Université d'Ottawa | University of Ottawa

Understanding the Threat and Challenge of Visual and Multimodal Disinformation (VMD)

SUMMARY OF A COLLABORATIVE STUDY BY THE COMPUTER RESEARCH INSTITUTE OF MONTREAL WITH THE INFORMATION INTEGRITY LAB OF THE UNIVERSITY OF OTTAWA

uOttawa
Laboratoire sur l'intégrité de l'information
Information Integrity Lab

CRIM

Cover image: Image generated by DALL-E using keywords: janus, one face real, one face digital

PREFACE

Disinformation poses a significant threat to society, with devastating impacts on institutions and individuals alike. Its effects span multiple sectors including the conduct of democratic processes, education, science, national security, and defense. By thwarting public discourse it hinders efforts to tackle urgent and existential global challenges such as those related to climate change and public health. The rise, propensity and intensity of disinformation is alarming, with large-scale campaigns becoming more frequent. Campaigns can spread more swiftly than the dissemination of verifiable facts, and thereby skew and influence public opinion and societal discourse.

Numerous accelerants and factors exacerbate the issue of disinformation: It spreads quickly on social networks, dodging regulatory efforts and making the development of reliable verification processes challenging. Advanced emerging technologies are also enabling the realistic falsification of text, images, audio, and video, which blurs the distinction between fake and real content, and these tools are becoming more user-friendly and widely available. However, disinformation detection methods lag behind these creation and falsification tools, particularly in addressing visual and multimodal disinformation (VMD), posing a significant obstacle in effectively managing disinformation.

In response to this growing challenge, the Information Integrity Lab (InfoLab) at the University of Ottawa and the Computer Research Institute of Montreal (CRIM) have formed a strategic partnership. This collaboration aims to study, develop, and disseminate knowledge on new tools and techniques specifically tailored to address the evolving field of visual and multimodal disinformation. Recognizing the rapid advancement in this area, the partnership is committed to continuous research and development, mirroring the evolution of disinformation technologies and strategies.

This report is a summary of an in-depth study produced by CRIM, in partnership with the uOttawa InfoLab, entitled "Visual and Multimodal Disinformation: Analysis, Challenges, Solutions", provides a comprehensive view of VMD, exploring methods to contain and combat it. It covers social, scientific, and technological aspects of VMD, offering a reference for those involved in confronting or tackling the phenomena, from academic researchers to technical solution developers, policymakers, media professionals, educators, and public awareness groups.

We begin by defining VMD and discussing its impact on society, emphasizing the need for adapted solutions. We survey various current initiatives aimed at studying and confronting disinformation, reflecting the multidisciplinary and evolving nature of the field. We go on to review the current methods and tools for producing and detecting VMD. This includes academic approaches, commercial tools, open-source libraries, and a focus on new AI generative methods that are changing the ways disinformation is created in the first place. This report concludes by looking at potential technological solutions. However, it acknowledges that the current tools for countering VMD are not yet fully capable of addressing the problem, highlighting the need for ongoing development and innovation in this area.

Disinformation is a critical threat of our times, and stakeholders across domains require a comprehensive understanding of the technological aspects underpinning the production and detection of VMD. This understanding is vital not only for developing effective countermeasures but also for fostering better "digital hygiene" practices and critical understanding among our publics. To ensure that our understanding and responses to VMD remain current and effective, the Info Integrity Lab and CRIM are committed to providing regular updates about ongoing work and developments in this ever-evolving field.

Jennifer Irish
Director, Information Integrity Lab,
Professional Development Institute of
the University of Ottawa

Françoys Labonté
Chief Executive Officer
Computer Research Institute of Montreal

The following is a summary report of a comprehensive study produced by the CRIM, in partnership with the InfoLab, entitled "Visual and Multimodal Disinformation: Analysis, Challenges, Solutions", and authored by Marc Lalonde, Houman Zolfaghari, Ph.D., Mohamed Dahmane, Ph.D., Hamed Ghodrati, Ph.D., Gilles Boulianne, Ph.D., Nicolas Rutherford, Richard Pinet.

Edited by Nicolas Rutherford and Marc Lalonde

Evolving Challenges of Digital Disinformation

Visual document manipulation, with the intention of influencing public opinion, has existed for a long time, with some early notable examples dating back to the American Civil War [1]. Since the early 1990s and the advent of Photoshop, image retouching has become more accessible than ever. The rapid spread of disinformation is increasingly likened to an overwhelming flood of inaccurate and misleading content, and is often termed as an "infodemic" due to its extensive and pervasive impact *(Brennen et al., 2021)*. This challenge has been significantly intensified by the proliferation of social media, where content, including disinformation, quickly spreads across various platforms, often eluding sufficient checks. Social media companies, whose revenue models are heavily reliant on high user engagement and content sharing, have faced criticism for not effectively curbing this trend. Complicating matters further are advanced AI tools like ChatGPT, which, while beneficial in many aspects, significantly increase the complexity of verifying the authenticity of information. This is due to the fact that not only do these tools make it easier to create malicious messages, but ChatGPT can also generate an infinite number of variants of the same message, and in multiple languages. These technologies thus open new channels for generating and disseminating disinformation, making the distinction between genuine and false content increasingly challenging.

Visual and multimodal disinformation (VDM) is distinct from traditional text-based disinformation, incorporating images, videos, and audio along with text. This integration of several types of media presents unique challenges, as it can be more persuasive and harder to fact-check than text alone.

The sophistication of VMD production can range from simple "shallowfakes", where texts are paired with out-of-context images or videos, to slightly more sophisticated "cheapfakes", which rely on basic image and video manipulation tools. These manipulations include simple image cropping, inserting parts of other images, altering brightness or color, blurring, changing the speed of vocal delivery in videos, or distorting graphical data. In addition to these, there are "deepfakes", which represent a more advanced level of manipulation. Deepfakes utilize artificial intelligence and deep learning techniques to fabricate or significantly alter audiovisual content, making it appear genuine and convincingly realistic.

[1] A British CDEI report mentions the existence of an image in which a "face swap" involving President Lincoln was carried out at the height of the American Civil War : https://www.gov.uk/government/publications/cdei-publishes-its-first-series-of-three-snapshot-papers-ethical-issues-in-ai/snapshot-paper-deepfakes-and-audiovisual-disinformation#about-this-cdei-snapshot-paper

Societal Impacts of VMD

The impacts of visual and multimodal disinformation are wide-ranging, affecting areas such as politics, the economy, and social cohesion. While the overarching negative effects of disinformation are widely recognized and apply to various forms, the specific impact of visual and multimodal elements, such as VMD, is less documented. Although the mentioned issues of influencing political discourse and decision-making, distorting economic stability, and harming social relations by spreading false information are true for disinformation in general, VMD uniquely amplifies these effects due to its increased persuasive power, making it a particularly potent form of disinformation. A thorough understanding of these impacts is essential for developing effective strategies to counter the unique challenges posed by VMD.

Political Impacts of VMD

VMD significantly affects political institutions. The U.S. Department of Homeland Security and the Australian Strategic Policy Institute have reported risks associated with deepfakes, including online propaganda and its influence on elections and legislative processes[2]. These reports also note the erosion of public trust in institutions due to VMD.

The U.S. Intelligence Community recognizes the geopolitical implications of VMD, with potential adversaries using deepfakes to destabilize the United States and its allies *(Appel & Prietzel, 2022)*. Political advertisements that utilize new imaging and video technologies, despite disclaimers, can have profound psychological effects.

Health Sector Impacts

In healthcare, mis/disinformation can lead to public mistrust in health recommendations from experts, and belief in unproven or fraudulent treatments. The role of visual elements in misinformation and disinformation during the Covid-19 pandemic, particularly in contentious discussions about vaccination, was notably reported by *Brennen et al. (2020)*. They observed that much of the questionable information circulating at the time was accompanied by enticing visuals, emphasizing the impact of VMD in such public health debates.

Financial and Commercial Sector Impacts

The financial and commercial sectors are also vulnerable to VMD. A 2022 report[3] by the U.S. Federal Trade Commission (FTC) highlights the range of online threats to consumers, including deepfakes used to damage the reputation of companies and government agencies, as was the case with Tesla[4] or with China's tax authority[5]. While there's no direct evidence yet of VMD affecting global financial systems, it can severely impact organizations during public relations crises and destabilize economies with weak financial institutions.

Marginal Communities Impacts

VMD can disproportionately harm marginalized communities, targeting women, LGBTQ2+ individuals, people of color, and in some cases even researchers who study online hate and racism *(Paris & Donovan, 2019)*. It can exacerbate vulnerabilities and widen social inequalities, highlighting the need for targeted interventions in these communities.

[2] https://www.aspi.org.au/report/weaponised-deep-fakes

[3] https://www.ftc.gov/reports/combatting-online-harms-through-innovation

[4] https://core.ac.uk/download/pdf/276953172.pdf

[5] https://findbiometrics.com/fraudsters-use-deepfake-biometrics-hack-chinas-taxation-system-040103/

Navigating the Complexities of VMD

Visual and multimodal disinformation can have a more significant impact on the public than textual disinformation. Research shows that content with visual elements, particularly on social media, garners more engagement, as evidenced by the increased clicks, likes, and retweets on tweets containing images *(Cao et al., 2020)*. A key reason for this is that interpreting images usually requires less cognitive effort than text, which often involves language or literacy barriers. Additionally, the sensory experience of viewing an image has a lasting impact on perceptions of credibility, reality, and engagement with the message. This impact is believed to increase with the richness of the message's representation, as seen when moving from text to image, or image to video.

Studies have shown that multimodal messages are generally perceived as more credible than textual ones *(Dan et al., 2021)*. They also evoke stronger emotional responses and are considered to have greater evidentiary value. The heightened emotional impact of visuals can significantly influence behavior, especially in a world dominated by visual information where the potential for image selection and manipulation is intrinsically linked to manipulating perceptions and opinion.

Detecting VMD, however, is challenging both for algorithms and for humans. The general public often struggles to identify disinformation, with a significant portion overestimating their ability to distinguish between legitimate and false information *(Lyons et al., 2021; Languein, 2022)*. Detection complexity arises not only from the content itself but also from the context of its production, including the creator's intent. The common-place, intuitive belief that "images do not lie" leaves citizens ill-equipped to combat VMD on their own, especially as AI advancements make VMD productions look increasingly sophisticated and realistic.

Meanwhile, VMD spreads faster, further, and more deeply on social networks than truthful information *(Morrow et al., 2020; Svahn & Perfumi, 2022)*. Studies on deepfakes have shown they are perceived as more convincing and credible than false news articles, leading to a higher likelihood of sharing on social media. The rapid, near-effortless spread of such news negatively impacts political attitudes and decision-making *(Dan et al. 2021)*.

Recent years have seen significant advancements in AI models for language processing, followed by rapid developments to make these techniques more practical and accessible. Examples include Transformers (Google, 2017), Bert (Google 2018), GPT2 (OpenAI 2019), GPT3 (OpenAI 2020), and ChatGPT (OpenAI 2022). In 2023, many of these models were accessible to the public. The combination of these large language models with other modalities for producing and manipulating images and videos, such as DALL-E 2 (OpenAI, 2022), GPT4 (OpenAI, 2023), and Gen-2 (RunwayAI, 2023), is leading to a rapid transformation in technology and society, with significant implications for the proliferation and refinement of disinformation.

Methods of Producing VMD

Tools and techniques can be divided into two categories: those that transform existing documents and those that generate them.

1 - Document Transformation Tools

Various tools are available for transforming existing documents:

- Face Swapping Software: Programs like FaceSwap and DeepFaceLab, both hosted on GitHub and ranking among the top 250 repositories out of 28 million, enable users to replace one person's face with another's in a video.
- Voice Conversion and Text-to-Speech Software: These tools can alter a person's voice recording or generate new spoken content from text in a specific voice.
- Lip Syncing Software: This technology modifies existing videos to make a person appear to say something different, aligning the mouth movements with the supposed spoken phonemes.
- Appearance Manipulation Software: These programs can alter a person's facial appearance, such as aging or rejuvenating them.
- Virtual Performance Synthesis Software: Entire body movements of a person can be altered or created by transposing the movements of another person.

2 - Generation and Manipulation of Documents

The field of content generation has seen significant growth in recent years. Tools that generate images of non-existent people are used for creating fake social media profiles, thereby avoiding the need to scrape and reuse real people's photos, which could be traced back using reverse image searches. Large Language Models (LLMs) like GPT are also used to create fake social media profiles, combining photos of non-existent people with made-up biographies, interests, and hobbies. These LLMs can also produce accompanying text for images or videos in social media posts. Finally, it should be noted that LLMs can help to generate harmful internet memes, another popular type of visual and multimodal disinformation *(Pramanick et al., 2021)*.

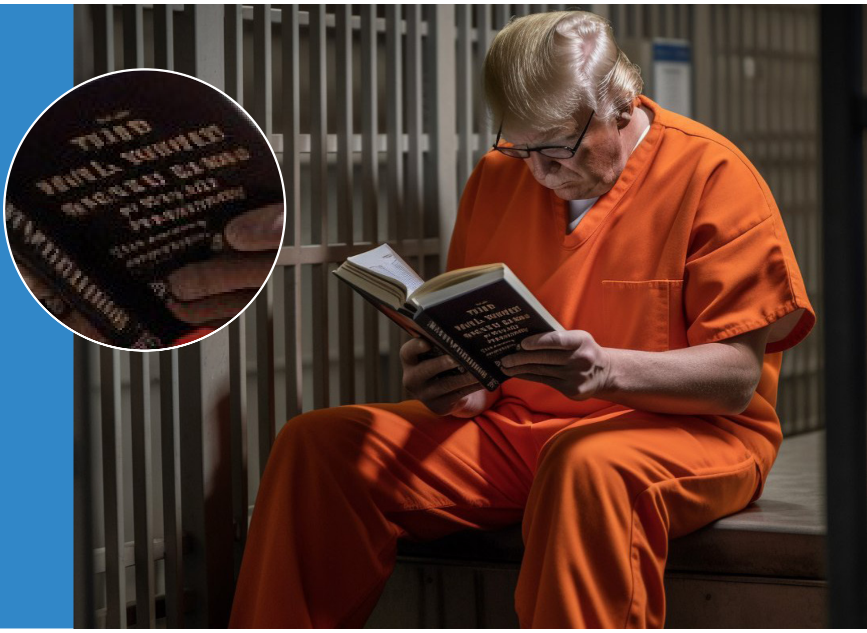
Recent progress in deep learning[6] has improved image synthesis technology. Now, new techniques allow for the creation of completely synthetic content, unrelated to reality, based on text descriptions or text prompts. Some well-known text-to-image models include:

- Generative Adversarial Networks (GANs): Though complex and less popular recently, GANs were among the first generative models developed, capable of synthesizing realistic face images.
- LLM-based Generative Models: These models, trained with millions of image-text pairs, generate hyper-realistic images from simple phrases. Companies like OpenAI (with ChatGPT and DALL-E) and Stability AI (with Stable Diffusion) are leading in this area. Google's Parti model, producing highly realistic images, is another significant contribution.
- Diffusion Models: OpenAI's GLIDE and DALL-E 2, and Google's Imagen, are based on diffusion techniques, advancing the generation of realistic images and improving text rendering within images.

[6] Deep learning is a subset of machine learning where artificial neural networks, algorithms inspired by the human brain, learn from large amounts of data. This approach enables the system to automatically learn complex patterns and make decisions based on its training.

THIS IMAGE IS A HYPER-REALISTIC IMAGE ENTIRELY CREATED BY THE MIDJOURNEY MODEL. THE ARTIFICIAL NATURE OF THE IMAGE IS EXPOSED BY THE UNINTELLIGIBLE TEXT FEATURED ON THE BOOK COVER.

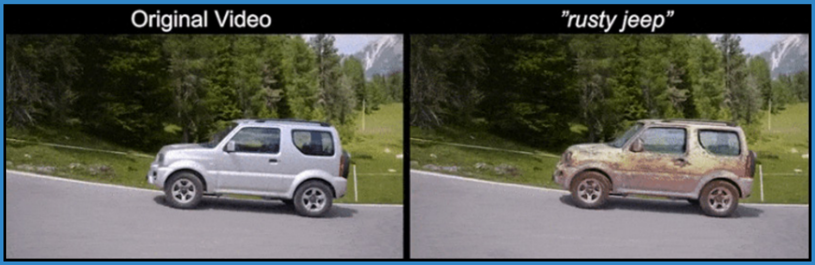
Source: https://twitter.com/EliotHiggins/status/1638187198162821127?s=20



These models not only generate images but also allow users to modify them easily with textual prompts, enabling changes like background alteration, object addition, and style modification without direct pixel manipulation. Companies are increasingly aware of the potential misuse of these technologies, as evidenced by DALL-E 3.0's refusal to process requests involving public figures[7].

The rapid advancements in language models for image generation have also spurred research in video generation. Here, the complexity of maintaining temporal coherence in content and visual rendering has been a challenge, but the latest tools can now create high-resolution videos of substantial length.

Video editing based on text prompts is also evolving, allowing users to describe changes to an object and have the generative model implement them. This field is rapidly progressing, with significant advancements in image quality and frame length. Already, political organizations have utilized AI-generated images for partisan advertising. The ease of use and online availability of these tools, requiring no programming experience, is accelerating the production of dubious content. With these tools' advancements in quality and user-friendliness, the potential for producing questionable content is increasing.



Source: https://www.unite.ai/consistent-ai-video-content-editing-with-text-guided-input/

IN THIS EXAMPLE, USING THE TEXT PROMPT "RUSTY JEEP" LEADS A MODEL TO AUTOMATICALLY ALTER THE APPEARANCE OF THE VEHICLE IN THE ORIGINAL VIDEO, WITHOUT ANY FURTHER HUMAN INTERVENTION.

[7] https://www.theverge.com/2023/9/20/23881241/openai-dalle-third-version-generative-ai

Current Efforts to Counter VMD

Combating visual and multimodal disinformation, and disinformation in general, is undeniably a complex and challenging task. This complexity is reflected in the extensive lists of recommendations found in reports such as the one from the Canadian Public Policy Forum[8] and the thirty-five recommendations by Wardle.[9] Research agrees that no single solution or stakeholder can fully address the challenge, and that effective strategies require a combination of new technologies, organizational practices, and societal changes, as per *Bateman (2020)*. This necessitates a multi-faceted approach, incorporating various types of measures developed and deployed jointly by multiple actors.

For instance, *Helmus (2022)* emphasizes the need to focus on five key areas: developing detection tools, implementing certification standards for audiovisual document authenticity, considering regulatory approaches, promoting intelligence-based approaches like open-source intelligence (OSINT) in journalism, and enhancing media literacy.

AI plays a critical role in filtering the vast amount of messages posted daily on social networks. Its use as a tool to identify online disinformation is on the rise, despite challenges related to biases, performance, and false positives, according to *Svahn & Perfumi (2022)*.

Amidst these varied strategies, the role of social media platforms is pivotal. They are at the forefront of detecting and managing VMD content through their internal systems. Social media approaches to VMD include:

- **Content Moderation**: This process involves reviewing and monitoring the content on online platforms to ensure compliance with the platform's rules. It often combines algorithmic tools, human moderators, and user reports, with the exact mix depending on how the platform is managed *(Morrow et al., 2020)*. Inappropriate content may be removed, made less visible, or lose its ability to generate revenue. However, there are challenges to this approach. Effective moderation requires human reviewers to be adequately compensated, legally protected, thoroughly trained, and supported, including mental health support. Additionally, despite striving for impartiality, human moderators can have biases, and there's a risk of over-relying on automated systems for decision-making, a concern highlighted by the *U.S. FTC (2022)*.

- **Labeling**: This method entails adding visual or textual annotations to user-generated content to give additional context. The concept of enhancing social media posts with fact-checked information is appealing but brings up ethical concerns such as censorship and the impact on free speech. Users' responses to labels are mixed. Some users consider them essential for platforms to provide fact-checking, whereas others perceive them as overbearing and potentially infringing on freedom of expression. The effectiveness of labels depends heavily on their design, including factors like their size and the wording used. Detailed guidelines are often employed to help in creating labels that are impactful and clear, ensuring they convey the intended message without being intrusive or overly directive[10].

- **Contextualization**: This is a form of labeling where extra information, not present in the original post, is added to provide more background. This might include details about the source of an image or information about the author of the post. Contextualization aims to give users a broader understanding of the content, helping them to better assess its credibility. For instance, knowing the origin of an image can clarify its authenticity, while information about the author can offer insights into potential biases or credibility. This approach is meant to support a more informed user base, capable of critically engaging with the content they encounter online.

[8] https://ppforum.ca/wp-content/uploads/2022/01/DEMX-R2.pdf

[9] https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

[10] https://firstdraftnews.org/articles/it-matters-how-platforms-label-manipulated-media-here-are-12-principles-designers-should-follow/

Other approaches to confronting and mitigating VMD can be classified as follows:

Societal Approaches

Enhancing digital and media literacy across the general population, with a focus on youths and journalists, is considered essential for contemporary information consumption.

Journalists, in particular, face a constant influx of information that must be quickly validated, yet they often lack the necessary IT and image/video editing skills for such verifications. The verification of DVM such as deepfakes, where the details that identify an image as a deepfake are often not easily noticeable, poses a significant challenge.

Meanwhile, programs like the News Literacy Project aim to enhance digital literacy from primary education to university level, especially in journalism and communication schools. However, as *Vaccari & Chadwick (2020)* point out, traditional fact-checking models may fall short when it comes to more sophisticated forms of disinformation, such as deepfakes. Efforts directed at enhancing critical thinking and understanding may go further.

In addition to these educational programs, there's a growing recognition of the need for resources targeted at other demographic groups, like the elderly, who are particularly susceptible to disinformation *(Brashier & Shacter, 2020)*. Online platforms and organizations are increasingly developing resources and tools to help these segments of the population navigate the digital landscape more safely and discerningly[11].

Political/Legal Solutions

Addressing VMD through legal measures is often viewed as a less promising approach, primarily because of the robust legal protections surrounding speech and the inherently slow nature of legal proceedings, as highlighted by *Bateman (2020)*. There is also a concern that quickly implemented legal frameworks might inadvertently encourage censorship among social media platforms, as they may over-regulate content to avoid legal repercussions *(Langguth et al., 2021)*.

Furthermore, while existing laws may offer frameworks to tackle the issue, they are often limited in scope and effectiveness due to jurisdictional challenges. These challenges exist both on an international level, where different countries have varying legal standards and enforcement capabilities, and on a national level, where laws may differ across states or regions. This patchwork of regulations creates a complex legal landscape for effectively addressing the spread and impact of VMD.

In addition to jurisdictional issues, the nuanced nature of content like deepfakes—which can straddle the line between legitimate expression and malicious disinformation—makes it difficult for legislation to comprehensively cover all aspects without overstepping into the realm of impinging on free speech. Thus, while legal responses have a role to play, their effectiveness is limited and must be complemented by other measures.

Technological Solutions

Algorithmic solutions for VMD detection are increasingly being proposed, with AI playing a significant role. Machine learning approaches, for instance, analyze image signal inconsistencies to detect cheapfakes, which are created by inserting pixel patches from other images into the original. However, sophisticated tools for creating deepfakes and generating images and videos from text prompts challenge these strategies.

[11] https://www.buffalo.edu/cii/projects/DART.html

Overview of VMD detection strategies

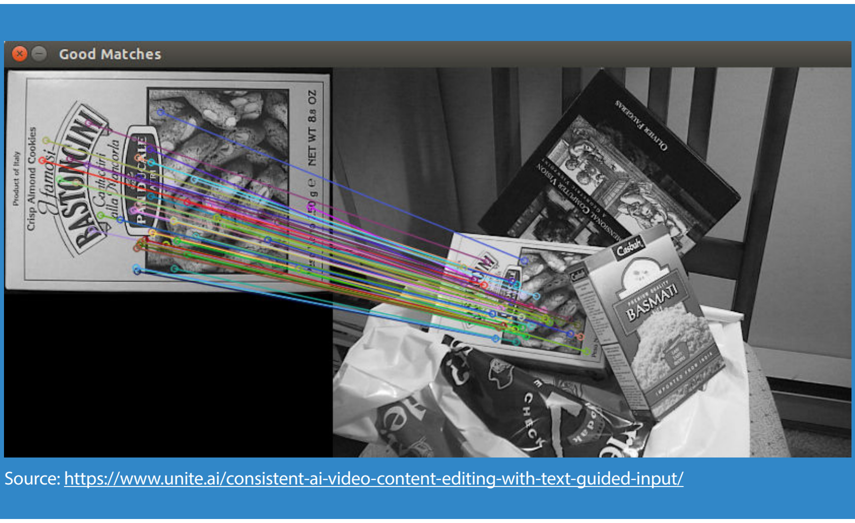
Artificial Intelligence approaches designed to detect VMD are becoming increasingly crucial due to their ability to process the growing volume of content exchanged on the web. These approaches can be categorized into four main categories: Semantic integrity analysis, reverse image search, artifact analysis, detection of generated images:

Semantic Integrity Analysis

This kind of analysis involves validating the semantic coherence of an image. For example, confirming the identity of people in an image can help detect face swaps[12] or historical inaccuracies. A more complex task is the text-visual semantic integrity analysis, a new field leveraging advancements in multimodal neural networks capable of processing both images and text. The goal is to ensure that the textual message aligns semantically with the visual content of the image, detecting scenarios where a real, unaltered image is taken out of context and paired with misleading text. Technologies like PROVES[13] *(Xie et al., 2022)* are being developed to certify and verify the semantic information of an image while allowing for simple editing operations like cropping and color adjustments. DARPA's SemaFor program director acknowledges the difficulty in aligning all semantics correctly, from a news story's text to the accompanying image and the elements within the image[14]. The ELSA project also focuses on semantic analysis, exploring novel ways of understanding and detecting fake data through machine learning approaches that blend syntactic and perceptive analysis.

Reverse Image Search

Reverse image search methods are designed to find similar images based on visual content. They are more effective than simple keyword searches in image metadata, which are often removed by social media platforms during upload. The basic idea is to construct a digital fingerprint from the "stable" visual features of the searched image and then compare this fingerprint to those in an image database. This process typically involves automatically identifying perceptually important keypoints in an image and attaching a digital signature to each point based on the surrounding pixels. Services like Google Image Search and Tineye continuously crawl the web to collect new images, calculate their fingerprints, and store them for comparison against billions of stored fingerprints.



Source: https://www.unite.ai/consistent-ai-video-content-editing-with-text-guided-input/

[12] https://par.nsf.gov/servlets/purl/10346314

[13] https://openaccess.thecvf.com/content/WACV2022/papers/Xie_PROVES_Establishing_Image_Provenance_Using_Semantic_Signatures_WACV_2022_paper.pdf

[14] https://www.darpa.mil/news-events/2021-03-02

Artifact Analysis

Manipulating digital images leaves traces that can be used to detect falsified images. Artifact analysis, a forensic image verification technique, considers artifacts generated by post-capture operations and those induced during acquisition by the sensor itself. These artifacts provide useful information for establishing an image's authenticity. The analysis encompasses artifacts caused during image acquisition (by the lens and electronic sensor), physical scene integrity (inconsistencies like incompatible shadows), generative model artifacts, and visual artifacts (unnatural visual inconsistencies). Research on detecting these artifacts is necessary, as such solutions have significant detection potential because they don't rely on explicit analysis or assumptions about the image content.

Detection of Generated Images

Generative models like GANs and text-to-image models can create images and videos from text prompts. To differentiate synthetic images from real ones, analysis methods attempt to detect artifacts produced during image creation by these generative models. It's also possible to trace back the class of generative model used to create a synthetic image based on a detailed analysis of artifacts. However, there's a need for universal approaches capable of detecting images generated by any model. Progress is being made with GANs, but more work is needed to include other generative models. Public datasets like CIFAKE and COCOFake, inspired by popular computer vision datasets, are crucial for developing these methods.

The effort to counter visual and multimodal disinformation encompasses other strategies as well, ranging from authentication and provenance to digital watermarking and beyond:

Authentication/Provenance

The Content Authenticity Initiative (CAI) is developing infrastructure to ensure the traceability of visual documents (images, videos) from capture to viewing, providing users access to metadata and a list of modifications applied. This "secure end-to-end system" could extend to generative models, allowing creators to disclose the use of Generative AI in content creation. The Coalition for Content Provenance and Authenticity (C2PA) is also leading the creation of technical standards for media document source/provenance certification. However, the success of this initiative depends on widespread industry adoption, including sensor manufacturers, software developers, and heavy users like media outlets. Key industry players like Adobe, Microsoft, and Intel are actively involved in this initiative.

Digital Watermarking

Digital watermarking embeds a visually imperceptible signature in an image, detectable via software. Long used in intellectual property management, it's becoming a vital part of provenance infrastructure, being robust against benign manipulations like compression or contrast adjustments, but fragile against deepfake manipulations. AI is also contributing to watermarking solutions, like the University of Chicago's Glaze, protecting images from style copying, and MIT's PhotoGuard, which injects imperceptible disturbances to immunize images against AI-driven editing manipulations. However, the effectiveness of digital watermarking is limited by the vast amount of non-watermarked material on the internet and its ineffectiveness for images used out of context.

Researcher/Developer Responsibility

With the increasing social pressure, researchers and developers in AI are being urged to consider the potential misuse of publicly shared models on platforms like GitHub. Big industry players can host their models behind access-controlled APIs, unlike smaller entities. Potential interventions include injecting signatures into pre-trained models to detect their involvement in DVM production and making datasets "radioactive" to verify if they've been used to train models suspected of disinformation production. The success of these measures hinges on widespread developer community participation and robust tracking solutions for models and datasets.

Disinformation and Speech Technologies

Speech technology advancements have significantly increased the impact of audio disinformation. Voice conversion and text-to-speech synthesis technologies can create realistic videos with fabricated speech. Recent developments have removed many technical constraints, such as data quantity, allowing high-quality voice synthesis from just a few seconds of voice recording. Free and easy-to-use tools now make these technologies accessible to anyone. Voice identity theft research, especially through the biennial ASVSpoof campaigns, focuses on detecting whether a voice sample is authentic. The challenge is to perform this detection blindly, without knowledge of the synthesis techniques used or exposure to them during training, and to remain reliable in noisy environments. Although impressive in speaker verification, detection systems currently have less success with deepfakes. The future of speech disinformation technology aims for even more realistic results, introducing subtle emotions and non-verbal elements, making detection harder while production becomes easier.

The Evolving Challenge and Need for Multidisciplinary Responses

The landscape of visual and multimodal disinformation (VMD) is vast and continuously evolving. It presents significant challenges across various sectors, from politics and public health to finance and social cohesion. The increasing sophistication of tools for creating and manipulating digital content, especially with advancements in AI, makes distinguishing between authentic and falsified content increasingly difficult. This complexity is further compounded by the rapid dissemination of information via social media and other digital platforms, often outpacing the capabilities of current detection and countermeasures.

Countering VMD and its effects requires a multifaceted approach, encompassing technological, legal, educational, and societal strategies. While AI and machine learning provide promising avenues for detecting and mitigating VMD, they also raise concerns about biases and ethical use. Legal frameworks must balance effective deterrence against misuse with the protection of free speech. Educational initiatives, particularly in media literacy, are crucial in equipping the public to critically evaluate the information they encounter.

Collaboration among various stakeholders – including governments, tech companies, academia, and civil society – is essential in developing comprehensive and adaptable solutions to this complex problem. As technology continues to advance, staying ahead of the curve in the fight against disinformation will require continuous innovation, vigilance, and a commitment to upholding the integrity of information in our digital age.

It is against this backdrop that CRIM and the uOttawa Information Integrity Lab will continue their collaborative study of visual and multimodal disinformation and provide regular publicly-available updates. Further, we commit to advance the development of tools and methodologies to identify and counter the significant threats of mis- and disinformation to our political discourse, social cohesion and economic and national security.

Bibliography

Appel, Markus, and Fabian Prietzel. "The detection of political deepfakes." Journal of Computer-Mediated Communication 27, no. 4 (July 2022): zmac008. https://academic.oup.com/jcmc/article/27/4/zmac008/6650406.

Bateman, Jon. Deepfakes and synthetic media in the financial system: Assessing threat scenarios. Carnegie Endowment for International Peace., (2020).

Brashier, Nadia M, and Schacter, Daniel L. "Aging in an Era of Fake News." Current directions in psychological science vol. 29,3 (2020): 316-323. https://pubmed.ncbi.nlm.nih.gov/32968336/

Brennen, J. Scott, Felix M. Simon, and Rasmus Kleis Nielsen. "Beyond (Mis)Representation: Visuals in COVID-19 Misinformation." The International Journal of Press/Politics 26, no. 1 (January 2021): 277-299. https://journals.sagepub.com/doi/10.1177/1940161220964780.

Cao, Juan, Peng Qi, Qiang Sheng, Tianyun Yang, Junbo Guo, and Jintao Li. "Exploring the Role of Visual Content in Fake News Detection." In Disinformation, Misinformation, and Fake News in Social Media, edited by Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu, 141-161. Lecture Notes in Social Networks. Cham: Springer International Publishing, (2020). https://link.springer.com/chapter/10.1007/978-3-030-42699-6_8.

Dan, Viorela, Britt Paris, Joan Donovan, Michael Hameleers, Jon Roozenbeek, Sander van der Linden, and Christian von Sikorski. "Visual Mis- and Disinformation, Social Media, and Democracy." Journalism & Mass Communication Quarterly 98, no. 3 (September 2021): 641-664. https://journals.sagepub.com/doi/10.1177/10776990211035395.

FTC. "Combatting Online Harms Through Innovation; Federal Trade Commission Report to Congress." s.d. https://www.ftc.gov/reports/combatting-online-harms-through-innovation.

Helmus, Todd. Artificial Intelligence, Deepfakes, and Disinformation: A Primer. RAND Corporation, (2022). https://www.rand.org/pubs/perspectives/PEA1043-1.html.

Langguth, Johannes, Konstantin Pogorelov, Stefan Brenner, Petra Filkuková, et Daniel Thilo Schroeder. « Don't Trust Your Eyes: Image Manipulation in the Age of DeepFakes ». Frontiers in Communication 6 (24 mai 2021): 632317. https://www.frontiersin.org/articles/10.3389/fcomm.2021.632317/full.

Languein, Adela. "Combatting Visual Misinformation on Social Media: A Review of Strategies and Concepts." Master's thesis, Concordia University, (2022). https://spectrum.library.concordia.ca/id/eprint/990735/1/Languein_MA_S2022.pdf.

Lyons, Benjamin A., Jacob M. Montgomery, Andrew M. Guess, B. Nyhan, J. Reifler. Overconfidence in news judgments is associated with false news susceptibility. Proceedings of the National Academy of Sciences of the United States of America, 118, (2021) Article e2019527118. https://www.pnas.org/doi/full/10.1073/pnas.2019527118

Morrow, Garrett, Briony Swire-Thompson, Jessica Polny, Matthew Kopec, and John Wihbey. "The Emerging Science of Content Labeling: Contextualizing Social Media Content Moderation." SSRN Electronic Journal, (2020). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3742120.

Paris, Brit and Joan Donovan. "Deepfakes and cheap fakes: the manipulation of audio and visual evidence". Data & Society. (2019). https://datasociety.net/library/deepfakes-and-cheap-fakes/

Pramanick, Shraman, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, et Tanmoy Chakraborty. « Detecting Harmful Memes and Their Targets ». In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2783 96. Online: Association for Computational Linguistics, (2021). https://aclanthology.org/2021.findings-acl.246/.

Svahn, Mattias, and Serena Coppolino Perfumi. "Towards a Positioning Model for Evaluating the Use and Design of Anti-Disinformation Tools." JeDEM - EJournal of EDemocracy and Open Government 14, no. 2 (December 23, 2022): 109-129. https://jedem.org/index.php/jedem/article/view/746.

Vaccari, Cristian, and Andrew Chadwick. "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News." Social Media + Society 6, no. 1 (January 2020): 205630512090340. https://journals.sagepub.com/doi/10.1177/2056305120903408.

uOttawa

Laboratoire sur l'intégrité de l'information

Information Integrity Lab

CRIM