0.708
w/o attention	0.827	0.778	0.853	0.814
w/o Bi-GRU	0.828	0.772	0.841	0.805
w/o branches	0.803	0.752	0.830	0.789

64 neurons respectively, and we add a dropout layer with a rate of 0.4 after the last fully connected layer to avoid overfitting. In the pixel domain sub-network, the number of hidden units of the GRU and the fully connected layer is 32 and 64, and each fully connected layer is followed by a dropout layer with a rate of 0.5. The number of branches is also a hyperparameter, and we find that 4 performs best after a lot of experiments. The frequency and pixel domain sub-networks are pre-trained and further fine-tuned during the jointly training process. When pre-training the pixel domain sub-network, we use the data augmentation strategy to improve its generalization performance. In the whole network, we add a lot of batch normalization layers to accelerate its convergence, and we set the batch size to be 64. The model is trained for 300 epochs with early stopping to prevent overfitting. We use ReLU as the non-linear activation function, and use Adam algorithm to optimize the loss function.

D. Performance Comparison

In this subsection, we conduct some experiments to compare the performance of MVNN and existing baselines described in Section V-B (**EQ1**). We use the accuracy, precision, recall and F1 score of the fake-news image class as evaluation metrics.

From Tab. I, we can see that: 1) MVNN is much better than other baselines, which validates that MVNN can effectively capture the intrinsic characteristics of fake-news images. Specifically, MVNN achieves an accuracy of 84.6%, outperforming existing approaches by 9.2% at least. 2) Intuitively, Fine-tuned VGG performs better than Pre-trained VGG, indicating that the learned features are more relevant to the task of fake news detection after fine-tuning the model on the fake news dataset. 3) The performance of ConvAE is slightly better than Pre-trained VGG. This demonstrates that ConvAE has the ability of understanding universal semantics of images, which is similar to models pre-trained in a supervised manner. 4) The performance of FF+LR is the worst among these compared methods because the information captured by these forensics features is very limited.

E. Ablation Study

In this subsection, we aim to evaluate the effectiveness of multiple domains and other network components in improving the performance of MVNN (**EQ2**) from the quantitative and qualitative perspectives.
Quantitative Analysis

To intuitively illustrate the effectiveness of different domains and other network components, we design several internal models for comparison, which are simplified variations of MVNN by removing certain components:

- **w/o frequency domain (fd):** The frequency domain sub-network is removed from MVNN while the attention mechanism is saved for fusing features of the pixel domain sub-network.
- **w/o pixel domain (pd):** We removed the pixel domain sub-network and associated attention mechanism. The remaining structure is the frequency domain sub-network and the binary classifier.
- **w/o attention:** MVNN without the attention mechanism. Features from the frequency and pixel domain sub-networks are concatenated to make the classification.
- **w/o Bi-GRU:** MVNN without the Bi-GRU in the pixel domain sub-network.
- **w/o branches:** In addition to removing the Bi-GRU, we also remove the branches in the pixel domain sub-network. The input image is fed into the pixel domain sub-network and obtains an output vector from the last block, which is further used to make the classification.

The results of the ablation study are reported in Tab. II. We have the following observations:

1) Multiple domains: The frequency and pixel domain both play important roles for fake news detection, especially the pixel domain. Specifically, the accuracy drops by 5.2% and 10.9% without the frequency and pixel domain sub-network correspondingly, which further demonstrates that physical and semantic visual information are both important for detecting fake news. Besides, the performance of MVNN without the pixel domain is obviously worse than the frequency domain. It illustrates that the pixel domain plays a major role in fake news detection while the frequency domain is auxiliary.

2) Network Components: Other than multiple domains, other network components: attention, Bi-GRU and branches in pixel domain sub-network, are all important for achieving the best performance by MVNN. If we remove one or several of them, the performance would drop by a certain degree. Specifically, i) the accuracy is 1.9% lower than the whole model if we remove attention, which means that the attention mechanism can fuse physical and semantic visual feature vectors better than simply concatenating, further proving our assumption that these features contribute to the task dynamically; ii) When we remove the Bi-GRU from the pixel domain sub-network, the performance of MVNN reduces 1.8%; when
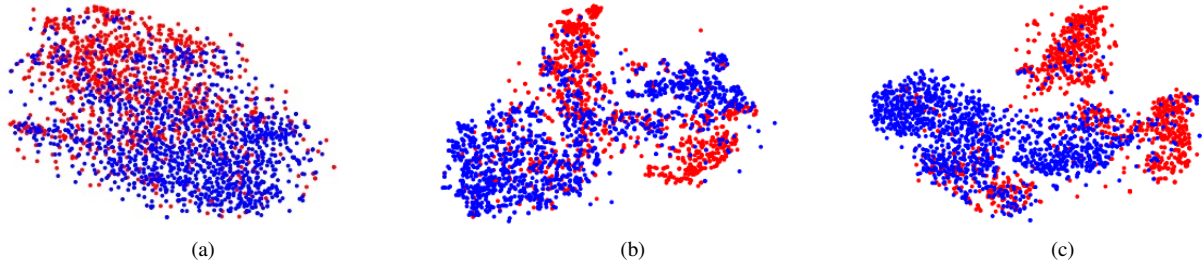
Fig. 7: Visualization of learned latent visual feature representations on the testing data. Dots in red and blue represent fake-news images and real-news images, respectively. (a) Frequency domain sub-network. (b) Pixel domain sub-network. (c) MVNN.

we further remove the branches, the performance drops by 4.3%. Therefore, we can conclude that incorporating different levels of features and considering the dependencies between these features both help capture the semantic characteristics of visual contents.

Qualitative Analysis

In the previous ablation experiments, we have intuitively evaluated the utility of multiple domains in improving the performance of MVNN. In this subsection, we qualitatively visualize the visual features learned by the frequency domain sub-network, the pixel domain sub-network and MVNN which considers both two domains on the testing set with t-SNE [36], as shown in Fig. 7. The label for each image is a real-news image or fake-news image.

From Fig. 7, we can easily observe that the separability of the feature representations learned by these three networks can be sorted as: MVNN > the pixel domain sub-network > the frequency domain sub-network. Specifically, in the visualization graph of the frequency domain sub-network, the feature representations of positive and negative samples overlap a lot. This is because all images uploaded to social media platforms will be compressed, which covers the original compression or tampering traces to some extent, reducing the difference between fake-news images and real-news images in the frequency domain. For the pixel domain sub-network, it can learn discriminable features, but the learned features are still twisted together, especially in the middle part of Fig. 7b. In comparison, in the visualization graph of MVNN, there is a relatively visible boundary between samples with different labels. From the above phenomena, we can conclude that: i) pixel domain is more effective than frequency domain in distinguishing fake-news images and real-news images; and ii) frequency and pixel domains are complementary for detecting fake-news images, consequently MVNN which fuses the information of multiple domains can learn better and more distinctive feature representations, and thus achieves better performance than single domain sub-network.

F. Application on Multi-modal Fake News Detection

In the previous parts, we have proven the effectiveness of MVNN in detecting fake news utilizing visual contents alone. Considering that a complete news story consists of text and visual content, the latest researches usually utilize multi-modal information to detect fake news. Following the convention, we compare the performance of different visual representations obtained from above visual modeling methods in fake news detection under the condition of fusing text information (**EQ3**), as described in Tab. III. We experiment with three state-of-the-art fusing methods as follows:

- **attRNN**: Reference [16] proposes an innovative RNN with an attention mechanism for effectively fusing the textual, visual and social context features. In detail, the neuron attention from the output of the LSTM is utilized when fusing with the visual features. In our experiments, we remove the part dealing with social context features.
- **EANN**: Reference [17] designs an event adversarial neural network to learn event-invariant multi-modal features through adversarial training techniques. It consists of a multi-modal feature extractor, a fake news detector and an event discriminator.
- **MVAE**: Reference [18] utilizes a multi-modal variational autoencoder trained jointly with a fake news detector to learn a shared representation of textual and visual information for fake news detection. MVAE is composed of an encoder, a decoder and a fake news detector.

Tab. III shows the similar trend as Tab. I. From Tab. III, we observe that MVNN consistently outperforms other baselines

TABLE III: The performance comparison of fake news detection in multi-modality.

|	Method	Accuracy	Precision	Recall	F1
attRNN	FF+LR	0.735	0.801	0.665	0.727
	Pre-trained VGG	0.821	0.813	0.862	0.837
	Fine-tuned VGG	0.849	0.888	0.818	0.852
	ConvAE	0.816	0.848	0.796	0.821
	MVNN	**0.901**	**0.911**	**0.901**	**0.906**
EANN	FF+LR	0.780	0.840	0.724	0.778
	Pre-trained VGG	0.821	0.861	0.791	0.824
	Fine-tuned VGG	0.841	0.883	0.807	0.843
	ConvAE	0.823	0.863	0.794	0.827
	MVNN	**0.897**	**0.930**	**0.872**	**0.900**
MVAE	FF+LR	0.777	0.776	0.815	0.795
	Pre-trained VGG	0.813	0.893	0.737	0.804
	Fine-tuned VGG	0.832	0.875	0.798	0.835
	ConvAE	0.827	0.831	0.847	0.839
	MVNN	**0.891**	**0.896**	**0.898**	**0.897**
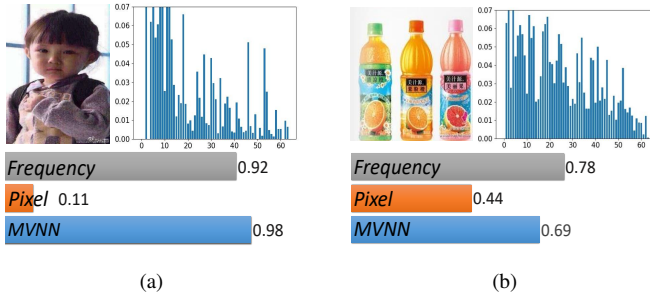
Fig. 8: Some fake-news images detected by the frequency domain sub-network and MVNN but missed by the pixel domain sub-network.

Fig. 9: Some fake-news images detected by the pixel domain sub-network and MVNN but missed by the frequency domain sub-network.

by a large margin on all fusion methods. Specifically, MVNN exceeds compared methods by over 5.2% in accuracy, which suggests that MVNN can easily replace existing methods to obtain the representations of visual contents, for achieving significant improvements in detecting fake news. Moreover, the performance of FF+LR in attRNN is obviously worse than EANN and MVAE, because attRNN can hardly utilize the semantic alignment between the text and the forensics features to fuse the textual and visual information.

G. Case Studies

To further illustrate the importance of multiple domains for fake news detection, we compare the results reported by MVNN and the single domain sub-networks, and exhibit some successful examples which are captured by MVNN but missed by the single domain sub-networks.

Fig. 8 shows two fake-news images detected by the frequency domain sub-network and MVNN but missed by the pixel domain sub-network. Fig. 8a is used to make up a fake news post about a lost child, and Fig. 8b is an illustration of a piece of fake news which says that juice drinks can cause leukemia in children. The image in the upper left is the original image attached to the fake news post, and the histogram in the upper right shows the variance of DCT coefficients on 64 frequencies, which reflects the characteristics of this image in the frequency domain. In general, the larger the variance is, the heavier the image has been re-compressed. From the perspective of semantics, the two images in Fig. 8 do not show evidence of fake news while their frequency histograms look quite suspicious, showing that they are very likely to be outdated images. The scores obtained from MVNN and single domain sub-networks are listed at the bottom of each figure. We can observe that these two examples are correctly classified by MVNN while the results would totally change without the frequency information, which further proves the importance of the frequency domain information in fake news detection.

Fig. 9 shows another two fake-news images detected by the pixel domain sub-network and MVNN but missed by the frequency domain sub-network. Fig. 9a tells a piece of fake news that the South Korean President Lee Myung-bak knelt
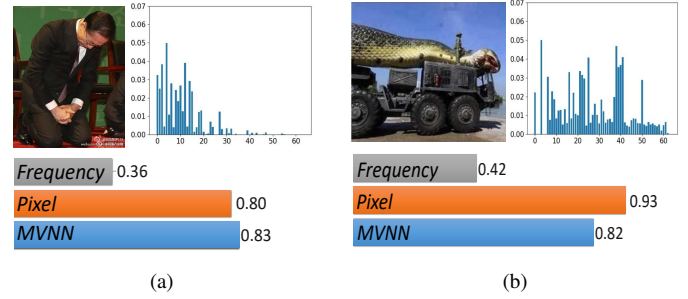
down to apologize to the nation for the sexual assault of a seven-year-old girl. In fact, he was just praying. Fig. 9b is used to fabricate a fake news story about a huge snake which is just a mock-up. We can see that the two frequency histograms in Fig. 9 show little evidence of fake news, however, the image contents are eye-attracting and rather dubious. By combining the information of the frequency and pixel domain, MVNN can easily detect that this is a fake-news image with high confidence. Therefore, the information of the pixel domain is also important for detecting fake news.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a novel framework MVNN to model the visual contents for fake news detection, which takes advantage of the visual information of the frequency and pixel domains to effectively capture and fuse the characteristics of fake-news images at physical and semantic levels. Experiments conducted on Weibo dataset validate the effectiveness of MVNN, and further prove the importance of multiple domains in fake news detection.

There are several future works that need further investigations. First, MVNN is a general framework for extracting effective visual representations to detect fake news, which is not limited to platforms. In this paper, we only evaluate the proposed model on Weibo data due to the limited distinctive images in the existing Twitter multimedia dataset. In future work, we can construct a larger multimedia dataset from the Twitter platform and explore the generalization capacity of the proposed model on different datasets. Further, we can compare the similarities and differences between the visual contents of Weibo and Twitter data. Second, although there are already many studies focusing on fusing multi-modal information for fake news detection, it is still a challenging problem that needs further investigations. For example, we can use the semantic alignment between images and text to explore the role of different modalities. At last, how to explain the decisions made by existing models based on multi-modal information is worth considering, since it can help us further understand and defend fake news.

REFERENCES

[1] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[2] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[3] S. Kumar and N. Shah, "False information on web and social media: A survey," *arXiv preprint arXiv:1804.08559*, 2018.

[4] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media: A survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 2, p. 32, 2018.

[5] L. Wu, F. Morstatter, X. Hu, and H. Liu, "Mining misinformation in social media," *Big Data in Complex and Social Networks*, pp. 123–152, 2016.

[6] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel visual and statistical image features for microblogs news verification," *IEEE Transactions on Multimedia*, vol. 19, no. 3, pp. 598–608, 2017.

[7] C. Boididou, S. Papadopoulos, D.-T. Dang-Nguyen, G. Boato, M. Riegler, S. E. Middleton, A. Petlund, Y. Kompatsiaris *et al.*, "Verifying multimedia use at mediaeval 2016." in *MediaEval*, 2016.

[8] S. Sun, H. Liu, J. He, and X. Du, "Detecting event rumors on sina weibo automatically," in *Asia-Pacific Web Conference*. Springer, 2013, pp. 120–131.

[9] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th International Conference on World Wide Web*. ACM, 2011, pp. 675–684.

[10] K. Wu, S. Yang, and K. Q. Zhu, "False rumors detection on sina weibo by propagation structures," in *IEEE 31st International Conference on Data Engineering (ICDE)*. IEEE, 2015, pp. 651–662.

[11] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks." in *International Joint Conference on Artificial Intelligence*, 2016, pp. 3818–3824.

[12] J. Ma, W. Gao, and K.-F. Wong, "Detect rumors on twitter by promoting information campaigns with generative adversarial learning," in *Proceedings of the 2019 World Wide Web Conference*. ACM, 2019.

[13] K. Wu, S. Yang, and K. Q. Zhu, "False rumors detection on sina weibo by propagation structures," in *2015 IEEE 31st International Conference on Data Engineering*. IEEE, 2015, pp. 651–662.

[14] K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," *Web Search and Data Mining*, pp. 312–320, 2019.

[15] C. Boididou, S. Papadopoulos, D.-T. Dang-Nguyen, G. Boato, and Y. Kompatsiaris, "The certh-unitn participation@ verifying multimedia use 2015." in *MediaEval*, 2015.

[16] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 795–816.

[17] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "Eann: Event adversarial neural networks for multi-modal fake news detection," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 849–857.

[18] K. Dhruv, G. Jaipal Singh, G. Manish, and V. Vasudeva, "Mvae: Multimodal variational autoencoder for fake news detection," in *Proceedings of the 2019 World Wide Web Conference*. ACM, 2019.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[20] Z. Jin, J. Cao, J. Luo, and Y. Zhang, "Image credibility analysis with effective domain transferred deep networks," *arXiv preprint arXiv:1611.05328*, 2016.

[21] C. R. Sunstein, *On rumors. How falsehoods spread, why we believe them, what can be done*. Farrar, Straus and Giroux, 2009.

[22] P. J. Lang, "A bio-informational theory of emotional imagery," *Psychophysiology*, vol. 16, no. 6, pp. 495–512, 1979.

[23] K. C. Dzmitry Bahdanau and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[24] F. Yang, Y. Liu, X. Yu, and M. Yang, "Automatic detection of rumor on sina weibo," in *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*. ACM, 2012, p. 13.

[25] C. Boididou, K. Andreadou, S. Papadopoulos, D.-T. Dang-Nguyen, G. Boato, M. Riegler, Y. Kompatsiaris *et al.*, "Verifying multimedia use at mediaeval 2015." in *MediaEval*, 2015.

[26] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[27] Z. Lin, J. He, X. Tang, and C.-K. Tang, "Fast, automatic and fine-grained tampered jpeg image detection via dct coefficient analysis," *Pattern Recognition*, vol. 42, no. 11, pp. 2492–2501, 2009.

[28] Y.-L. Chen and C.-T. Hsu, "Detecting recompression of jpeg images via periodicity analysis of compression artifacts for tampering detection," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 2, pp. 396–406, 2011.

[29] T. Bianchi and A. Piva, "Image forgery localization via block-grained analysis of jpeg artifacts," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 1003–1017, 2012.

[30] Q. Wang and R. Zhang, "Double jpeg compression forensics based on a convolutional neural network," *EURASIP Journal on Information Security*, vol. 2016, no. 1, p. 23, 2016.

[31] J. Yang, Y. Sun, J. Liang, Y.-L. Yang, and M.-M. Cheng, "Understanding image impressiveness inspired by instantaneous human perceptual cues," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[32] X. Zhu, L. Li, W. Zhang, T. Rao, M. Xu, Q. Huang, and D. Xu, "Dependency exploitation: A unified cnn-rnn approach for visual emotion recognition." in *International Joint Conference on Artificial Intelligence*, 2017, pp. 3595–3601.

[33] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 202–211.

[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[35] J. Masci, U. Meier, C. Dan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *International Conference on Artificial Neural Networks*, 2011.

[36] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.