Self-Supervised Distilled Learning for Multi-modal Misinformation Identification

Michael Mu Sreyasee Das Bhattacharjee Junsong Yuan
The State University of New York at Buffalo, NY, USA
{msmu, sreyasee, jsyuan}@buffalo.edu

Abstract

Rapid dissemination of misinformation is a major societal problem receiving increasing attention. Unlike Deepfake, Out-of-Context misinformation, in which the unaltered unimode contents (e.g. image, text) of a multi-modal news sample are combined in an out-of-context manner to generate deception, requires limited technical expertise to create. Therefore, it is more prevalent a means to confuse readers. Most existing approaches extract features from its uni-mode counterparts to concatenate and train a model for the misinformation classification task. In this paper, we design a self-supervised feature representation learning strategy that aims to attain the multi-task objectives: (1) task-agnostic, which evaluates the intra- and inter-mode representational consistencies for improved alignments across related models; (2) task-specific, which estimates the category-specific multi-modal knowledge to enable the classifier to derive more discriminative predictive distributions. To compensate for the dearth of annotated data representing varied types of misinformation, the proposed Self-Supervised Distilled Learner (SSDL) utilizes a Teacher network to weakly guide a Student network to mimic a similar decision pattern as the teacher. The two-phased learning of SSDL can be summarized as: initial pretraining of the Student model using a combination of contrastive self-supervised task-agnostic objective and supervised task-specific adjustment in parallel; finetuning the Student model via self-supervised knowledge distillation blended with the supervised objective of decision alignment. In addition to the consistent outperformances over the existing baselines that demonstrate the feasibility of our approach, the explainability capacity of the proposed SSDL also helps users visualize the reasoning behind a specific prediction made by the model.

1. Introduction

The spread of misinformation whether in the form of a full-fledged news article or just a small tweet has raised significant concern in various domains e.g., politics, finance, society, and others[1, 2]. According to Weibo's 2020 annual report, [42], $76,107$ news contents shared on Weibo social media platform were identified as false by the authority all year round. As an emerging field of research, evaluating misinformation has attracted attention of researchers across multiple disciplines (Social Science, Communication, Journalism, Computer Science). To ensure maximum impact in its audience, content creators of such misleading news articles frequently utilize multi-modal information, e.g. texts and images, to describe topics. A specific type of malicious multi-modal manipulation efforts, deep fakes [27, 39, 6, 12], has received significant attention from researchers, who attempt to develop automated methods for detecting such distortions. Nevertheless, a common phenomenon in recent years, popularly known as Out-of-Context images [15, 36], is far more prevalent a means to spread misinformation. It leverages existing unaltered images as is, but represents an irrelevant and misleading fact via newly coupled text.

Unlike deep fakes, generating an out-of-context multi-modal news content requires very limited technical expertise [23]. In fact, such a manipulation is more difficult to identify as none of its unimode contents is distorted in itself, and there are humongous ways of generating such misleading contents. While the present practices for verification rely significantly on manual fact-checking efforts, an automated mean to facilitate the process is a need of the hour. A set of existing methods [40, 45, 46] attempt to identify such misinformation by leveraging available evidence like entities, context, social media responses/reactions to posts, etc. However, many of these methods restrict their focus to text-based metadata to validate the claim, whereas validating the cross-modal correspondences may be critical to success in detecting such image repurposing events. While some recent works [41, 3] aim to approximate this multi-modal relation, many of these rely on the consistency of non-optimal factual information (e.g. named entities information) or external open-domain evidences to support the fact-checking task. Nonetheless, in reality the availability of such auxiliary information may not be a feasible option for all kinds of news contents due to many reasons including being cost-prohibitive.

In contrast, we aim to design an explainable classifier that categorizes multi-modal news content as 'Falsified' (if their unimodal contents are not pairwise consistent among one other) or 'Pristine' (if pairs of unimodal contents are consistent). Unlike previous work [5], no image in the training collection is required to have two or more captions to illustrate an inconsistency. In addition to being explainable, which helps the model justify its decision by highlighting the query image-regions contributing (or detracting) to its veracity attributes, the proposed *Self-Supervised Distilled Learner* (SSDL) adopts a two-phased Self Supervised Learning (SSL) strategy. It utilizes a weak guidance from text-modal input on the accompanying image to build an initial classifier. This then works as a baseline *Student* model for the second phase of learning. To further enhance the generalization capacity of this initial *Student* network, at the second phase, the proposed self-supervised knowledge distillation strategy leverages a *Teacher* network (separately pretrained in a SSL setting, but remains frozen during the knowledge distillation phase) and the baseline *Student* model is further finetuned to mimic a similar decision pattern as the *Teacher* over an identified set of data samples. In particular, the primary contributions of the work include the followings:

1. *A multi-modal multi-task SSL framework that combines language driven in-content information and self supervision* to evaluate the veracity factor of a given news content.

2. *A process of Knowledge Distillation within a self-supervised scenario* that helps transfer knowledge from a larger *Teacher* network to a specialized yet smaller *Student* model, for multi-modal misinformation identification.

3. *Evaluation with Explanation Visualization* scheme that enables the model to attribute the decision making (e.g. image sub-regions or text segments influencing inconsistency decision).

2. Related Work

A significant number of works have focused on detecting fakenews and rumor, wherein the objective has primarily been on evaluating a uni-modal news content[7, 43, 33, 24, 30]. Some recent works have leveraged multi-modal information to improve the decision precision[47, 8]. However, in this work we address the identification of another important kind of misinformation, in which none of the unimode content of a multi-modal news sample is altered, but the alteration appears only in its manipulated the image-text correspondence. This is often categorized as a type of Cheapfake, which being easy to create, is more prevalent and damaging than Deepfake [29]. In this section, we review of recent literature in this and other relevant topics to highlight the unique contribution of the proposed SSDL model.

Multi-modal Information Verification: A set of works [28, 23, 9, 5], closely related to the works on image repurposing [21], explore generic semantic correspondence between the constituent unimode components (i.e. text-image) of a multi-modal news content to verify its content veracity. To describe the huge spectrum of data patterns and the lack of a sufficiently representative data collection, oftentimes, existing literature [32, 26] use synthetic data collection, by randomly combining real images with real captions of other news contents (but not its own) to generate the out-of-context image samples. It is imperative that such synthetic data collection may not sufficiently reflect the challenges of a real-life problem scenario, as the existence of a weak/no relation between an image and a random text caption may provide an easy and explicit cue toward its inconsistencies. In a recent work, Aneja et al. [5] utilize a specially tailored data collection (wherein each news sample is a real image combined with a pair of captions collected from distinct news resources) and their objective is to establish if two captions accompanying an image are consistent. While the assumptions on the availability of such dataset may impact the model's plausibility in a generic test setting, its disproportionate reliance on text-mode may also lead to an increased decision level bias. Furthermore, none of these methods address the explainability issue, which may specifically be critical for such socially sensitive use-case settings.

However, many of these methods were evaluated in simulated (real captions are replaced by a random caption to generate a 'Falsified' sample) or specially tailored dataset (each image present in the dataset may be accompanied by two captions from two different sources), are not really appropriate to take on the real-world malpractice challenges. Another category of works introduces multi-modal fact checking methods, which leverage external knowledge base for information validation[26, 47]. Sahar et al. [3] collect evidences for both visual and textual components to perform the cycle consistency checks. However, relying on such external information for validity check makes the entire approach very expensive, complex, memory intensive, and difficult to deploy in a generic test setting. Furthermore, in this use case setting the primary objective is to early verify the content of a news item. Therefore, availability of enough evidences related to this news topic in web may not be assured in general.

Self-supervised Learning: With the recent advances in Contrastive learning, Self-supervised learning has emerged as an effective learning model, which leverages inter-

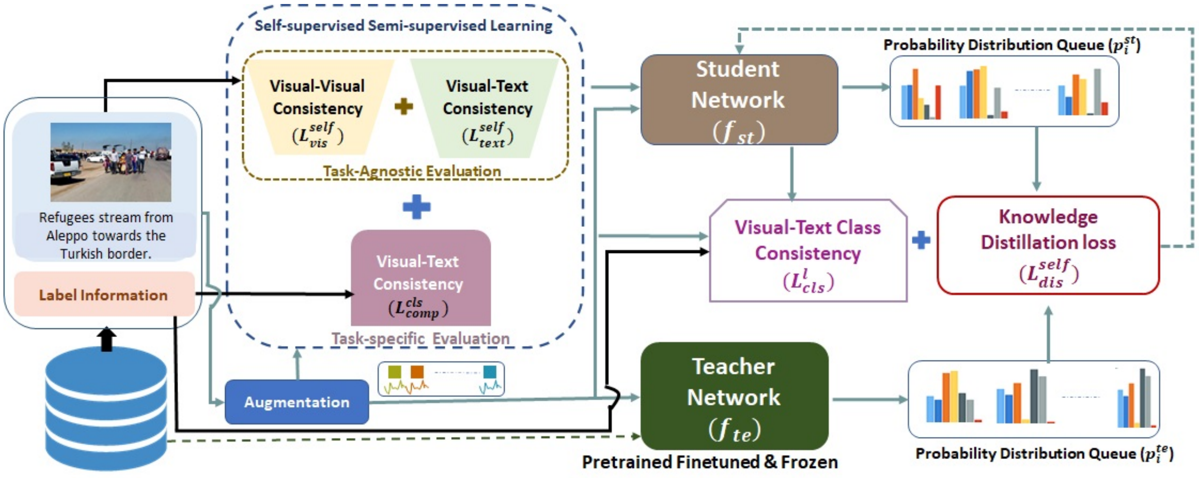
Self-supervised Semi-supervised Learning
Visual-Visual Consistency (L^{self}_{vis})
Visual-Text Consistency (L^{self}_{text})
Task-Agnostic Evaluation
Visual-Text Consistency (L^{cls}_{comp})
Task-specific Evaluation
Refugees stream from Aleppo towards the Turkish border.
Label Information
Augmentation
Student Network (f_{st})
Probability Distribution Queue (p^{st}_i)
Visual-Text Class Consistency (L^{l}_{cls})
Knowledge Distillation loss (L^{self}_{dis})
Teacher Network (f_{te})
Pretrained Finetuned & Frozen
Probability Distribution Queue (p^{te}_i)

Figure 1. Overview of our proposed method

nal structural and appearance consistencies among different data regions to drive training of a predictive model [16, 10, 18, 44, 25]. Different types of contrastive objectives have also been proposed to enhance the discriminability of SSL-based feature representation [17, 11]. To maximize the learning effect in this limited yet widely varied data environment, in this work we leverage the strength of SSL architecture in a semi-supervised multi-task learning setting with two complementary objectives: task agnostic (learning without class labels) and task-specific (learning with class labels).

Knowledge Distillation: To address the challenges of overfitting in a neural network model, knowledge distillation [20] attempts to transfer knowledge from a larger network (often referred as *Teacher Network*) to another smaller one (often referred as *Student Network*), without having to bear the burden of learning from scratch. A set of works formulates knowledge distillation using different learning objectives[4, 38, 34] to enhance feature discriminability. However, most of these methods adopt a supervised learning scenario, where the *Student Network* gets a task-specific transferred knowledge from the *Teacher Network*, rather than the task-agnostic knowledge components. The proposed SSDL introduces a multimodal representation learning in self-supervised manner, in which cross-modal consistency is verified at various levels of details.

3. Proposed Method

3.1. Problem Statement

The overview of the proposed model is shown in Figure 1. Given a multi-modal news content $(\mathbf{v}, \mathbf{t})$ (with its visual component $\mathbf{v}$ and text component $\mathbf{t}$), the objective of the proposed *Self Supervised Distilled Learner* (SSDL) is to estimate its cross-modal consistency status in terms of a binary class label ('Pristine' or 'Falsified'). To enable maximum utilization of the limited size of annotated data collection, SSDL designs a self-supervised learning approach, which uses both *Task Specific* and *Task Agnostic* constraints by leveraging the multi-modal training collection $\mathcal{D}_{train} = \{n_i\}_{i=1}^{|\mathcal{D}_{train}|}$ to train an initial *Student* classifier S_{init}. In $\mathcal{D}_{train}$, each multi-modal news sample $n_i := (\mathbf{v}_i, \mathbf{t}_i, l_i)$ represents an instance of the category l_i (which can be 'Pristine' or 'Falsified') using a visual component $\mathbf{v}_i$ and text component t_i. The baseline S_{init} is then finetuned following a knowledge distillation process for transferring the domain-specific multi-modal knowledge to further enhance the overall learner capacity.

3.2. Multi-modal Multi-task Self-Supervised Learning

Given a pre-trained vision-language model (e.g., CLIP [31]), we represent its encoder function as $f(.|\theta_\mathbf{c})$. Unless specified otherwise, for notation simplicity, we will omit $\theta_\mathbf{c}$ and denote the function as $f()$ instead. In fact, the encoder $f()$ combines knowledge of English-language concepts (represented using the text component $\mathbf{t}$) with semantic knowledge (represented using the image component $\mathbf{v}$) from a raw multimodal input $(\mathbf{v}, \mathbf{t})$ to the fixed-dimensional multi-mode descriptor $(f(\mathbf{v}), f(\mathbf{t}))$.While the model is generic and does not rely on the specific choice of the pretrained vision-language models, we have chosen CLIP encoder due to their improved performance compared to the other non-contrastive options [35, 13]. These descriptors are then used to learn S_{init} by means of a non-linear function g(·) (e.g. deep projection head). For each sample $n_i \in \mathcal{D}_{train}$, during this first learning phase, several views of its visual component $\mathbf{v}_i$ are created to be validated for consistency with $\mathbf{t}_i$ and their mutual self-consistencies via two types of loss components: *Task Agnostic* and *Task Spe-*

cific.

Given an image $\mathbf{v}_i$ each of its augmented (we have used random crop, color distortion, and Gaussian blur for augmentation) pairs $\mathbf{x}_i^j$ and $\mathbf{x}_i^k$ are encoded via CLIP to obtain their derived representations $\mathbf{c}_i^j$ and $\mathbf{c}_i^k$, which we represent with $\mathbf{c}_i^l = f(\mathbf{x}_i^l)$ for $l = 1, 2$. These descriptors are then used as input to $g()$ to generate the resulting contrast enhanced descriptors $\mathbf{z}_i^j$ and $\mathbf{z}_i^k$.

3.2.1 *Task Agnostic* Self-Supervised learning Objective

To learn a semantically relevant multi-modal representation with a training collection $\{(\mathbf{v}_i, \mathbf{t}_i)\}_i$, we adopt self-supervised learning approach SimCLRV2 [11] to contrastively learn a multimodal descriptor by maximizing two components: the intra-mode consistency between two different augmented visual components of the same image; the inter-mode consistency using the ratio of the pairwise consistency of each augmented visual component with the accompanying text component compared to the overall image-text cross-modal consistency. Following a mini-batch learning framework, the contrastive loss between $\mathbf{z}_i^j$ and $\mathbf{z}_i^k$ is defined as follows:

$$\mathcal{L}_{vis}^{self}(\mathbf{z}_i^j, \mathbf{z}_i^k) = -log\left(\frac{e^{sim(\mathbf{z}_i^j, \mathbf{z}_i^k)/\tau}}{\sum_{p=1}^{B+1}\mathbb{I}(p \neq j)(e^{sim(\mathbf{z}_i^j, \mathbf{z}_i^p)/\tau})}\right) \tag{1}$$

where $\mathbb{I}$ is the indicator function, τ is a temperature parameter, $sim()$ is the similarity function (e.g. scaled cosine similarity), and p iterates through $B + 1$ batch size. The relative consistency between $\mathbf{x}_i^j$ and its accompany text component $\mathbf{t}_i$ compared to the cross-modal consistency between the whole image $\mathbf{v}_i$ and $\mathbf{t}_i$ is defined as follows:

$$\mathcal{L}_{text}^{self}(\mathbf{z}_i^j, \mathbf{t}_i) = -log\left(\frac{e^{sim(\mathbf{z}_i^j, h(\mathbf{t}_i))/\tau}}{\sum_{p=1}^{B+1}\mathbb{I}(p \neq j)e^{sim(\mathbf{z}_i^p, h(\mathbf{t}_i))/\tau} + s_i}\right) \tag{2}$$

where $h = g \circ f$ and $s_i = exp(sim(h(\mathbf{v}_i), h(\mathbf{t}_i))/\tau)$ is the overall image-text cross-modal consistency.

3.2.2 *Task Specific* Consistency Learning Objective

To evaluate the model's task specific understanding and its cross-modal correspondence at the fine-grained details, we compare the consistencies between the category-specific predictive distributions of the augmented samples generated from the same dataset instance. In particular, as the multi-modal samples, both $(\mathbf{c}_i^j, f(\mathbf{t}_i))$ and $(f(\mathbf{v}_i), f(\mathbf{t}_i))$ should report similar category-specific prediction distributions. Toward this, we utilize cross-entropy based CLIP-like multi-modal losses [31], defined as:

$$\mathcal{L}_{cls}^{l}(\mathbf{z}_i^j, h(\mathbf{t}_i), l_i) = -l_i log(m_i^{l,j}) - (1 - l_i) log(1 - m_i^{l,j}) \tag{3}$$

for $l = 1, 2$ and $m_i^{1,j} = sim(\mathbf{z}_i^j, h(\mathbf{t}_i)^T)$, $m_i^{2,j} = sim(h(\mathbf{t}_i), (\mathbf{z}_i^j)^T)$, and l_i is the label. As mentioned before, $sim(,)$ is the scaled cosine similarity between its two argument vectors. Intuitively, we expect the function $g()$ to demonstrate higher discriminability via its learned unimode descriptors $\mathbf{z}_i^j$ and $h(\mathbf{t}_i)$, so that $m_i^{l,j}$ is higher when the corresponding augmented sample $(\mathbf{c}_i^j, f(\mathbf{t}_i))$ is generated from a training sample $(\mathbf{v}_i, \mathbf{t}_i)$ representing the category 'Pristine'. Then the comprehensive task-specific consistency loss component is computed as $\mathcal{L}_{comp}^{cls} = 1 - min(\frac{a_i^j}{a_i}, \frac{a_i}{a_i^j})$, where $a_i^j = \mathcal{L}_{cls}^1(\mathbf{z}_i^j, h(\mathbf{t}_i), l_i) + \mathcal{L}_{cls}^2(\mathbf{z}_i^j, h(\mathbf{t}_i), l_i)$ and $a_i = \mathcal{L}_{cls}^1(h(\mathbf{v}_i), h(\mathbf{t}_i), l_i) + \mathcal{L}_{cls}^2(h(\mathbf{v}_i), h(\mathbf{t}_i), l_i)$. While the term a_i quantifies the cross-modal category-specific similarity observed in the unimode components for $(\mathbf{v}_i, \mathbf{t}_i)$, we assume that a_i^j (which quantifies the cross-modal similarity of the unimode components for the augmented sample $(\mathbf{x}_i^j, \mathbf{t}_i)$) and a_i would typically demonstrate nearly identical cross-modal similarity patterns. Therefore, minimizing $\mathcal{L}_{comp}^{cls}$ is equivalent to maximizing the category-specific cross-modal consistency in the learned descriptor via $g()$.

Finally, the total loss function deployed for learning S_{init} is computed as: $\mathcal{L}_{init}^{tot} = \mathcal{L}_{vis}^{self} + \mathcal{L}_{text}^{self} + \mathcal{L}_{comp}^{cls}$. While several scaling configurations can be employed to weigh each of these components, we have not used any scaling in our experiments.

3.3. Self-supervised Knowledge Distillation

To further improve the generalization ability of this baseline *Student* classifier S_{init}, we introduce a self-supervised knowledge distillation that may, leveraging the limited-sized data collection in $\mathcal{D}_{train}$, transfer knowledge from a finetuned larger *Teacher* network to enhance discriminability of S_{init}. The pretrained vision-language encoder $f()$ (which we have designed using CLIP network) is finetuned by the annotated data collection $\mathcal{D}_{train}$ to define the *Teacher* network, $f_{te}(.|\theta_{te})$.

Given the sample $n_i := (\mathbf{v}_i, \mathbf{t}_i, l_i)$ from $\mathcal{D}_{train}$, its visual component $\mathbf{v}_i$ is used to generate a batch of its augmented versions $\{\mathbf{x}_i^j\}_{j=1}^B$. Then we expect the distribution of similarity scores between $\mathbf{t}_i$ and the elements of batch $\{\mathbf{x}_i^j\}_j$, obtained from the *Teacher* network and that computed by the *Student* network represented by $f_{st}(.|\theta_{st})$ (which was initialized by S_{init}), should be similar. For notation simplicity, again we will omit the corresponding learnable parameters of *Teacher* (and *Student*) network and represent it as $f_{te}()$ (and f_{st}). In particular, this intuitive understanding is formulated by minimizing the Kullback-Leibler(KL) divergence between the *Student* and the *Teacher*'s similarity score distributions.

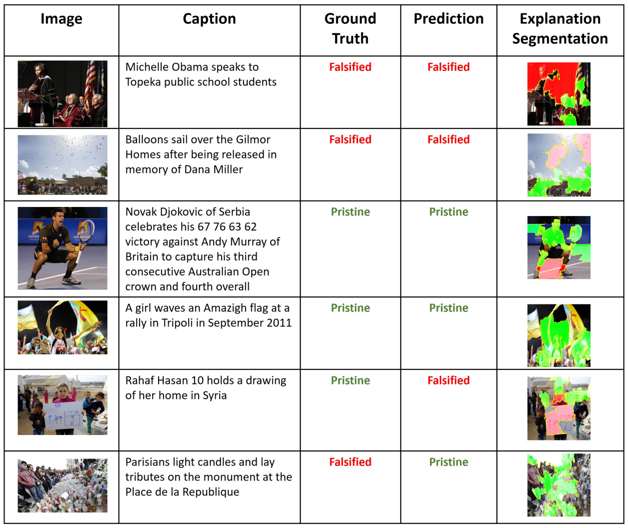

Image	Caption	Ground Truth	Prediction	Explanation Segmentation
	Michelle Obama speaks to Topeka public school students	Falsified	Falsified	
	Balloons sail over the Gilmor Homes after being released in memory of Dana Miller	Falsified	Falsified	
	Novak Djokovic of Serbia celebrates his 67 76 63 62 victory against Andy Murray of Britain to capture his third consecutive Australian Open crown and fourth overall	Pristine	Pristine	
	A girl waves an Amazigh flag at a rally in Tripoli in September 2011	Pristine	Pristine	
	Rahaf Hasan 10 holds a drawing of her home in Syria	Pristine	Falsified	
	Parisians light candles and lay tributes on the monument at the Place de la Republique	Falsified	Pristine	

Figure 2. Some example results from the NewsClipping Dataset[23]

For the batch $\{\mathbf{x}_i^j\}_{j=1}^B$, we define their pairwise similarity distribution with $\mathbf{t}_i$, as obtained by the *Teacher* network as $\mathbf{p}_i^{te} = [p_i^{te,1}, ..., p_i^{te,j}, ..., p_i^{te,B}]$, where $p_i^{te,j} = sim(f_{te}(\mathbf{x}_i^j), f_{te}(\mathbf{t}_i))$. Similarly, the pairwise similarity distribution of the batch of augmented visual components with $\mathbf{t}_i$, as obtained by the *Student* network, is defined as $\mathbf{p}_i^{st} = [p_i^{st,1}, ..., p_i^{st,j}, ..., p_i^{st,B}]$, where $p_i^{st,j} = sim(f_{st}(\mathbf{x}_i^j), f_{st}(\mathbf{t}_i))$.

Then the proposed knowledge distillation loss in a self-supervised learning scenario is formulated by optimizing the Kullback-Leibler (KL) Divergence[22] between $\mathbf{p}_i^{te}$ and $\mathbf{p}_i^{st}$, as:

$$\mathcal{L}_{dis}^{self}(\mathbf{v}_i, \mathbf{t}_i) = \mathbf{p}_i^{st} \cdot log\left(\frac{\mathbf{p}_i^{st}}{\mathbf{p}_i^{te}}\right) \tag{4}$$

Nevertheless, to leverage the available label information from $\mathcal{D}_{train}$, we also combine the distillation loss with the ground-truth labels of $n_i \in \mathcal{D}_{train}$ and define the total distillation loss ($\mathcal{L}_{dis}^{tot}$) is as the combination of the visual-text class inconsistency ($\mathcal{L}_{cls}^{l}$) and the knowledge distillation loss ($\mathcal{L}_{dis}^{self}$) and computed as follows:

$$\mathcal{L}_{dis}^{tot}(n_i) = \sum_{l \in \{1,2\}} \sum_{j=1}^{B} \mathcal{L}_{cls}^{l}(f_{st}(\mathbf{x}_i^j), f_{st}(\mathbf{t}_i)), l_i) + \mathcal{L}_{dis}^{self}(\mathbf{v}_i, \mathbf{t}_i) \tag{5}$$

4. Experiments

In this section, we will discuss the experimental details and the performance of the proposed method using large-scale public dataset.

Dataset: The proposed SSDL is evaluated using the recent, large-scale NewsCLIPpings Dataset[23], which contains multimodal (i.e. each sample has a text caption accompanied by an image component) news samples from two categories: 'Pristine' and 'Falsified'. A sample representing 'Falsified' category is comprised of an image, which does not align with its text caption component. It leverages the recently introduced VisualNews corpus that contains news from four different sources: BBC, The Guardian; The Washington Post; USA Today. Based on the details of how these samples were generated, the entire collection is considered as four mutually disjoint subsets: *Split 1 (or Semantics / CLIP Text-Image subset)* was created by using CLIP embeddings to find the highest similarity between nonmatching text-image pairs to create a falsified pairing; *Split 2 (or Semantics/CLIP Text-Text subset)* was created by using CLIP embeddings to find samples with similar textual embeddings to create out-of-context pairings; *Split 3 (or Person / SBERT-WK Text-Text subset)* was created by acquiring person entities, then matching an out-of-context image by finding the most semantically different, corresponding caption as determined by SBERT-WK score; *Split 4: (or Scene / ResNet Place subset)* was created by matching scenes with high Places365 image similarity as determined by the dot product of ResNet embeddings. Finally, the *Balanced* split mixes equal number of samples from all subsets to develop a more realistic sample collection and consists of $71,072$ train, $7,024$ validation, and $7,264$ test examples.

Experimental Settings: The proposed method relies on CLIP [31] to build the baseline for the study. The classification performances over the comprehensive *Balanced* split as well as all the other individual splits are reported using the *Accuracy* metric. To compare the performance of SSDL against the existing baselines, we report the *Accuracy* scores for the entire test collection. We also separately report the 'Falsified' category to specifically evaluate the system ability to identify misinformation.

Initial pretrained models are obtained from OpenAI and Facebook [25],[31]. The OpenAI implementation is pretrained on a dataset of 400 million image-text pairs. The Facebook model is pretrained on a filtered YFCC100M dataset [37], [31], which is dubbed YFCC15M [25] and consisted of 15 million image-text pairs. The learning using Adam optimization technique is based on the learning rate in the range $[10^{-6}, 10^{-5}]$. The batch-size is used as 16. For the finetuning process, the pretrained descriptor is fed into a 2-layer multi-layer perceptron (MLP) classifier. The learning process uses Cross-entropy loss, repeated for 90 epochs with an option for early stopping, and the learning rate is initiated to $5 \times e^{-5}$ with AdamW optimization. For the entire set of experiments, we have used $\tau = 30$.

Results: Figure 2 shows some qualitative results. As we observe, there is a clear correlation between the entities mentioned in the text and the objects present in their respective visual components. The top two examples represent the system predictions using two test queries from the 'Falsified' category. While the text components appear to

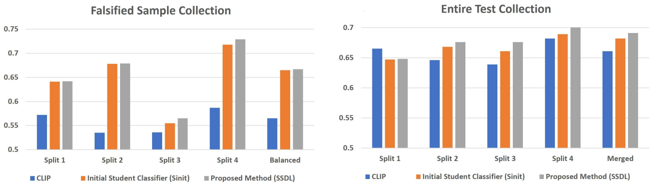

Figure 3. Comparative performance of the proposed SSDL against CLIP [23] in the combined test set shown using *Accuracy* metric, when a single model (using ViT-B/16 as the backbone) was learned using all the available training samples, available from all the splits in the dataset. The left plot shows the performance of SSDL in the 'Falsified' category and the right plot shows the performance of SSDL in the entire test collection.

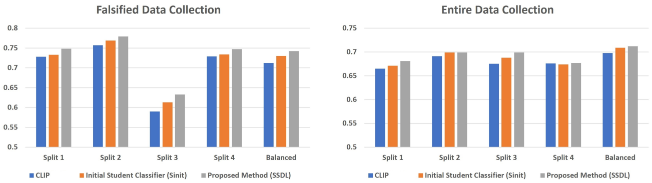

Figure 4. Comparative performance of the proposed SSDL in the combined test set shown using *Accuracy* metric, when a single model (using ResNet50 as the backbone) was learned using all the available training samples, available from all the splits in the dataset. The left plot shows the performance of SSDL in the 'Falsified' category and the right plot shows the performance of SSDL in the entire test collection.

have evidences of veracity, due to lack of correspondences between its two mode-specific representations, the system is able to correctly identify the multimodal queries as 'Falsified', a system characteristic that is aligned to the goal of this project, which is to combine truths and falsehoods to create a more convincing message. For the third and fourth queries, there are no serious mismatches between textual and visual information. The resulting "Pristine" classifications are therefore unsurprising. The Fifth and Sixth examples represent two misclassifications. For example, in the Fifth case, where the text component is "Rahaf Hasan ... holds a drawing", the cues related to the drawing being held up, contributes to a 'Pristine' classification. Meanwhile the buildings and people in the background contribute to a 'Falsified' classification decision. A clear connection can be drawn between the word "drawing" and the highlighted region that covers drawing, contributes as detractions. However, the contributions towards 'Falsified' is less clear. One possible reason is the lack of a clear corresponding entity in the text, leading to an inconsistency. In the Sixth example, the model finds several consistencies between text and image: candles, tributes, monument, and a little bit of Parisians. Although, this sample was actually 'Falsified,' there are very few indicators that the text and image are inconsistent. As humans, we can notice that a few people in the audience are smiling and region does not seem like the Place de la Republique. Nevertheless we can't be sure with so few scenic indicators. Since such facial expression analysis or external information on the entities were not taken into consideration by the proposed SSDL, with such small indicators, the system was unable to discern the true label.

The performance of the proposed SSDL, is compared against the recent CLIP [31] model with different backbones ViT-B/16 [14] and RN50 [19]. Figure 3 compares the general performance of the proposed multimodal misinformation identification method using ViT-B/16 as the backbone, which learns an initial *Student* classifier and later improves it via proposed self-supervised knowledge distillation module to build the full SSDL model.

Unified Model Performance: Following the experimental protocol by Radford et al. [31], we evaluate a single model trained on all the combined training set from all the splits, so that it is balanced with respect to both the categories. Based on the results reported in the left plot of the figure, SSDL shows a remarkable performance gain (**around 7% improvement over all the four splits and 10% improvement in the *Balanced* split**) in the 'Falsified' category. The right plot of the figure, which shows the comparative performance of SSDL in the entire test collection, reports **around $2-4\%$ improved accuracy score in 3 out of 4 splits**. We also note that in the *Balanced* split (which by its very structural definition, may be regarded as an aggregated snapshot of all types of misinformation samples available in the dataset), **SSDL reports around 3% improvement compared to CLIP in the experiments conducted using the entire test collection**. A similar performance is also observed in Figure 4, wherein the model is built using Resnet50 as its backbone and the performance of SSDL is compared against the baseline CLIP model. As observed in the right plot of the Figure 4, SSDL shows **around 1.2% average performance gain across all four splits and also reports around 1.5% improvement in the *Balanced* split**. In fact, per the statistics in the left plot of Figure 4, while CLIP shows a deteriorated performance trend in the 'Falsified' category of various dataset splits (e.g. **CLIP reported average *Accuracy* of around 70.1% Vs. SSDL reported average *Accuracy* of 72.7%** using Resnet50 backbone), SSDL demonstrates uniformly superior performances across all splits. Finally, we see **a boost of 3% in the performance of SSDL for the *Balanced* split**. In both figures 3 and 4, the initial *Student* classifier, learned in a multi-task self-supervised, semi-supervised scenario, exhibits consistently improved performances over CLIP. Then the consequential self-supervised knowledge distillation module, which specifically enables the model amplify its discriminability capacity in the challenging 'Falsified' category, helps SSDL enhance ts generalization capacity further.

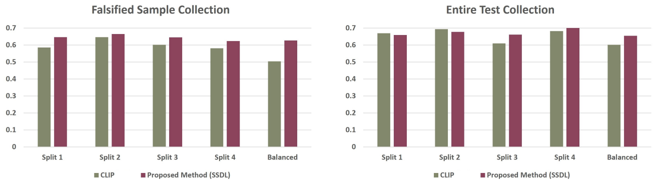

Figure 5. Comparative performance of the proposed SSDL in the split-specific test sets against the baseline [23] shown using *Accuracy* metric. We train a distinct classifier (using ViT-B/16 as the backbone) for each split. The left plot shows the performance of SSDL in the 'Falsified' category and the right plot shows the performance of SSDL in each split specific test collection.

Split Specific Model Performances: We report the performance of the multimodal SSDL classifier in Figure 5, in which we train distinct classifiers for each split individually. There is a noticeable tendency by CLIP to over-predict Pristine labels, which was also discussed by Luo et al. [23]. This indicates the model's confusion in correctly classifying many falsified samples as real, with Split 3 often being the most difficult. We note that this split models a threat scenario that queries for a specific person, with the intent to portray them in a false context. As observed in the left plot of the figure, the proposed SSDL (using ViT-B/16 as backbone) exhibits considerably improved performance ($2-6\%$ **across all splits**) in identifying 'Falsified' samples, compared to CLIP in all splits. Specifically in split 3, SSDL reports 5% improvement gain compared to its baseline. While several existing methods leverage external information on named entities to recognize the validity of the news relating to them, SSDL leverages the cross-view contexts at the regional level for validation. As we find from the right plot of the figure, SSDL dominates over CLIP in 3 (split 3, split 4, and *Balanced* split) out of all 5 test splits. Finally, SSDL reports **an average of** 2% **improved accuracy score compared to CLIP** (65.1% **Vs** 67.1%) **in all** 5 **test split collections**. While external information as an extra information source could definitely be useful to further enhance the performance, our objective in conducting these experiments was to evaluate the effectiveness of the SSDL without assuming an access to any auxiliary information sources.

Stability Analysis in Limited Data Environment As shown in the Figure 6, in a self-supervised semi-supervised learning scenario, the proposed SSDL attains a comparable performance with the baseline, by using only a smaller subset of whole training collection. More specifically, SSDL achieves an average of 65.1% (and 68.2%) test *Accuracy* highlighted in red (and in green) in the split-specific collections using ViT-B16 (and Resnet50) as the backbone, whereas **SSDL requires only** 50% **(and** 75%**) of the whole training collection to cross these benchmarks**.

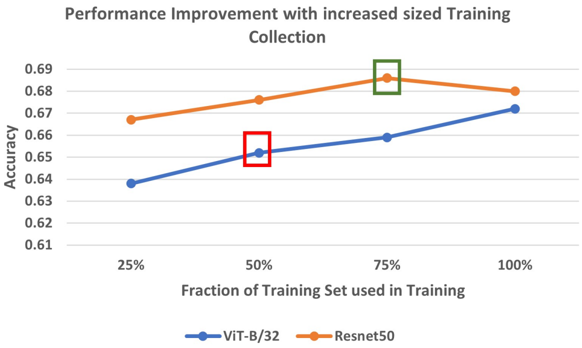

Figure 6. Performance Improvement of the proposed SSDL over increased sized training collection. Important to note that the CLIP-based baseline model used 100% of the training collection to attain an average of 65.1% (and 68.2%) test *Accuracy* highlighted in red (and in green) in the split-specific collections using ViT-B16 (and Resnet50) as the backbone, whereas SSDL requires only 50% (and 75%) of the whole training collection to cross these benchmarks.

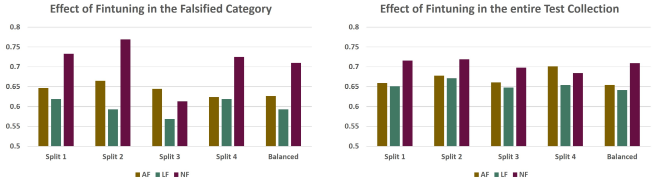

Figure 7. Comparing different finetuning strategies for SSDL classification performance (test set), where AF:=All frozen, LF:=Lower frozen, and NF - None frozen

Encoder Finetuning: In a set of experiments, we explore the performance improvement due to the finetuning of the pre-trained vision-language model. CLIP as our pre-trained encoder, we analyze the results by finetuning a varying number of CLIP layers. In the Figure 7, we use three fine-tuned models to report the results: RN50-all-frozen (AF, no CLIP layers fine-tuned); RN50-lower-frozen (LF, final few layers fine-tuned) 10, and RN50 (all layers fine-tuned). From both the plots in the figure, we note that finetuning all layers (RN50) does positively influence the performance in general, except for the Split 3, in which the performance slightly deteriorates. This may be due to the fact that the other training splits are comparably larger and therefore we can meaningfully finetune all layers whereas in the split 3, we do not have enough contextual evidences to do so. Nevertheless, the partial freezing does not appear to impact the overall performance of SSDL much.

Explainability Analysis The proposed explanation visualization module uses Local Interpretable Model-Agnostic Explanations (LIME)[1] to explain the decision made by the proposed SSDL. Using Lime, we found the words and visual features that contributed most to the final decision process. Figure 8 shows two example explanation segmenta-

[1] https://github.com/marcotcr/lime

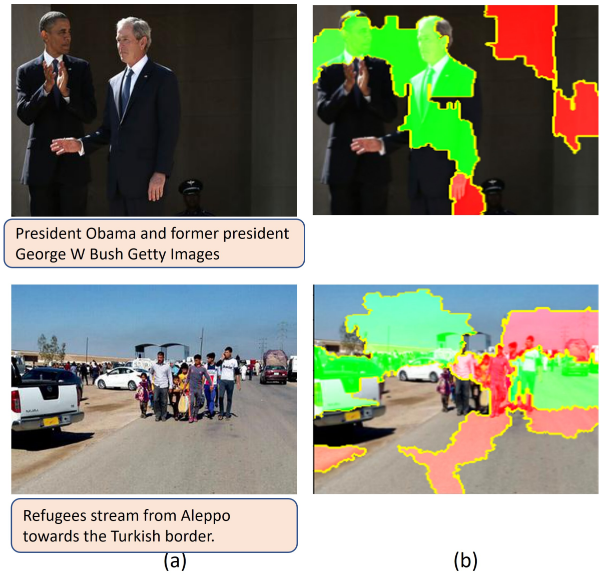
President Obama and former president George W Bush Getty Images
Refugees stream from Aleppo towards the Turkish border.
(a)
(b)

Figure 8. Examples of Explanation Visualization that highlights the words to explain the decision made by the proposed model in a multimodal environment. Each row in the column (a) represents an original query and each row in the column (b) represents the query specific explanation segmentation. The example in the top row is correctly classified by the system as 'Pristine' with probability 0.9, while the example in the bottom row is correctly classified as 'Falsified' with probability 0.54. The regions contributing to the model decision is highlighted in Green. The detracting regions are highlighted in Red.

tion results used to interpret the classification decision of the system. The contributing Green regions tend to focus on objects and subjects that act as evidence for the captions. The detracting regions are highlighted in Red. As observed, the example of the first row represents a correct classification of 'Pristine' with a strong likelihood of 90%, with the weight distribution to the caption words is nearly uniform. The row in the bottom of the figure shows an example, which was classified as 'Falsified' with a likelihood of 54%. The word "refugee" has received a weight of 24% followed by the word "border" receiving a weight of 15%.

Intuitively in the first example (top row of Figure 8), the captions mention both "President Obama" and "President George W Bush." Both of these individuals are commonly found in the news and the NewsCLIPpings dataset, so the model has been trained to recognize their features. This is further exemplified by the fact that the model focuses heavily on the defining features of each individuals, namely their face and parts of their outfits, which were identified by the system as the contributing Green. The detracting Red regions, show regions that seem inconsistent with the captions. Notably, there is a Red region encompassing George Bush's hand near a man's face in the background, and the other Red regions encompass a plain background. While SSDL appears confident in its prediction for the example as 'Pristine', the small Red regions in its image component are detracting, possibly because they lack an audience or vibrant background, which are typical in presidential photos, and there is also the lack of a gesticulating hand that most prominent speakers possess.

For the second example (bottom row of Figure 8), the caption mentions refugees traveling towards the Turkish border. The image-caption pair was correctly identified as "Falsified", but it was a much more difficult decision. In this case, the contributing regions focuses on cars in the background. Intuitively, a news visual rarely depicts cars when reporting on refugees, so the proposed SSDL has found the presence of cars in the image component to be inconsistent with the presence of the word "refugees" in its accompanying text component. The detracting red regions, on the other hand, tend to focus more on the individuals and the road itself. The presence of these human subjects are consistent with how the term "refugee" is represented in visual component, so the region encompassing the refugees are identified as detracting factors to the predicted category "Falsified".

5. Conclusion

In this paper, we propose a two-phased multimodal multi-task self-supervised semi-supervised learning strategy that evaluates both intra-and inter-mode self-consistencies, in conjunction with a category-specific supervised objective to build an initial *Student* classifier. This is later finetuned by leveraging a distilled guidance from a larger *Teacher* network in a self-supervised manner and thereby enhancing the model's generalized fact-checking capacity in a limited yet widely varied training data environment. Our work outperforms the baselines and offers an innovative benchmark of multi-modal fact-checking, which is not just more accurate, but also better explainable. In future we would like to extend our methods to evaluate information veracity in video-text multimedia content. To attain this, we intend to utilize the explanation feedback (more specifically cross-modal consistency between the caption and the contributing as well as detracting regions separately) further to improve the classification capacity of the *Student* classifier in an iterative manner.

Acknowledgment

This work is supported in part by the Center for Computational Research and Experiential Learning Network of the University at Buffalo and the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0124. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

References

[1] Coronavirus: The human cost of virus misinformation. `https://www.bbc.com/news/stories-52731624`.

[2] Youtube to remove all anti-vaccine misinformation. `https://www.bbc.com/news/technology-58743252`.

[3] Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14940–14949, 2022.

[4] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9163–9171, 2019.

[5] Shivangi Aneja, Christoph Bregler, and Matthias Nießner. Cosmos: Catching out-of-context misinformation with self-supervised learning. *arXiv preprint arXiv:2101.06278*, 2021.

[6] Shivangi Aneja and Matthias Nießner. Generalized zero and few-shot transfer for facial forgery detection. *arXiv preprint arXiv:2006.11863*, 2020.

[7] Sreyasee Das Bhattacharjee, Ashit Talukder, and Bala Venkatram Balantrapu. Active learning based news veracity detection with feature weighting and deep-shallow fusion. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 556–565. IEEE, 2017.

[8] Sreyasee Das Bhattacharjee and Junsong Yuan. Multimodal co-training for fake news identification using attention-aware fusion. In *Asian Conference on Pattern Recognition*, pages 282–296. Springer, 2022.

[9] Giscard Biamby, Grace Luo, Trevor Darrell, and Anna Rohrbach. Twitter-comms: Detecting climate, covid, and military multimodal misinformation. *arXiv preprint arXiv:2112.08594*, 2021.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[11] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

[12] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15108–15117, 2021.

[13] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11162–11173, 2021.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[15] Lisa Fazio. Out-of-context photos are a powerful low-tech form of misinformation. *The Conversation*, 14, 2020.

[16] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *Proceedings of the ieee/cvf International Conference on computer vision*, pages 6391–6400, 2019.

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

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

[20] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[21] Ayush Jaiswal, Yue Wu, Wael AbdAlmageed, Iacopo Masi, and Premkumar Natarajan. Aird: Adversarial learning framework for image repurposing detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11330–11339, 2019.

[22] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[23] Grace Luo, Trevor Darrell, and Anna Rohrbach. Newsclippings: Automatic generation of out-of-context multimodal media. *arXiv preprint arXiv:2104.05893*, 2021.

[24] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. Detecting rumors from microblogs with recurrent neural networks. 2016.

[25] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021.

[26] Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, and Ralph Ewerth. Multimodal analytics for real-world news using measures of cross-modal entity consistency. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 16–25, 2020.

[27] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2307–2311. IEEE, 2019.

[28] Christian Otto, Matthias Springstein, Avishek Anand, and Ralph Ewerth. Understanding, categorizing and predicting semantic image-text relations. In *Proceedings of the 2019*

on International Conference on Multimedia Retrieval, pages 168–176, 2019.

[29] Britt Paris and Joan Donovan. Deepfakes and cheap fakes. *United States of America: Data & Society*, 1, 2019.

[30] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, et al. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535*, 2020.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[32] Ekraam Sabir, Wael AbdAlmageed, Yue Wu, and Prem Natarajan. Deep multimodal image-repurposing detection. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1337–1345, 2018.

[33] Ronald Salloum, Yuzhuo Ren, and C-C Jay Kuo. Image splicing localization using a multi-task fully convolutional network (mfcn). *Journal of Visual Communication and Image Representation*, 51:201–209, 2018.

[34] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[35] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *European Conference on Computer Vision*, pages 153–170. Springer, 2020.

[36] Dana Statton Thompson, Stephanie Beene, Katie Greer, Mary Wegmann, Millicent Fullmer, Maggie Murphy, Sara Schumacher, and Tiffany Saulter. A proliferation of images: Trends, obstacles, and opportunities for visual literacy. *Journal of Visual Literacy*, 41(2):113–131, 2022.

[37] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[38] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

[39] Luisa Verdoliva. Media forensics and deepfakes: an overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020.

[40] William Yang Wang. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

[41] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing 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*, pages 849–857, 2018.

[42] Weibopiyao. Weibo 2020's annual report on refuting rumors. `https://weibo.com/weibopiyao?profile_ftype=1&is_all=1&is_search=1&key_word=2020`, 2020.

[43] Yue Wu, Wael Abd-Almageed, and Prem Natarajan. Deep matching and validation network: An end-to-end solution to constrained image splicing localization and detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1480–1502, 2017.

[44] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[45] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *Advances in neural information processing systems*, 32, 2019.

[46] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020.

[47] Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. Fact-checking meets fauxtography: Verifying claims about images. *arXiv preprint arXiv:1908.11722*, 2019.