Towards Automated Detection of Risky Images Shared by Youth on Social Media

Jinkyung Park
Vanderbilt University
Nashville, USA
jinkyung.park@vanderbilt.edu

Joshua Gracie
University of Central Florida
Orlando, USA
joshua_gracie@knights.ucf.edu

Ashwaq Alsoubai
Vanderbilt University
Nashville, USA
ashwaq.alsoubai@vanderbilt.edu

Gianluca Stringhini
Boston University
Boston, USA
gian@bu.edu

Vivek K. Singh
Rutgers University
New Brunswick, USA
v.singh@rutgers.edu

Pamela Wisniewski
Vanderbilt University
Nashville, USA
pamela.wisniewski@vanderbilt.edu

ABSTRACT

With the growing ubiquity of the Internet and access to media-based social media platforms, the risks associated with media content sharing on social media and the need for safety measures against such risks have grown paramount. At the same time, risk is highly contextualized, especially when it comes to media content youth share privately on social media. In this work, we conducted qualitative content analyses on risky media content flagged by youth participants and research assistants of similar ages to explore contextual dimensions of youth online risks. The contextual risk dimensions were then used to inform semi- and self-supervised state-of-the-art vision transformers to automate the process of identifying risky images shared by youth. We found that vision transformers are capable of learning complex image features for use in automated risk detection and classification. The results of our study serve as a foundation for designing contextualized and youth-centered machine-learning methods for automated online risk detection.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in HCI*; • **Security and privacy** → *Human and societal aspects of security and privacy*;

KEYWORDS

Youth Online Risk, Instagram, Private Message, Vision Transformer, Self-supervised Learning, Semi-supervised Learning

ACM Reference Format:
Jinkyung Park, Joshua Gracie, Ashwaq Alsoubai, Gianluca Stringhini, Vivek K. Singh, and Pamela Wisniewski. 2023. Towards Automated Detection of Risky Images Shared by Youth on Social Media. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3543873.3587607

WWW '23 Companion, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9419-2/23/04…$15.00
https://doi.org/10.1145/3543873.3587607

1 INTRODUCTION

Internet and social media have become a significant part of youth's daily lives. In 2022, 97% of U.S. teens are reported to be online daily, and 46% of them are online almost constantly [66]. While social media platforms provide opportunities to youth such as social connections [47], learning [23], and representing themselves [60], they can also expose youth to online risks. A recent survey confirmed that almost half of U.S. teens experienced online risks ranging from harassment, sexual messages, and threats [65]. Given the adverse impacts of such risks on youth [5], the phenomena of youth online risk are now one of the critical concerns to many scholars [20, 35, 40].

Recently, the popularity of media-based social media platforms has grown rapidly among youth [66]. In 2022, YouTube was used by 95% of U.S.teens and TikTok was by 67%, followed by Instagram (62%) and Snapchat (59%) [66]. Following the popularity, risks associated with media content (e.g., using someone else's photos to create fake profiles [57], or sharing explicit photos or pornography [67]) have increased [59]. Therefore, monitoring and validating media content shared online and preventing the spread of risky media content is pivotal for promoting youth online safety.

One way to mitigate youth online risk is the use of machine-learning algorithms to identify risky content. In this work, we leverage a computational framework to detect contextualized online risk. Instead of binary classification (e.g., risky vs. non-risky), we focused on contextual descriptions based on the risk type (e.g., harassment, sexual risk), media content type (e.g., screenshots, memes), and direction of risks (e.g., personally targeted vs. non-targeted). Given that risk is highly subjective and contextualized [14], these descriptions are critical for a better understanding of the youth online risk phenomena and for creating customized intervention (support) mechanisms. Obtaining these descriptions using an automated method is also important to tackle the scale, cost, and timeliness issues associated with the mitigation of youth online risk. The overarching question that we address in this study is:

RQ: How can we use human-centered insights for designing automated risky image detection methods to support youth online safety?

To explore our research question, we first collected 10,000 private conversations (i.e., Instagram Direct Messages) donated by 100 youth participants and asked them to annotate their own messages for risk and risk types. Then we had research assistants (RAs) annotate the same conversations for risk and risk types. Through this process, we collected risk labels for 686 media messages flagged by

youth and RAs. Next, we conducted qualitative analyses of risky media messages and identified two contextual risk dimensions that emerged. The key risk dimensions identified in our work are **risk types** (e.g., sexual messages, harassment, hate speech), **media content types** (e.g., meme, screenshot, art illustration), and the **nature of risk** (e.g., humor, broadcast, personally targeted). Then we labeled 686 media messages in terms of contextual dimensions. After excluding audio/video content, the final set of data that was prepared for automated analysis consisted of 550 images. From there, we applied self- and semi-supervised machine learning methods to identify various dimensions of youth online risk. A linear classifier trained with a self-supervised method yielded 82.2% accuracy in identifying the media content type, while a classifier trained with a semi-supervised method achieved 83.1% accuracy in classifying risk type. Our study makes the following contributions to the field, particularly to the youth's online safety literature:

- We introduce key contextual dimensions to understand youth online risks
- We design an automated process to identify those contextual dimensions of risks (with self- and semi-supervised learning methods).

2 BACKGROUND

2.1 Youth Online Safety

A wealth of research on youth's online safety has emerged over the past couple of decades, especially within the intersection of computer science and social science domain [54, 55]. In previous literature, researchers empirically explored various types of youth online risk perceptions and behaviors [15, 32, 41, 52, 54, 69, 70, 74] to address the risks associated with youth online interactions. Particularly, the popularity of media-based social media platforms has grown rapidly among youth [66] and the prevalence of risks associated with media content has grown [59]. There is a wide range of risks associated with media content sharing on social media such as posting an image of someone else without their consent [38], sharing a meme with the intent to embarrass or harass someone else [61], using someone else's photos to create fake profiles [57], or sharing explicit photos or pornography [67]. Addressing the issues of risks posed by media-based content sharing is important because the victims of media content risk can perceive more harm and trauma than they do from text-only content risk [43, 58], and it is expected that the prevalence of media-based content risks will only continue to grow as more social network platforms emphasize sharing multimedia content [59]. Hence, monitoring and validating media content shared online and preventing the spread of risky media content is pivotal for promoting youth online safety.

Meanwhile, the traditional way to monitor media content via manual content moderation has become almost impossible due to the massive volume of content that is being generated and shared online. One of the technological approaches to mitigate youth online risk is the use of algorithms to identify risky content. Researchers applied various computational approaches to detecting online risks such as cyberbullying [34, 58, 59], sexual risks [13, 33, 50], many of which involve simple binary classification tasks (e.g., risky or non-risky). The risk is highly subjective and contextualized in nature [14], especially when it comes to media content youth share privately on social media. Hence, binary classification tasks alone would not allow us to identify contextualized online risks youth experience. Recently, contextual dimensions of online risks have been addressed by a few studies [25, 64]. For instance, using publicly available Twitter data, Hassan et al. [25] developed automated models to identify the types of sexual violence (e.g., unwanted contact, nonphysically forced penetration, alcohol/drug-facilitated penetration, etc.), and the relationship between the perpetrators and the victims (e.g., intimate partner, family member, acquaintance, etc.) of sexual risks.

Although prior studies provided valuable insights into automatically identifying online risk, they often relied on *public* datasets, in part, due to legal and/or ethical challenges of obtaining private data. There has been relatively little work on exploring private interactions which are prone to more severe and variegated types of risks [1]. However, understanding the nature of online risk that occurs in private online spaces is important because many of the risks, such as sexual risks and the sale/promotion of illegal products are known to be greater in private spaces [24, 44, 50]. Building upon the prior work, we examined a rich and challenging-to-obtain dataset of youths' private social media data (i.e., direct messages) and defined online risk detection as multiple multi-class classification tasks (e.g., harassment, sexual solicitation, violence, etc.) to identify contextual descriptions of youth online risks.

Additionally, recent work on automated online risk detection often relied on the perspective of third persons to annotate ground-truth data [27, 51] without reflecting the perspectives of those who experienced the risks [34, 50]. Given that risk is highly subjective [14], relying solely on the perspective of a third person (e.g., third-party annotators) would not provide an accurate view of the risks youth encounter online. To tackle this problem, we explored risky media content (e.g., images, illustrations) flagged by both youth participants who experienced risk (first-person perspective) and research assistants (third-person perspective) to annotate ground truth data from multi-perspectives. With risk annotations from both youth participants and RAs, we conducted qualitative analyses to understand the contextual dimensions of media risk youth experienced through private conversations on Instagram. Based on our findings, we applied automated methods to identify risky images at scale and to create youth-centered intervention mechanisms. Below, we provide background on automated approaches to identify online risks.

2.2 Automated Approaches to Detecting Risks in Images

In the previous literature, various efforts have been made to automatically detect risky images. In a line of work, scholars have applied hashing technologies (i.e., the use of file hashes to identify abusive content) to identify sexually abusive images or videos or minors [7, 48]. This approach has been applied by various law enforcement agencies to scale investigation of child sexual abuse investigations [48]. However, the main issue with this approach was that the techniques are not robust to content-preserving operations applied to modify the images (e.g., adding watermarks) [22]. Recent developments in the field of computer vision resulted in methods

that could be used for identifying risky content in images. These methods are based on the visual information extracted from the original images (e.g. skin color, nudity, texture, shape) [46, 59, 72, 76]. Particularly, deep learning approaches were applied to detect child pornography materials by utilizing pre-trained adult pornography detection models [39, 42, 46] or transfer learning to fine-tune models on pornography data [21]. For example, Nian, et al. [46] developed pornographic image detection where a dataset with more than 13000 pornographic images with adults was used to train a deep convolutional neural network. One of the main reasons for the success of deep learning methods to detect risky images was the availability of massive labeled datasets for training the models. However, in many cases, the challenge with youth online risk detection is the lack of extensive labeled datasets to train the models.

Recently, Vision Transformer (ViT) has been proposed as a state-of-art technique for image recognition tasks [16]. The key aspect that sets transformers apart from previous convolutional models is the use of self-attention, a mechanism by which each pixel (or patch of pixels) 'attends' to every other pixel (or patch of pixels) [63]. This mechanism allows the transformer to relate pixel features across the image and make sense of objects within an image. The ability of self-attention to focus on important areas/objects within a context made it a critical component in neural network models for multiple modalities [49], hence, made transformers achieve outstanding performance in various fields, including computer vision [37, 68], image classification [16, 31, 37, 62], object detection and segmentation [9, 77], as well as action recognition in videos [3, 6].

Although previous studies yielded accurate results in object detections with vision transformers, the same approaches may not be applied to complex tasks, such as the detection of risky images shared by youth. In recent work, a visual attention mechanism was applied to detect child pornographic content [22]. The classifier built upon a deep learning architecture with an attention mechanism was able to accurately discern child pornography vs. non-child pornography. However, given the subjective and contextualized nature of risk [14], binary classification tasks alone would not allow us to identify contextualized online risks youth experience. Therefore, we first need to understand the contextual dimensions of online risk in order to detect nuanced risks in images.

In this work, we enhance previous work by utilizing vision transformers with contextual dimensions of online risk to automatically identify risky images youth share online. We first identified the three risk dimensions based on the literature review (risk types) and a set of qualitative analyses (media content type and nature of risks). From there, we trained the vision transformers with self- and semi-supervised learning methods to identify risk images. We applied these two learning methods because the traditional supervised classification method would not be appropriate to train a vision transformers model due to an imbalance in our dataset (risk-label vs non-risk label). Below, we describe the semi- and self-supervised approaches for risky image classification.

2.2.1 Self-Supervised Image Classification. Self-supervised learning is the practice of learning feature representations *without* the use of pre-existing labels (e.g., learning the features that make up an image of a dog or cat without telling the model which images contain dogs or cats). By training without labels, the model itself must determine what features differentiate one image from another. Various methods have been created to facilitate self-supervised learning, with most methods focused on contrastive learning [11, 12, 26, 71]. The premise behind contrastive learning is to treat each image as its own separate class and compare them to augmentations of themselves as well as to augmentations of other images [10]. In this way, a model can learn the features that make up an image and how they differ from the features of other images.

In this work, we utilized Distillation with No Labels (DINO), a self-supervised co-distillation method [10]. DINO works by utilizing two networks, a student and a teacher, both of which have the same architecture but different weights. During training, the student and teacher are passed augmented variants of the input image, where the teacher receives two global crops of the image, and the student receives both the global crops as well as several local crops of smaller resolution [10]. DINO has been shown to be an effective method of self-supervised learning with self-attention maps that are capable of nearly 'segmenting' the objects in the images it is given. We used DINO as a general pre-text method of extracting features from our images for use in downstream classification tasks to identify risk types, media content types, and the nature of risks.

2.2.2 Semi-Supervised Image Classification. Semi-supervised learning differs from both supervised learning and self-supervised learning in that it makes use of some labels instead of all or none. Semi-supervised learning is typically done through the use of pseudo-labeling [36], a process by which a model is trained on a subset of a dataset that has labels. With semi-supervised learning, models produce pseudo-labels for the subset of unlabeled data and then retrain themselves on the larger set of labeled and pseudo-labeled data. The difficulty with this approach, however, is determining a good pseudo-label; in many cases, network confidence is used as a way of estimating the quality of a potential pseudo-label [56]. Still, an issue with the use of network confidence is the chance of producing noisy pseudo-labels due to poor network calibration. Since neural networks are generally poorly calibrated at the start of training, models can produce highly confident pseudo-label predictions that are very far off the actual target value.

To address this problem, scholars have applied the uncertainty regularization technique [45, 73, 75]. For instance, Uncertainty-Aware Pseudo-Label Selection (UPS) [53] utilizes network confidence *and* certainty to filter out noisy pseudo-labels during training. Pseudo-labels are produced using network output predictions that are chosen using a threshold value γ, which in the case of multi-class classification is the maximum probability in the output. To select the best pseudo-labels, UPS selects the labels that have a confidence value (network prediction output) $\geq \tau_p$ in the case of positive labels and $< \tau_n$ for negative labels. UPS then calculates an uncertainty estimation $u(p)$ which is then compared against threshold values κ_p and κ_n for positive and negative labels respectively. Once the pseudo-labels have been chosen, the loss for positive labels is the cross-entropy of the network output and the positive label, and the loss for negative labels is the negative cross-entropy loss of the network and the negative label. All losses are then summed to get the total loss for the network. In summary, the training sequence can be described as 1) initially train on the known dataset, 2) produce

positive and negative pseudo-labels, 3) train again using the new labels, and 4) repeat until all images are labeled.

In this work, we utilized UPS to train a linear classifier for downstream classification tasks with the features generated by DINO. During the training process, the UPS produced pseudo-labels for our dataset in terms of all three different risk dimensions (risk types, media types, and nature of risks).

3 METHODS

3.1 Data Annotation Process

3.1.1 Dataset. We collected Instagram Direct Messages (DMs) from youth between the ages of 13 and 21, who were then asked to flag private message conversations (a set of DMs) that made them or someone else feels uncomfortable or unsafe. We selected Instagram as it is one of the most popular social media platforms among youth [2]. Each participant was required to have an active Instagram account during the ages of 13-17. The participants were also required to have had at least 15 Direct Message (DM) conversations (private conversations that would not appear in users' feed, search, or profile [30]), two of which must have made them or someone else feel uncomfortable. To recruit youth participants, we promoted our study on social media, especially Facebook and Instagram. We also contacted more than 650 youth-serving organizations for offline recruitment. We developed a web admin tool to manually verify the collected data to ensure that participants met inclusion criteria and that the shared messages were from real participants. The total dataset was comprised of over 10,000 private conversations had by 100 youths between the ages of 13 and 21. For our automated analysis, we filtered the dataset to focus only on images and this gave us 50,442 images donated by 41 youth participants. Out of 50,422 images, 550 images were labeled as risky (described below), while the rest were labeled as non-risky. We used a set of 443 labeled images (80%) to train and a set of 107 labeled images (20%) to test a linear classifier with Uncertainty-Aware Pseudo-Label Selection (UPS). Below, we describe the labeling process in detail.

3.1.2 Ethical Considerations. Due to the sensitive nature of the dataset, we took the utmost care to ensure the confidentiality and privacy of the participants. This study was approved by the authors' Institutional Review Board (IRB). We disclosed ourselves as mandated child abuse reporters in the case of imminent risk posed to a minor and our federal obligation to report child pornography to the proper authorities. As we were unable to make diagnostic clinical decisions about participants' mental health conditions, we provided participants with help and support resources. We explicitly warned the participants not to upload digital imagery depicting the nudity of a minor and gave them clear instructions on how to remove such media from their data before uploading it to our system. In addition, we obtained a National Institute of Health Certificate of Confidentiality, which further ensures participant privacy and prevents the subpoena of the data during legal discovery.

For the data and analysis, we took special care by removing all personally identifiable information in any publication resulting from this dataset to ensure the confidentiality of our participants. We did not use any cloud-based services (e.g., Google Vision API) when analyzing our data to avoid sharing the data with third parties. We followed our data management plan which included only storing data in safe and restricted data storage approved by the university's information technology security audit team; researchers were not permitted to download the data on any personal devices. All researchers analyzing the data completed the Collaborative Institutional Training Initiative (CITI) human subjects research training and the initiation protection of minors training program. We also provided mental health support and adequate breaks for RAs who helped verify and qualitatively analyze the data as some of the content could be triggering or explicit.

3.1.3 Youth's Risk Annotations. First, each private conversation (a set of Direct Messages) was labeled by youth participants as either risky or non-risky. If the conversation was labeled as risky, participants were then asked to flag the specific messages that made that conversation risky, as well as to identify the type of risk(s) in those messages. Although we provided pre-defined risk types, we mentioned to participants to not limit the unsafe conversations to these categories and to flag any content that seems unsafe to them. Each risky media message was labeled with one or more risk types. The type of risks was assessed by seven categories identified in the existing literature [70] and the existing Instagram reporting feature categories similar to participants' experience on Instagram [29]. The seven categories included:

- Nudity/porn: Photos or videos of a nude or partially nude people or person
- Sexual messages/solicitations: Sending or receiving a sexual message ("sexting") - being asked to send a sexual message, revealing, or naked photo
- Harassment: Messages that contain credible threats, aim to degrade, or shame someone, contain personal information to blackmail or harass someone or threaten to post nude photos of someone
- Hate speech: Messages that encourage violence or attack anyone based on who they are; specific threats of physical harm, theft, or vandalism
- Violence/threat of violence: Messages, photos, or videos of extreme violence, or that encourage violence or attack anyone based on their religious, ethnic, or sexual background
- Sale or promotion of illegal activities: Messages promoting the use or distribution of illegal material such as drugs
- Self-injury: Messages promoting self-injury such as suicidal thoughts, cutting, and/or eating disorders

3.1.4 Research Assistants' Risk Annotation. Next, we enlisted six RAs to identify risky media messages. RAs were undergraduate research assistants at the last authors' institution, who ranged in age from 18 to early twenties. We had an interdisciplinary team of RAs from computer science, psychology, criminology, and sociology majors. After training on data annotation, each RA was assigned participants with which to review and annotate all of their DMs for risks. We had two RAs code each conversation to make sure that we captured as many risky media messages as possible. We had a shared codebook for RAs to expand the definitions and examples of risks, as well as to document discrepancies and difficult-to-judge

scenarios. We held weekly workshops for RAs to discuss their perspectives and come to a consensus on the conversations if they had conflicting ideas. After both youth participants and RAs completed risk annotation, we filtered the data to focus on media messages that had risk types flagged by the youth participants and/or RAs. Through this process, we collected 686 unsafe media messages from 127 private conversations annotated by 18 different participants and 6 research assistants.

3.1.5 Qualitative Analyses for Contextual Risk Dimensions. From there, we (researchers) conducted a set of qualitative analyses on the 686 risky media messages annotated by the youth participants and the RAs to determine the contextual risk dimensions. We started the qualitative analyses with a content analysis [17] to code each risky media message for manifest content: media content type. We did this process by familiarizing ourselves with the dataset and categorizing the major media content types among the entire risky media messages. Through the content analysis, we came up with the first contextual dimension, "media content type," which consisted of five different codes including:

- Meme: Digitally altered/created images usually containing both images and text
- Screenshot: Images of device screens
- Natural image of the person: Images of a person or body part in the natural world
- Natural image of objects: Images of an object or animal in the natural world
- Art Illustration: Drawn or illustrated artworks

Then, we performed a thematic analysis [8] to determine more nuanced patterns and/or themes that emerge. We began this process by revisiting the dataset and noting down some initial codes based on our observations, considering the larger conversation around the shared risky media. From there, we began the full coding process for two more rounds to refine the codes for potential themes. Through this iterative and comparative process, we came up with three key themes that emerged in terms of the "nature of risks":

- Humor: Risky images that contained a humorous undertone (non-serious),
- Broadcast: Risky images that were not directed toward any particular individual
- Personal: Risky images that were sent personally (i.e., to target or address the individual).

Based on the codes above, we labeled 686 risky media messages in terms of the "media content types" and "nature of risk." Each risky message was labeled one code from the media content type and the nature of risk dimensions. Some of the media messages contained one or more media content types (e.g., image of a person and image of objects in one media message). In this case, we assigned one code that is the most relevant to the context of the message. Finally, we excluded the videos and audio recordings from 686 risky media messages to focus only on the risky images for the image classification tasks. In total, we had 550 risky images labeled for risk types, media types, and the nature of risks.

In summary, "risk types" was annotated by youth participants and RAs using 7 risk type categories we identified based on the previous literature [70] and Instagram risk reporting feature [29].

Table 1: The number of labels for risk types (annotated by participants and RAs)

Risk Types	Number of Label
Harassment	94
Hate speech	30
Nudity/porn	95
Sale of illegal activities	33
Self-injury	17
Sexual messages	301
Violence/threat	64
Total	634

Table 2: The number of labeled images for the media type (annotated by the researcher)

Media Type	Number of Label
Natural Image (person)	87
Natural Image (object)	22
Meme	231
Screenshot	139
Art illustration	71
Total	550

Table 3: The number of labeled images for the nature of risks (annotated by the researcher)

Nature of Risks	Number of Label
Personal	112
Broadcast	150
Humor	286
Total	550

The "media content type" and the "nature of risk" were annotated by the researcher; these two contextual risk dimensions were generated by a set of qualitative analyses conducted by the researcher. Tables 1-3 show the number of labeled images for the three key risk dimensions that we identified in this work.

3.2 Automated Detection using Machine Learning

3.2.1 Analytical Approaches. In many cases, the challenge with computer vision (and machine learning as a whole) is the lack of extensive labeled datasets with which to create supervised classifiers. Our dataset had such a challenge, with only 550 of our 50,442 images being labeled as risky. The ground truth annotation to label youth's online risk is challenging due to the sensitivity of the topic and the difficulty in recruiting those who experienced online risks [18]. We faced additional challenges when labeling ground-truth data because we were working with private media messages, and therefore, every researcher and RA had to be approved and trained on privacy regulations before labeling the data. Due to such challenges, instead of traditional supervised classification, we explored the use of self- and semi-supervision to train a vision transformer model to detect risky images using our dataset. We first trained our

model with self-supervised learning, DINO [10], to extract features. With the features extracted by DINO, we trained a single-layer linear classifier (before any semi-supervised learning) and evaluated the performance of this linear classifier. Next, with the features extracted by DINO (fine-tuned model), we trained a single-layer linear classifier with semi-supervised learning, UPS [53], and evaluated the performance of trained linear classifiers. Below, we debrief the methods that we used to build risky image classification algorithms.

3.2.2 Feature Extraction using Self-Supervised Learning. We started the model training with feature extraction using DINO, a self-supervised learning [10] method. We performed two types of training with DINO. The first one was the raw training with the entire dataset (N = 50,442) without any pre-trained features. We did this training for 300 epochs with a 16x16 patch size using the same parameters and ViT architecture as in the original work that proposed the DINO model [10]. Then we downloaded the pre-trained DINO model with ImageNet weights [19] and trained a fine-tuned DINO model with both our dataset and ImageNet weights. We used ImageNet for pre-training because it was the largest dataset available by the time we pre-trained the DINO model (over 10 million) [28]. We did this training for 100 epochs with both 8x8 and 16x16 patch sizes using the same parameters as in the original paper [10]. Note that we did not make use of risk labels to train and extract features from DINO. Instead, we used the risk labels to train a linear classification model with UPS using the features extracted by DINO.

3.2.3 Linear Classification. To build a downstream classification model, we trained a single-layer linear classifier both with and without UPS (semi-supervised learning) using the same parameters as in the previous literature [10]. We split the dataset randomly, where 80% of the dataset (n = 433) was set aside for training the machine learning algorithms and the remaining 20% (n = 107) was utilized as a test set. For training without UPS, we trained the classifier for 100 epochs using the features extracted from the 1) raw-trained DINO using our dataset, 2) DINO pre-trained with ImageNet features, 3) and fine-tuned DINO trained with both our dataset and ImageNet (combined). Next, for training with UPS, we trained our classifier with features extracted by fine-tuned DINO model for 20 iterations with 20 epochs each for a total of 400 epochs of training using the same thresholds as in the original UPS paper [53]. When training a linear classifier to identify risk types, we included a non-risky class because we are classifying for risk type class for the entire set of images (N = 50,442), many of which are non-risky. We did not include a non-risky class for media types and direction of risk classification tasks because they are not necessarily tied to whether the images are risky or non-risky (e.g., being meme does not necessarily mean it is risky or non-risky). Figure 1 presents the visualized overview of the automated approaches that we applied in this study.

4 RESULTS

4.1 Linear Classification with DINO

A single-layer linear classifier with the DINO model pre-trained with our dataset resulted in reasonable performance with respect to media types (75.7% accuracy). However, it did not perform as accurately as a linear classifier with the DINO model pre-trained on ImageNet (82.2% accuracy). The best-performing model to classify media types was a linear classifier with the fine-tuned DINO model (83.2%). We note that these are noticeably higher than a baseline majority class classifier for the media type (one that classifies all instances into the majority class, 42%).

Table 4: Accuracy results from the linear classifiers using frozen features generated by different DINO models

Pre train	Risk Type	Media Type	Nature of Risk
Our dataset	**59.8%**	75.7%	71%
ImageNet	47.7%	82.2%	66.4%
Fine-tuned	56.1%	**83.2%**	**75.7%**

Table 5: Results from training a linear classifier with UPS on frozen DINO features

Metric	Risk Type	Media Type	Nature of Risk
Accuracy	83.1%	28%	61.3%
Precision	19.6%	20.9%	62.9%
Recall	16.9%	17.4%	69.4%
F1	17.3%	8.3%	61.3%

In terms of the nature of risks, a linear classifier with the DINO model pre-trained with our data also resulted in reasonable performance (71% accuracy). Contrary to the media type, it did perform more accurately than a linear classifier with the DINO model pre-trained on ImageNet (66.4% accuracy). A linear classifier fine-tuned with our data and ImageNet performed the best for classifying the nature of risks (75.7% accuracy). Regardless of pre-training methods, the accuracy results of all classifiers were noticeably higher than a baseline majority class classifier for the nature of risks (52%).

Meanwhile, compared to other classification tasks, performance on the risk types was less accurate for all DINO models, though the DINO pre-trained with our dataset performed better (59.8% accuracy) than the other two models (47.7% and 56.1% accuracy, respectively). However, the accuracy results were still higher than a baseline majority class classifier for the risk types (47%). Table 4 summarizes the accuracy results of this experiment.

Overall, the results show that the linear classifiers trained with features extracted from DINO models were able to accurately label the risky images in terms of all three risk dimensions: risk types, media content type, and nature of risks. That is, DINO was able to produce quality features from our dataset to be used for multi-class classification tasks.

4.2 Linear Classification with UPS

Table 5 summarizes the performance of the linear classifiers with pseudo-labels produced by UPS. For the risk types classification task, a classifier trained with UPS performed well in terms of accuracy (83.1%), but it did not in terms of precision and recall (19.5%, 16.9% respectively). When labeling the non-risky class, UPS achieved a 92% accuracy. However, when labeling the risk types, UPS only achieved a 12.5% on average across seven different categories. That is, UPS was not able to accurately identify risk types with the small number of labeled images and features extracted from DINO.

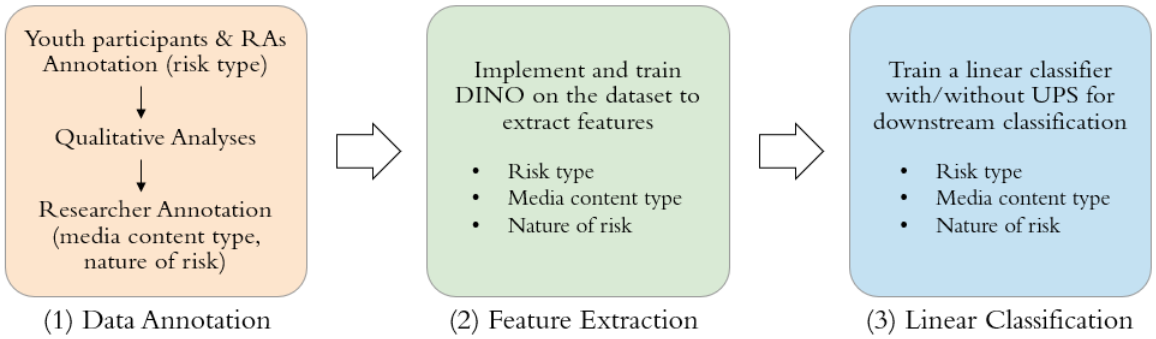

Figure 1: Overview of Automated Analysis Approach

UPS did perform fairly well when labeling the nature of risks (61.3% accuracy), with most of the messages being labeled as humor. Of those messages labeled as humor, 61% of them were labeled correctly, whereas 88% of the messages labeled as broadcast were correct and only 36% of the messages labeled as personally targeted were correct. On further review, most of the messages labeled by UPS as personally targeted were incorrectly attributed to broadcast messages with very few of the targeted messages being missed by UPS. This indicated that our linear classifier trained with UPS is sensitive to personally targeted messages and had a very low false negative rate for detecting targeted messages.

Meanwhile, our results indicated that a linear classifier trained with UPS struggled to properly label media types of the images in our dataset (28% accuracy). Most of the images labeled by UPS for media type were labeled as memes, the majority class of media type. All memes were correctly classified as memes, while the majority of the images with other media content types were incorrectly classified as memes.

5 DISCUSSION

5.1 Linear Classification

The results of our study show that DINO, a self-supervised method, is potentially capable of producing quality features from our dataset. Compared to the media content type and the nature of risks, performance on the risk types was less accurate for all DINO models. Given the number of classes for the risk types (N = 7), we consider the accuracy level of the classifier to be reasonable. Specifically, we observed that the DINO pre-trained with our dataset did perform better than the DINO pre-trained with ImageNet for the risk type classification task. This could be due to the ImageNet pre-trained DINO being better trained for less common image types in our dataset such as natural images of people and objects which are both common media types in ImageNet.

For downstream classification, our results showed that UPS (semi-supervised method) trained with the features extracted by DINO (self-supervised method) performed reasonably well to classify the nature of risks (humor, broadcast, and personal). However, the results also indicated that a linear classifier trained with UPS struggled to accurately label the media types of the images in our dataset. This could be due to the low number of labeled data in our study (N = 550) compared to the original work where UPS was trained with 1000-4000 labeled images [53].

When we closely examined the classification results, most of the images labeled by UPS for media types were labeled as memes, despite being of a different media type. This could be due to first, the majority of the labels for media types belonged to memes (n = 231/550), hence, the classifier was able to learn patterns better for memes compared to other media types. Second, the way we assigned media type labels and the synthetic nature of memes could have contributed to our results. During the data annotation process, we assigned media type labels mutually exclusively (i.e., assigning one media type label to one image), while memes can exhibit the characteristics of several different types of media (e.g., memes that include natural images of people/objects, or memes that include artist renders). This could lead to confusion for linear classifiers when trying to generalize the patterns with the unlabeled data that contained numerous examples of this cross-media type. Had we assigned more than one media content type label to memes, the classifiers could have been able to learn the patterns better and yield more accurate classification results. Similarly, if we included more complex examples of memes and other media types in our training dataset, the classifier may have been able to distinguish between the patterns for different media types more precisely.

Overall, our results indicated that more labeled data is needed to have machine-learning models to generalize contextual information of risks manifested in images. However, given the challenges with labeling contextual information on youth risk experience, particularly in the private sphere, we suggest that a better alternative would be to create a cascade or ensemble model. For instance, the ensemble model can make use of our classification models for the contextual risk dimensions (media types such as memes, screenshots, etc., and the nature of risks including humor, personal, and broadcast) by combining their outputs into a single input with the images themselves to a risk classifier. By combining this contextual information, the ensemble model may be able to achieve better classification results than just using the images themselves as input.

5.2 Implications for Youth-centered Online Risk Detection

In 2022, a large majority of U.S. teens perceived that social media sites and government officials are doing a poor job of addressing

the youth online risk. More importantly, teens who have experienced online risks - the most vulnerable - are more skeptical about how various groups fail to curtail online risks [65]. It is critical to build a youth-centered framework to intervene with online risk and put it into practice. Our work provides important insights into designing such youth-centered automated detection of online risks. First, part of our ground truth data was collected from youth who have experienced online risks in private conversations. We did so because a critical aspect of machine learning development is establishing ground truth that is reflective of the phenomenon in the real world [50]. For instance, adult users' data cannot be considered ecologically valid grounds for automated detection of cyberbullying or sexual solicitation targeted toward youth users because the behavioral patterns and linguistic styles would be different for the two populations. Our work adds novel insights into the field of online risk detection by collecting and analyzing risky images labeled by youths themselves and providing contextual dimensions for risk classification tasks.

Additionally, instead of a binary (e.g., risky vs. non-risky) classification, our work yielded contextual descriptions based on the risk type, media content type, and direction of risks (e.g., "harassment", "screenshot", and "personally targeted"). These descriptions are essential for a better understanding of the youth online risk phenomena and for creating customized intervention (support) mechanisms. Furthermore, applying these descriptions using an automated method is vital to tackle the scale, cost, and timeliness issues associated with youth online risk. The key risk dimensions that we generated in our study could serve as foundations for designing machine-learning models to detect youth online risks in a different context. Also, semi- and self-supervised learning methods could be applied to build more sophisticated models (e.g., cascade or ensemble models) to detect nuanced and contextualized online risks youth experience.

Finally, our findings have implications for the design of safety features for social media platforms, especially those with private messaging features. Given that risk is highly subjective and contextual [14], social media platforms need to consider risk context carefully when designing and applying filters to moderate privately-shared risky images. At the same time, our results indicate that identifying contextual risk dimensions with high accuracy is not trivial. To tackle this issue, social media platforms can implement feedback features to the filtering system so that youth can provide interactive feedback on contextual information about risky images to the system. They can even consider adding reward features for youth users to encourage them to actively participate in this interactive feedback process. This way, social media platforms can reflect unique perspectives of youth to design intervention systems to promote their online safety.

Taken together, we argue for designing youth-centered online risk detection models that can benefit those who are likely to be victimized in risky online interactions. Such youth-centered approaches to designing online risk detection systems will be more translatable in the real world. Our work is a step forward in this space so that translational research has a real-world impact on youth online safety.

5.3 Limitations and Future Work

The first and most critical challenge that we faced was a small number of risk labels (N = 550) to build automated risk detection algorithms. Although we tackled this problem with the use of semi and self-supervised methods, our results indicate that more labeled data is needed to have machine-learning models to generalize contextual information of risks manifested in images. Future research can investigate ways to increase the number of labeled examples of risk to ensure a diverse pool of risks to better aid machine learning models with generalization. In addition, we exclude the videos and audio from our dataset and focused exclusively on the images in this work. We acknowledge this as a limitation since online content is increasingly becoming multi-modal and popular social media platforms (e.g., YouTube and TikTok) are being nearly exclusively video-based. Future work can look to video action recognition or video object segmentation using vision transformers.

Another potential limitation would be sampling bias. Participants of our study must have registered as active users on Instagram for a certain time and signed up to donate their data for research. In addition, we consciously recruited youth who have experienced online risks on Instagram. Thus, we recognize that the ground-truth annotation from this study may not be the same for other youth populations. At the same time, this allowed us to work with those who are likely to experience online risks and hence to analyze a more ecologically valid dataset. Future research could endeavor to work with a more diverse pool of youth to label ground-truth data.

Even with the utmost care, we recognize ethical concerns that can arise from our research. While understanding the first-person perspective is valuable, studying online risks with youth can unintentionally put an "already vulnerable population at greater risk" [4]. For instance, reviewing and flagging risky media could have triggered youth participants to have uncomfortable feelings. Having said that, addressing the issues of youth online risks is critical, and research such as our work is necessary to identify a path forward toward youth-centered online safety measures. Additionally, since personal information could be easily traceable even in aggregated data, extra care for the privacy of the youth is needed. Future research should address ethical and privacy-preserving ways to work with sensitive datasets generated by the most vulnerable youth.

6 CONCLUSION

In this work, we identify key contextual dimensions of online risk youth experience online. Based on those dimensions, we explored the state-of-the-art automated methods (self- and semi-supervised methods) for identifying contextualized risks in images. The results of our study serve as a foundation for understanding contextualized online risk youth experience and designing youth-centered machine-learning methods for automated online risk detection.

ACKNOWLEDGMENTS

This research is supported in part by the U.S. National Science Foundation under grants #IIP-1827700, #IIS-1844881, and by the William T. Grant Foundation grant #187941. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the research sponsors.

REFERENCES

[1] Shiza Ali, Afsaneh Razi, Kim Seunghyun, Ashwaq Alsoubai, Joshua Gracie, Munmun De Choudhury, Pamela J Wisniewski, and Gianluca Stringhini. 2022. Understanding the Digital Lives of Youth: Analyzing Media Shared within Safe Versus Unsafe Private Conversations on Instagram. (2022).

[2] Monica Anderson and Jingjing Jiang. 2018. Teens, Social Media and Technology 2018. https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6836–6846.

[4] Karla Badillo-Urquiola, Zachary Shea, Zainab Agha, Irina Lediaeva, and Pamela Wisniewski. 2021. Conducting risky research with teens: co-designing for the ethical treatment and protection of adolescents. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–46.

[5] Fanni Bányai, Ágnes Zsila, Orsolya Király, Aniko Maraz, Zsuzsanna Elekes, Mark D Griffiths, Cecilie Schou Andreassen, and Zsolt Demetrovics. 2017. Problematic social media use: Results from a large-scale nationally representative adolescent sample. *PloS one* 12, 1 (2017), e0169839.

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding?. In *ICML*, Vol. 2. 4.

[7] Rubel Biswas, Victor González-Castro, E Fidalgo, and Deisy Chaves. 2019. Boosting child abuse victim identification in Forensic Tools with hashing techniques. *V Jornadas Nacionales de Investigación en Ciberseguridad* 1 (2019), 344–345.

[8] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*. Springer, 213–229.

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9650–9660.

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

[12] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297* (2020).

[13] Arijit Ghosh Chowdhury, Ramit Sawhney, Puneet Mathur, Debanjan Mahata, and Rajiv Ratn Shah. 2019. Speak up, fight back! detection of social media disclosures of sexual harassment. In *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Student research workshop*. 136–146.

[14] Jody Clay-Warner. 2003. The context of sexual violence: Situational predictors of self-protective actions. *Violence and victims* 18, 5 (2003), 543–556.

[15] Patricia De Santisteban and Manuel Gámez-Guadix. 2018. Prevalence and risk factors among minors for online sexual solicitations and interactions with adults. *The Journal of Sex Research* 55, 7 (2018), 939–950.

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

[17] James W Drisko and Tina Maschi. 2016. *Content analysis.* Pocket Guide to Social Work Re.

[18] Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: triangulating diagnostic signals. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–16.

[19] facebookresearch/dino. 2022. facebookresearch/dino. https://github.com/facebookresearch/dino

[20] David Finkelhor, Kerryann Walsh, Lisa Jones, Kimberly Mitchell, and Anne Collier. 2021. Youth internet safety education: Aligning programs with the evidence base. *Trauma, violence, & abuse* 22, 5 (2021), 1233–1247.

[21] Abhishek Gangwar, Eduardo Fidalgo, Enrique Alegre, and Víctor González-Castro. 2017. Pornography and child sexual abuse detection in image and video: A comparative evaluation. (2017).

[22] Abhishek Gangwar, Víctor González-Castro, Enrique Alegre, and Eduardo Fidalgo. 2021. AttM-CNN: Attention and metric learning based CNN for pornography, age and Child Sexual Abuse (CSA) Detection in images. *Neurocomputing* 445 (2021), 81–104.

[23] Christine Greenhow. 2011. Youth, learning, and social media. *Journal of Educational Computing Research* 45, 2 (2011), 139–146.

[24] Joel W Grube and Elizabeth Waiters. 2005. Alcohol in the media: content and effects on drinking beliefs and behaviors among youth. *Adolescent medicine clinics* 16, 2 (2005), 327.

[25] Naeemul Hassan, Amrit Poudel, Jason Hale, Claire Hubacek, Khandaker Tasnim Huq, Shubhra Kanti Karmaker Santu, and Syed Ishtiaque Ahmed. 2020. Towards automated sexual violence report tracking. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 250–259.

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

[27] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Analyzing labeled cyberbullying incidents on the instagram social network. In *International conference on social informatics*. Springer, 49–66.

[28] ImageNet. 2021. ImageNet. https://www.image-net.org/

[29] Instagram. 2022. Abuse and Spam. https://help.instagram.com/165828726894770

[30] Instagram. 2022. How do I send a message to someone on Instagram? https://help.instagram.com/155540431448273

[31] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. In *International conference on machine learning*. PMLR, 4651–4664.

[32] Lisa M Jones, Kimberly J Mitchell, and David Finkelhor. 2012. Trends in youth internet victimization: Findings from three youth internet safety surveys 2000–2010. *Journal of adolescent Health* 50, 2 (2012), 179–186.

[33] Aparup Khatua, Erik Cambria, and Apalak Khatua. 2018. Sounds of silence breakers: Exploring sexual violence on twitter. In *2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*. IEEE, 397–400.

[34] Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela J Wisniewski, and Munmun De Choudhury. 2021. A Human-Centered Systematic Literature Review of Cyberbullying Detection Algorithms. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–34.

[35] Robin M Kowalski, Susan P Limber, and Annie McCord. 2019. A developmental approach to cyberbullying: Prevalence and protective factors. *Aggression and Violent Behavior* 45 (2019), 20–32.

[36] Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, Vol. 3. 896.

[37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.

[38] Megan K Maas, Kyla M Cary, Elizabeth M Clancy, Bianca Klettke, Heather L McCauley, and Jeff R Temple. 2021. Slutpage use among US college students: the secret and social platforms of image-based sexual abuse. *Archives of sexual behavior* 50, 5 (2021), 2203–2214.

[39] Joao Macedo, Filipe Costa, and Jefersson A dos Santos. 2018. A benchmark methodology for child pornography detection. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 455–462.

[40] Sheri Madigan, Vanessa Villani, Corry Azzopardi, Danae Laut, Tanya Smith, Jeff R Temple, Dillon Browne, and Gina Dimitropoulos. 2018. The prevalence of unwanted online sexual exposure and solicitation among youth: A meta-analysis. *Journal of Adolescent Health* 63, 2 (2018), 133–141.

[41] Rose Maghsoudi, Jennifer Shapka, and Pamela Wisniewski. 2020. Examining how online risk exposure and online social capital influence adolescent psychological stress. *Computers in Human Behavior* 113 (2020), 106488. https://doi.org/10.1016/j.chb.2020.106488

[42] Felix Mayer and Martin Steinebach. 2017. Forensic image inspection assisted by deep learning. In *Proceedings of the 12th International Conference on Availability, Reliability and Security*. 1–9.

[43] Ersilia Menesini, Annalaura Nocentini, and Pamela Calussi. 2011. The measurement of cyberbullying: Dimensional structure and relative item severity and discrimination. *Cyberpsychology, behavior, and social networking* 14, 5 (2011), 267–274.

[44] Kimberly J Mitchell, David Finkelhor, and Janis Wolak. 2007. Youth Internet users at risk for the most serious online sexual solicitations. *American Journal of Preventive Medicine* 32, 6 (2007), 532–537.

[45] Subhabrata Mukherjee and Ahmed Awadallah. 2020. Uncertainty-aware self-training for few-shot text classification. *Advances in Neural Information Processing Systems* 33 (2020), 21199–21212.

[46] Fudong Nian, Teng Li, Yan Wang, Mingliang Xu, and Jun Wu. 2016. Pornographic image detection utilizing deep convolutional neural networks. *Neurocomputing* 210 (2016), 283–293.

[47] Gwenn Schurgin O'Keeffe, Kathleen Clarke-Pearson, Council on Communications, and Media. 2011. The impact of social media on children, adolescents, and families. *Pediatrics* 127, 4 (2011), 800–804.

[48] Claudia Peersman, Christian Schulze, Awais Rashid, Margaret Brennan, and Carl Fischer. 2016. iCOP: Live forensics to reveal previously unknown criminal media on P2P networks. *Digital Investigation* 18 (2016), 50–64.

[49] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. 2019. Stand-alone self-attention in vision models. *Advances in Neural Information Processing Systems* 32 (2019).

[50] Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Gianluca Stringhini, Thamar Solorio, Munmun De Choudhury, and Pamela J Wisniewski. 2021. A Human-Centered Systematic Literature Review of the Computational Approaches for Online Sexual Risk Detection. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–38.

[51] Andrew G Reece and Christopher M Danforth. 2017. Instagram photos reveal predictive markers of depression. *EPJ Data Science* 6 (2017), 1–12.

[52] Lauren A Reed, Margaret P Boyer, Haley Meskunas, Richard M Tolman, and L Monique Ward. 2020. How do adolescents experience sexting in dating relationships? Motivations to sext and responses to sexting requests from dating partners. *Children and Youth Services Review* 109 (2020), 104696.

[53] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. 2021. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329* (2021).

[54] Tara L Rutkowski, Heidi Hartikainen, Kirsten E Richards, and Pamela J Wisniewski. 2021. Family Communication: Examining the Differing Perceptions of Parents and Teens Regarding Online Safety Communication. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–23.

[55] Sarita Schoenebeck, Carol F. Scott, Emma Grace Hurley, Tammy Chang, and Ellen Selkie. 2021. Youth Trust in Social Media Companies and Expectations of Justice: Accountability and Repair After Online Harassment. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 2 (apr 2021), 18 pages. https://doi.org/10.1145/3449076

[56] Weiwei Shi, Yihong Gong, Chris Ding, Zhiheng MaXiaoyu Tao, and Nanning Zheng. 2018. Transductive semi-supervised deep learning using min-max features. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 299–315.

[57] Thiago H Silva, Pedro OS Vaz De Melo, Jussara M Almeida, Juliana Salles, and Antonio AF Loureiro. 2013. A picture of Instagram is worth more than a thousand words: Workload characterization and application. In *2013 IEEE International Conference on Distributed Computing in Sensor Systems*. IEEE, 123–132.

[58] Vivek K Singh, Souvick Ghosh, and Christin Jose. 2017. Toward multimodal cyberbullying detection. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2090–2099.

[59] Devin Soni and Vivek K Singh. 2018. See no evil, hear no evil: Audio-visual-textual cyberbullying detection. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–26.

[60] Catherine Cheng Stahl and Ioana Literat. 2022. # GenZ on TikTok: the collective online self-Portrait of the social media generation. *Journal of Youth Studies* (2022), 1–22.

[61] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, et al. 2021. Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 247–267.

[62] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*. PMLR, 10347–10357.

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[64] Nishant Vishwamitra, Hongxin Hu, Feng Luo, and Long Cheng. 2021. Towards understanding and detecting cyberbullying in real-world images. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*.

[65] Emily A Vogels. 2022. Teens and cyberbullying 2022. https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/

[66] Emily A Vogels, Risa Gelles-Watnick, and Navid Massarat. 2022. Teens, social media and technology 2022. https://www.pewresearch.org/internet/2022/08/10/teens-social-media-and-technology-2022/

[67] Kate Walker and Emma Sleath. 2017. A systematic review of the current knowledge regarding revenge pornography and non-consensual sharing of sexually explicit media. *Aggression and violent behavior* 36 (2017), 9–24.

[68] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 568–578.

[69] Yilin Wang and Baoxin Li. 2015. Sentiment analysis for social media images. In *2015 IEEE international conference on data mining workshop (ICDMW)*. IEEE, 1584–1591.

[70] Pamela Wisniewski, Heng Xu, Mary Beth Rosson, Daniel F Perkins, and John M Carroll. 2016. Dear diary: Teens reflect on their weekly online risk experiences. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3919–3930.

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

[72] Emilios Yiallourou, Rafaella Demetriou, and Andreas Lanitis. 2017. On the detection of images containing child-pornographic material. In *2017 24th International Conference on Telecommunications (ICT)*. IEEE, 1–5.

[73] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. 2019. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 605–613.

[74] Michele Zappavigna. 2016. Social media photography: construing subjectivity in Instagram images. *Visual Communication* 15, 3 (2016), 271–292.

[75] Zhedong Zheng and Yi Yang. 2021. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision* 129, 4 (2021), 1106–1120.

[76] Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J Miller, and Cornelia Caragea. 2016. Content-Driven Detection of Cyberbullying on the Instagram Social Network.. In *IJCAI*, Vol. 16. 3952–3958.

[77] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159* (2020).