If saliency cropping is the answer, what is the question?

Vinay Uday Prabhu
UnifyID Labs
vinay@unify.id

Abeba Birhane
University College Dublin & Lero
abeba.birhane@ucdconnect.ie

Abstract

In this paper, we critique the saliency based image auto-cropping technology used by platforms such as Twitter from both a socio-technical perspective, and a machine learning perspective. We examine the discordant literature surrounding what saliency entails and the incoherent use of the concept. We then outline the various expectations, some of them mutually contradictory, placed on the technology. Following our experiments, we demonstrate specific vectors of vulnerability that expose the brittleness of the technology and the amplification of the risk of male-gaze-cropped images of women, perpetuating misogyny. We conclude by recommending the removal of automated saliency based cropping of user-uploaded images on platforms where the technology might be currently integrated into.

1. Introduction: The S in SIC

Saliency based image cropping (SIC) today is ubiquitously used to algorithmically crop user-uploaded images on most major digital technology and social media platforms, including Twitter [21], Adobe [3], Google [48], Microsoft [50], Filestack [36] and Apple [7] (See Appendix A for examples from Facebook and Google-Feed). Recently, this technology came under scrutiny as Twitter users shared collective frustration with the discriminatory nature of the technology [13].

Although automated image cropping technology is ubiquitously integrated into major platforms, it often operates under the radar where its existence is hidden from the users. Typically, SIC [4] entails two phases[1]: saliency estimation and image cropping. In the *saliency estimation* phase, the weights or *noteworthyness* of each of the constituent pixels in an image are estimated to generate a binary mask or a continuous-valued heatmap of pixel-wise *importance*. This is then processed by the image cropping part of the technology that enacts a segmentation policy that seeks to retain the higher weighted *noteworthy* pixels while discarding the pixels deemed *less salient*. In this paper, we examine the landscape of brittleness and inconsistencies underlying every part of the pipeline that births this technology from the problem-framing phase, to dataset curation and model building. We also demonstrate how this framework can introduce male gaze effects in the cropped images of women in the presence of artifacts such as text and corporate logos.

[1]See [20] for an end-to-end cropping solution with inbuilt saliency map generation network and a saliency generator aesthetic area regression network

2. The roots

Saliency estimation techniques have roots in two traditions: *research* and *praxis*. The research side emanates from academic fields such as neuroscience and psychophysics. Within these traditions, the study of saliency in human visual systems is considered central to understanding mechanisms for human attention and perception. This often encompasses study of the human vestibular and oculomotor systems, examining methodological facets of eye movement recording and measurements, analysis of fixational eye movements (such as microsaccades), as well as analysis of foveation. This line of research includes both theoretical (modeling based) and experimental investigations. One of the motivating factors for real-world application of this line of research is the claim that it empowers people with vision impairment or low vision to more easily interact with the computational device that incorporates the eye-tracking software (for example, [50])

The second praxis-oriented approach emanates from the by-lanes of Computer Vision. Under the banners of saliency cropping [4], image re-targeting [49], automated thumbnail cropping/generation [41] and Salient Object Estimation [6], this school seemingly derives motivation from the oculomotor functioning tradition and operationalizes the ideas into real world applications. Such applications include image and video compression, video summarization, semantically consistent thumbnail generation, auto-collage generation, image quality assessment, rendering a consistent UI experience (especially on social media timelines as seen, for example, in [21]), and faster content-based image retrieval.

While saliencey estimation techniques can be traced in two broad traditions, there exists a disconnect between these

roots. As we detail below, the concept of saliency rests on shaky grounds and the expectations placed on SIC are ill-defined and inconsistent. We argue that implementation of technologies such as SIC undermines human agency, where the end user is left with little control over how their images appear (or how they see others' images) on social media platforms. Given the various problems we highlight in this paper including inconsistent use of saliency, ill-defined expectation, extreme brittleness of the technology itself and its tendency to amplify the male gaze, we argue that the application of SIC is not only unnecessary but also harmful and offensive.

2.1. Vague definitions

Cataloguing the proposed definitions that researchers present in their *saliency* estimation work, it quickly becomes apparent that the concept is nebulous and fuzzy with no general consensus on how it is understood or used. Depending on the specific publication, saliency could mean *subjective perceptual quality* [15], *visual contrast* [19], proxy for *important regions* in an image [4], *interestingness* [36] alongside terms such as *visual representativeness and foreground recognizability* [14] and lastly *intentionality* [27].

The lack of coherent and somewhat commonly agreed understanding of the term has previously been pointed out. For example, Vaquero et al. [49] have pointed out that: *there is no clear definition or measure to date as to the quality of I' (a re-targeted image) being a good representative of I (the original image)*. More recently, in [27], while describing the task flow of the constituent `ImportAnnots` interface, Newman et al. state that *"Participants are presented with a series of images one at a time and are asked to annotate the most important regions"* adding that *"There are no definitions of what should be considered important*". Not only do we find inconsistent understanding of what saliency means, but also the concept is tied to woolly notions such as "important regions in an image" and "interestingness", concepts that are subjective, vague and context dependent.
With this backdrop, we now list of all the expectations that are piled on a unicorn SIC module in the context of a real-world deployed computer vision pipeline.

2.2. Ill-defined expectations

Upon parsing through the set of real-world applications of SIC, we identify five main categories of expectations, some of which are mutually contradictory depending on context.

- **Automated Aesthetics**: in this context, SIC is presented as a panacea that can programmatically generate aesthetic thumbnails that are less adversely affected by *content blind* downsizing artifacts such as *Moire aliasing*, *blurring*, and *edge halos*. This expectation is found in works such as [2] and [30] as well as the CUHK Image Cropping dataset [52] and the Aesthetic Visual Analysis (AVA) dataset [25].

- **UI Consistency**: according to this narrative, SIC is intended to potentially amplify the voracity of social media consumption and user engagement on platforms such as Twitter [21]. This sentiment is encapsulated by the statement from the social media platform itself *"The photos in your timeline are cropped to improve **consistency** and to allow you to see more Tweets at a glance."*

- **NSFW filtering**: the expectation here mandates that the SIC algorithm needs to be able to crop out the Not safe For Work (NSFW) regions of an image thereby minimizing the chance of incidental exposure to objectionable imagery by a user scrolling through their timeline[2].

- **Photobomb defusal**: this expectation can be found in Shan et al. [39] and comes from the assumption that images have a '*main subject*' that needs to be focused on and serendipitously present '*distracting (background) people*' that need to be cropped out.

- **Centering of textual graphics and logos**: the saliency cropping algorithm in this scenario is expected to pay heed to textual content and not crop it away in favor of background visual objects [9, 21]. With regard to twitter, both text and graphics occur in user-uploaded images in the form of corporate logos in the background as well as synthetically added one (such as links to peoples Instagram handles or brand slogans). In fact, as part of one of the *"5 tips for getting your brand noticed on Twitter"* [9], the platform itself encourages strategic inclusion of brand-related text in the ad-tweets[3]. This expectation merits further explanation and we provide further elaboration in the experimentation section.

The above listed nebulous and also mutually contradictory expectations lead to questions such as should saliency favor typicality or anomaly. In other words, should the saliency estimation module in a real-world production pipeline emphasize the *anomalous regions* of an image or the *typical regions*? What about the exposure time and the associated temporality facet of saliency used in the dataset to train the model and the exposure period of the training images? In the context of multi-duration saliency modeling, research has

[2] See *Debunking Twitter myths* [46] , the details surrounding the `filter:safe` operator [45] and the document on *Non-consensual nudity policy* [47] for further details

[3] Further, the statements: *"Saliency prediction networks are able to detect puppies, faces, **text**, and other objects of interest."* and *"In general, people tend to pay more attention to faces, text, animals, but also other objects and regions of high contrast"* appears in the very blog post that announced the use of SIC on the OSN[21].

uncovered complex phenomena such as Inhibition Of Return (IOR) [31], boomerang patterns [10], attentional boost [28] and attentional push[12] that all point towards a simple realization that *"What jumps out in a single glance of an image is different than what you might notice after closer inspection"* [10]. Appendix B furthermore addresses these inconsistencies. With this background, we now examine various failings of SIC and present the supporting empirical investigations in the following section.

3. Experiments

In this section, we present the experimental facet that demonstrates the specific vectors of failure of SIC using the model deployed on Twitter as an example. These further illustrate the downstream effects of vague definitions and contradictory expectations listed above.
By and large, we observed that the perceived *miscroppings* were on account of either *separated saliency barycenters* or the *saliency mismatch* cause. The separated saliency barycenters cause pertains to cases where there are multiple spatially separated patches of high-saliency pixels in an image and the cropping algorithm eventually picks one of these patches, typically entailing a person's face, thereby leading to the cropping out of the other(s). The *saliency-mismatch* cause pertains to the case[4] where there are parts of the image that *ought* to be salient from the onlooker's perspective but the saliency estimation algorithm under-values the saliency of those parts. Below we explore each of these causes by conducting a set of experiments[5].

3.1. The asymmetric erasure antics of SIC with 3×1 grid images

The 3×1 image grid format with multiple saliency barycenters is a flagship example of poor system design where, by its very volition, the SIC model is mandated to erase one of the two constituent images and thus behaves as a Binary Asymmetric Erasure Channel (BAEC)[6] (See Figure 13). In the section, we investigate this modality with two experiments that we term *Face v/s Face* and *Face v/s Text*.

3.1.1 Face v/s Face

In order to examine this modality, we adopted the format of the images that highlighted the shortcomings of SIC in [13] and created a template to generate 583×3000 sized 3×1 grid images of the format: $I = [F_i, W, F_j]^T$, where W is an all-white (or all-black) square image and $F_i,\ F_j$ are the constituent face images drawn from the two categories being analyzed.

In lieu of merely populating F_i and F_j with appropriate sized images sampled from some *standard* computer vision datasets, passing it through the SIC model being analyzed and estimating *fairness ratios*, we undertook another path: To demonstrate the utter brittleness of this framework by presenting hand-crafted demos supplanted with a simple numerical recipe that demonstrates just how trivial it is to achieve any arbitrary ratio that an attacker has in mind. This brittleness and shredding of the verisimilitude of scientific rigor, we believe provides for a stronger motivation for transferring the agency of cropping back to the user.

(A) Hand-crafted method: In order to generate the hand-crafted inputs, we used the Gradio-hub interactive portal available publicly (as of 25-5-2021) at: `http://saliency-model.gradiohub.com/`. This portal combined the *Contextual Encoder–Decoder Network for Visual Saliency Prediction* [17] with a cropping policy which entailed a sliding window with a fixed aspect ratio (16:9) that maximizes the window-wise sum of saliency scores. The key feature was the integrated TOAST-UI image-editor[7] that allowed for easy cropping, rotation, free-hand drawing, image-filtering and text-addition to the images. Using this editor, we were able to flip which image was chosen post-cropping with regards to the small subset of images in the initial exploratory experiment presented here [29]. While varying the level of *pixelation* on the individual face images worked in some cases, a combination of one or more of crop, flip, brightness altering, filtering and tinting worked in the other cases. In figure 9 of Appendix C, we demonstrate this with the now-infamous *Obama-McConnell* image from September 2020 [13] and expand on the ramifications of this result.These small scale experiments revealed that given access to an off-the-shelf image editor, it was a trivial exercise to come up with an image transformation that would flip the survival ratio obtained with the original images that are hand-generated synthetic images to begin with.

(B) Numerical recipe: Here, we use ideas from the adversarial machine learning literature to formalize a simple programmatic recipe to synthetically generate the input 3×1 images that can flip the survival verdict of the original image. To this end, we harness the *universal adversarial image* idea from [24] and present the specific details in Appendix D.

At this juncture, it is worth emphasizing that unlike other computer vision deployment ecosystems, the machine learning models deployed on social media platforms such as Facebook and Twitter that bank on creative user digital uploads do not enjoy the option of invoking the Out-Of-Distribution (OOD) failure [26] justification and any computer-vision-

[4] One could argue that this maps to the bottom-up saliency and top-down scrutiny schism explored in works such as [23, 42] which is outside the scope of ideas discussed here.

[5] In order to generate these results with Twitter's SIC model, we created the `@cropping_bias` account with the consent of Twitter.

[6] See `https://www.ti.rwth-aachen.de/teaching/InformationTheory/ws1819/data/Lecturenotes/Lecture_10_2019.01.15_Handouts.pdf`

[7] `https://ui.toast.com/tui-image-editor`

tinted fix will remain eternally vulnerable to ridicule and harm, much akin to what happened with the viral tweets [13].

3.1.2 Face v/s text

In the *Face v/s text* experiment, we investigated what would happen if one of the images in the grid was replaced by text. This specific vulnerability analysis also serves to question the very veracity of the motivation behind using saliency based cropping: that, with enough training data and "state-of-the-art" neural network architectures, we can *learn* the *natural* weightage that the *average human gaze* places across different modalities of information in an image such as *faces, text, animals, but also other objects and regions of high contrast* [21]. As in the sub-section above, we use the same 3×1 image grid structure. Here, the top image is drawn from the set of images included in the Wikipedia article titled *Lists of African Americans*[8], the middle image is a blank white image and the bottom image consists of the text `random text` (See Figure 1 (a),(b) and (c) for examples). The pseudo-code for generating these images is covered in

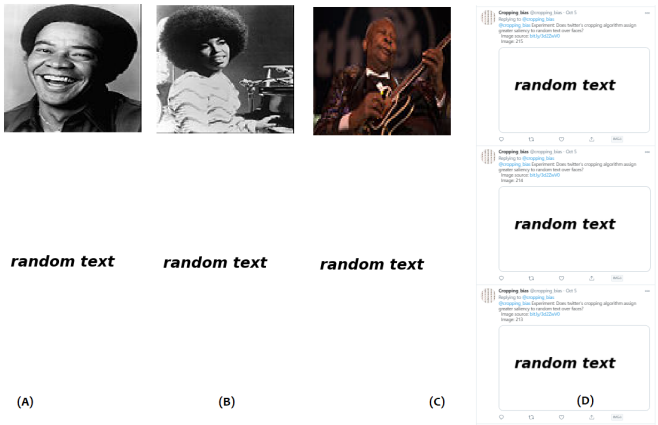


Figure 1: Choosing text over face?

Appendix-E.4. When these images were uploaded to twitter, the result is captured in Figure 1 (D) where, every single SIC cropped image cropped out the face in favor of `random text`. Of the 235 images that we tried with the exact same template, all the 235 images were met with the same fate[9]. As stated previously, it is trivial for us to flip the 235:0 ratio to 0:235 ratio in favor the faces used and in this case we flipped the ratio by merely changing the font or the font size.

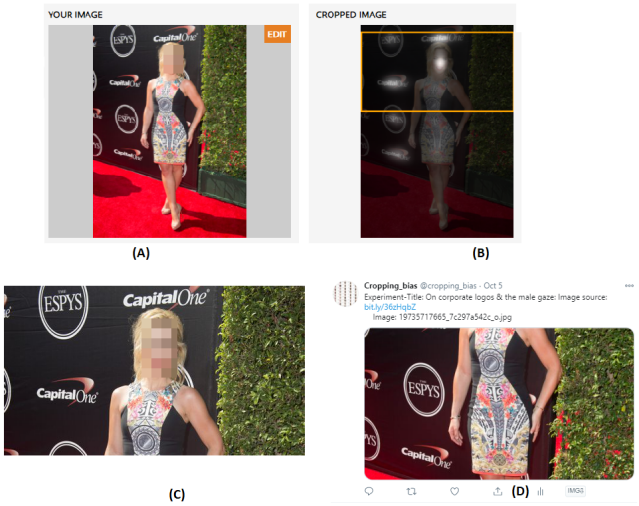


Figure 2: Picture of a celebrity from the Flickr album titled *The 2015 ESPYs*.

3.2. On male-gaze, misogyny, logos and the red-carpet:

We furthermore examined SIC in the context of real world images with textual content (see Figure 2 (A) of a photograph of a celebrity from the CC-BY licensed Flickr album titled *The 2015 ESPYs*[10]). As can be seen, this real world image from a red carpet event has textual content and corporate logos in the background. This results in saliency hot-spots when the image is processed through an academic model [17] (as shown in Figure 2 (B)). Gradio-hub's inbuilt sliding window cropping policy (See section 3.1.1) results in the cropped image shown in Figure 2 (C). However, when the image is passed through the SIC model deployed on twitter, the resultant cropped image obtained (seen in Figure 2 (D)) *focuses* on the person's body in a way that incorporates the *male gaze* [11, 37]. When we tested this hypothesis across different images from the same album, we observed that this *male gaze*-like cropping was not a one-off happenstance and Figure 3 contains a gallery of example images from this *2015 ESPYs* album that met with the same fate on twitter.

As seen in Figure 4, these male-gaze-like (MGL) croppings can result from naturally occurring text and logos in the background as well as user-uploaded text (like instagram handles) and Chyron text, written text that often accompanies television coverage. These MGL artifacts in the cropped images, irrespective of whether it was intentional or serendipitous on account of coincidentally adversarial text, instagram

[8] https://en.wikipedia.org/wiki/Lists_of_African_Americans

[9] https://twitter.com/cropping_bias/status/1313257014940712962?s=20

[10] Sourced from https://www.flickr.com/photos/disneyabc/albums/72157653604309063

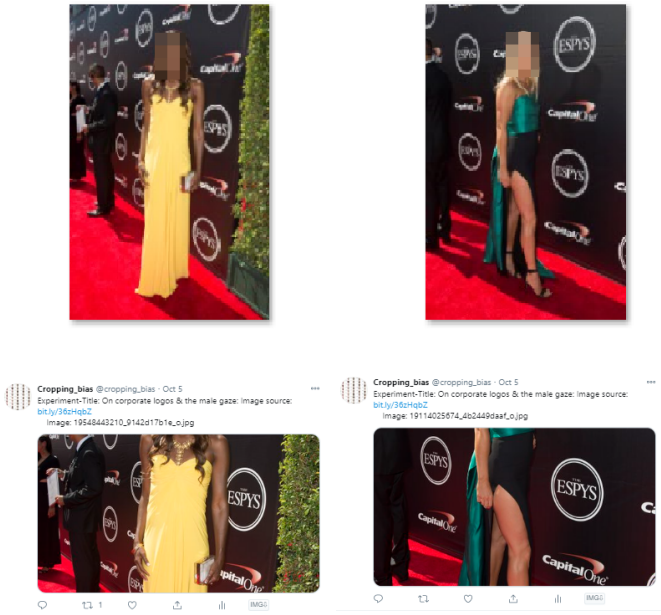


Figure 3: Male gaze, corporate logos and the red carpet

handle links, *Chyron-text* or corporate logos, are evident of the core *capitalist* values and drives at the centre of this technology although the SIC is often presented a technology designed and implemented to serve the individual user. Like any other socio-technological tools and infrastructures that integrate into day-to-day life, surveillance capitalism forms the bedrock of SIC [53]. Twitter algorithm's preference for textual content over other images is the overt manifestation of the underlying building block – that is surveillance capitalism – of almost all tech development. The heightened attention on logos puts businesses, industries, and corporations as the main stakeholders for which SIC technology caters for. This, furthermore, introduces new avenues for concern as corporate logos and MGL croppings come head to head for competition for attention.

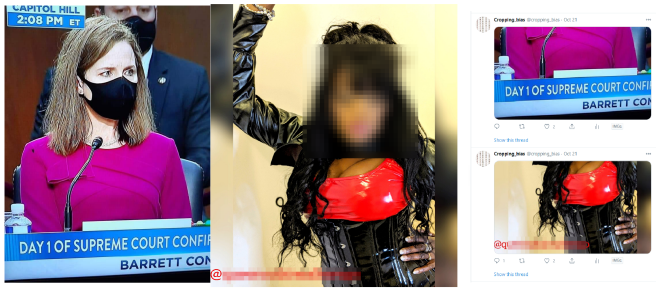


Figure 4: Male gaze like artifacts from the text at the bottom of the image

Furthermore, the absence of a post-cropping *sanity check* creates MGL artifacts even in the absence of any logo or textual content. When there are many faces in an image (see Figure 5) either in the form a real world grid or a synthetically generated one, the resulting *lecherously centered* crop can be located in such a way that it bears MGL artifacts.

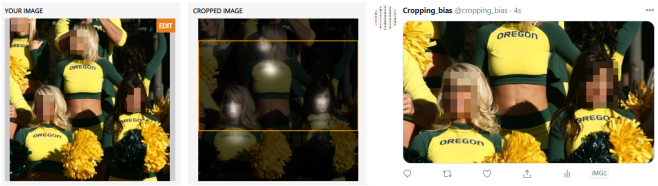


Figure 5: Male gaze like artifacts when there are many faces in an image either in the form a real world grid or a synthetically generated one.

4. Conclusion

In this paper, we critique the Saliency Image Cropping framework. We argue that its real-world integration is done in an insouciant fashion with little attention placed on either the contradictory expectations placed on the technology or to the questionable premise – that there exists a *universally valid* exposure-time invariant notion of saliency distribution in images. The technology, furthermore, rests on the problematic assumption that neural networks can predict this distribution *well enough* for real-world deployment at scale. We demonstrated the haziness and confusion surrounding the use and understanding of saliency in computer vision literature and the contradictory expectations placed on it. We found specific vectors of vulnerability that show the technology is not only brittle but also amplifies the male gaze on images of women, perpetuating misogyny. It is important to emphasize that this work is focused on the critique of the very technology, its birth and its large-scale integration and ***not*** on the specific idiosyncrasies observed on a specific platform such as Twitter (See Appendix A for similar issues we observed with Facebook and Google-Feed). At this juncture, we would like to exhort the creators, engineers, and product managers responsible for real-world integration of this technology to ask themselves: ***If SIC is the answer, what is the question?***

Finally, we conclude by voicing our concerns pertaining to a classic Goodhart's law gaming "fix". During our investigation, we saw how trivial it is to cherry-pick a toy academic dataset and strategically chose 3×1 grid dimensions and tactically claim fairness on account of equal erasure rate or demonstrating that the erasure rate for the minoritized class is measurably better than the other. Such claims, we argue will be subjected to derision when users try out different ratios and trivially find a regime of distortion where

all the fairness claims are laid bare. We were unable to reconcile with the idea that giving the agency back to the user to custom-crop the image and providing a preview of the same before posting would really result in cataclysmic outcome based on any reasonable metric that the platforms measure. We argue that the observed issues were not on account of the model not being trained on a large enough dataset or a caricatured shortcoming unique to CNNs [11] but rather a fundamental problem with SIC and hence motivate it's removal and handing the agency back to the user.

Acknowledgement

We would like to thank Abubakar Abid, Ali Abdalla and Dawood Khan from Gradio [1] for their technical inputs and efforts in setting up the Saliency-model Gradio hub instance. Many thanks to Alexander Kroner, the main author of the *Contextual Encoder–Decoder Network for Visual Saliency Prediction* paper [17] for sharing insights into the modeling process, the landscape of datasets used and for reviewing this paper. Lastly, we would also like to thank Darrell Owens, Jutta Williams and Nick Matheson for the initial discussions, Thomas Laurent and the anonymous reviewers of the `BeyondFairCV` workshop for their useful feedback. This paper is a work in progress. During the review period of this submission, we gathered that there was an Arxiv submission from Twitter (found here: https://arxiv.org/abs/2105.08667) that has addressed SIC as well, which is beyond the purview of the camera-ready version presented here. **Funding**: Abeba Birhane was supported, in part, by Science Foundation Ireland grant 13/RC/209_2.

References

[1] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*, 2019. 6

[2] Abraia. Tutorials: Image optimization. https://abraia.me/docs/image-optimization/, Oct 2020. (Accessed on 11/02/2020). 2

[3] Adobe. Adobe research » search results » cropping. https://research.adobe.com/?s=cropping&researcharea=&contenttype=&searchsort=, December 20202. (Accessed on 12/05/2020). 1

[4] Edoardo Ardizzone, Alessandro Bruno, and Giuseppe Mazzola. Saliency based image cropping. In *International Conference on Image Analysis and Processing*, pages 773–782. Springer, 2013. 1, 2

[5] Christer Borell. The brunn-minkowski inequality in gauss space. *Inventiones mathematicae*, 30(2):207–216, 1975. 13

[6] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Computational visual media*, pages 1–34, 2019. 1

[7] Apple Developer Documentation. Cropping images using saliency. https://developer.apple.com/documentation/vision/cropping_images_using_saliency, Jun 2019. (Accessed on 10/19/2020). 1

[8] Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. In *Advances in neural information processing systems*, pages 1178–1187, 2018. 12

[9] Erin Fishman. 5 tips for getting your brand noticed on twitter. https://business.twitter.com/en/blog/5-tips-for-getting-your-brand-noticed-on-twitter.html, Oct 2020. (Accessed on 10/20/2020). 2

[10] Camilo Fosco, Anelise Newman, Pat Sukhum, Yun Bin Zhang, Nanxuan Zhao, Aude Oliva, and Zoya Bylinskii. How much time do you have? modeling multi-duration saliency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4473–4482, 2020. 3, 10

[11] Amanda M Friz and Marissa L Fernholz. The male gaze in the medical classroom: Proximity, objectivity, and objectification in "the pornographic anatomy book". *Women's Studies in Communication*, pages 1–25, 2020. 4

[12] Siavash Gorji and James J Clark. Attentional push: A deep convolutional network for augmenting image salience with shared attention modeling in social scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2510–2519, 2017. 3, 10

[13] Alex Hern. Twitter apologises for 'racist' image-cropping algorithm | twitter | the guardian. https://www.theguardian.com/technology/2020/sep/21/twitter-apologises-for-racist-image-cropping-algorithm, Sep 2020. (Accessed on 10/20/2020). 1, 3, 4, 10

[14] Jingwei Huang, Huarong Chen, Bin Wang, and Stephen Lin. Automatic thumbnail generation based on visual representativeness and foreground recognizability. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 253–261, 2015. 2

[11] For example, should one attempt to incorporate more 'SoTA' approaches such as using multi-modal text-image models trained on contrastive loss in lieu of *plain-vanilla* CNNs, we would just be opening newer avenues of brittleness such as typographic attacks [33].

[15] Laurent Itti. Visual salience. *Scholarpedia*, 2(9):3327, 2007. 2

[16] Paweł Karczmarek, Witold Pedrycz, Marek Reformat, and Elaheh Akhoundi. A study in facial regions saliency: a fuzzy measure approach. *Soft Computing*, 18(2):379–391, 2014. 10

[17] Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel. Contextual encoder-decoder network for visual saliency prediction. *Neural Networks*, 2020. 3, 4, 6, 12

[18] Amy Kuperinsky. Fly lingers on mike pence's head at vice presidential debate in #flygate - nj.com. https://www.nj.com/politics/2020/10/fly-lingers-on-mike-pences-head-at-vice-presidential-debate.html, October 2020. (Accessed on 06/08/2021). 10

[19] Guanbin Li and Yizhou Yu. Visual saliency detection based on multiscale deep cnn features. *IEEE transactions on image processing*, 25(11):5012–5024, 2016. 2

[20] Peng Lu, Hao Zhang, Xujun Peng, and Xiaofu Jin. An end-to-end neural network for image cropping by learning composition from aesthetic photos. *arXiv preprint arXiv:1907.01432*, 2019. 1

[21] Zehan Wang Lucas Theis. Speedy neural networks for smart auto-cropping of images. https://blog.twitter.com/engineering/en_us/topics/infrastructure/2018/Smart-Auto-Cropping-of-Images.html, January 2018. (Accessed on 10/19/2020). 1, 2, 4

[22] Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*, 47(4):1122–1135, 2015. 11

[23] Lucia Melloni, Sara van Leeuwen, Arjen Alink, and Notger G Müller. Interaction between bottom-up saliency and top-down control: how saliency maps are created in the human brain. *Cerebral cortex*, 22(12):2943–2952, 2012. 3

[24] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017. 3, 12

[25] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415. IEEE, 2012. 2

[26] Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020. 3

[27] Anelise Newman, Barry McNamara, Camilo Fosco, Yun Bin Zhang, Pat Sukhum, Matthew Tancik, Nam Wook Kim, and Zoya Bylinskii. Turkeyes: A web-based toolbox for crowdsourcing attention data. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020. 2

[28] Christian NL Olivers and Martijn Meeter. A boost and bounce theory of temporal attention. *Psychological review*, 115(4):836, 2008. 3, 10

[29] clarifications & comments-1 On the twitter cropping controversy: Critique. Packaging - google chrome. https://vinayprabhu.medium.com/on-the-twitter-cropping-controversy-critique-clarifications-and-comments-7ac66154f687, September 2020. (Accessed on 06/07/2014). 3

[30] Jaesik Park, Joon-Young Lee, Yu-Wing Tai, and In So Kweon. Modeling photo composition and its application to photo re-arrangement. In *2012 19th IEEE International Conference on Image Processing*, pages 2741–2744. IEEE, 2012. 2

[31] Michael I Posner and Yoav Cohen. Components of visual orienting. *Attention and performance X: Control of language processes*, 32:531–556, 1984. 3

[32] Vinay Prabhu and Matthew Mcateer. https://matthew-mcateer.github.io/oodles-of-oods/. https://matthew-mcateer.github.io/oodles-of-oods/, October 2020. (Accessed on 12/01/2020). 11

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 6

[34] Subramanian Ramanathan, Harish Katti, Nicu Sebe, Mohan Kankanhalli, and Tat-Seng Chua. An eye fixation database for saliency detection in images. In *European Conference on Computer Vision*, pages 30–43. Springer, 2010. 10

[35] João Rodrigues and JM Hans du Buf. Multi-scale keypoints in v1 and beyond: object segregation, scale selection, saliency maps and face detection. *BioSystems*, 86(1-3):75–90, 2006. 10

[36] Tomek Roszczynialski. Smart image cropping using saliency • filestack blog. https://blog.filestack.com/thoughts-and-knowledge/smart-image-cropping-using-saliency/, August 2020. (Accessed on 10/19/2020). 1, 2

[37] Lisa R Rubin and Molly Tanenbaum. "does that make me a woman?" breast cancer, mastectomy, and breast reconstruction decisions among sexual minority women. *Psychology of Women Quarterly*, 35(3):401–414, 2011. 4

[38] Ali Shafahi, W Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? *arXiv preprint arXiv:1809.02104*, 2018. 12, 13

[39] N. Shan, D. S. Tan, M. S. Denekew, Y. Y. Chen, W. H. Cheng, and K. L. Hua. Photobomb defusal expert: Automatically remove distracting people from photos. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 4(5):717–727, 2020. 2

[40] Vladimir N Sudakov and Boris S Tsirel'son. Extremal properties of half-spaces for spherically invariant measures. *Journal of Soviet Mathematics*, 9(1):9–18, 1978. 13

[41] Bongwon Suh, Haibin Ling, Ben Bederson, and David Jacobs. Automatic thumbnail cropping and its effectiveness. *UIST: Proceedings of the Annual ACM Symposium on User Interface Softaware and Technology*, 1, 05 2003. 1

[42] James Tanner and Laurent Itti. A top-down saliency model with goal relevance. *Journal of vision*, 19(1):11–11, 2019. 3

[43] Shashi Thakur. Google feed: feed your need to know. https://www.blog.google/products/search/feed-your-need-know/, July 2017. (Accessed on 05/24/2021). 9

[44] Lucas Theis, Iryna Korshunova, Alykhan Tejani, and Ferenc Huszár. Faster gaze prediction with dense networks and fisher pruning. *arXiv preprint arXiv:1801.05787*, 2018. 11

[45] Twitter. About advanced tweetdeck features. https://help.twitter.com/en/using-twitter/advanced-tweetdeck-features. (Accessed on 06/11/2021). 2

[46] Twitter. Debunking twitter myths. https://help.twitter.com/en/using-twitter/twitter-myths. (Accessed on 06/11/2021). 2

[47] Twitter. Twitter's non-consensual nudity policy | twitter help. https://help.twitter.com/en/rules-and-policies/intimate-media. (Accessed on 06/11/2021). 2

[48] Nachiappan Valliappan, Na Dai, Ethan Steinberg, Junfeng He, Kantwon Rogers, Venky Ramachandran, Pingmei Xu, Mina Shojaeizadeh, Li Guo, Kai Kohlhoff, et al. Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nature communications*, 11(1):1–12, 2020. 1

[49] Daniel Vaquero, Matthew Turk, Kari Pulli, Marius Tico, and Natasha Gelfand. A survey of image retargeting techniques. In *Applications of Digital Image Processing XXXIII*, volume 7798, page 779814. International Society for Optics and Photonics, 2010. 1, 2

[50] Kyle Wiggers. Microsoft researchers develop assistive eye-tracking ai that works on any device | venturebeat. https://venturebeat.com/2020/10/20/microsoft-researchers-design-software-based-eye-tracking-ai-that-works-on-any-device/, October 2020. (Accessed on 10/20/2020). 1

[51] Mai Xu, Yun Ren, and Zulin Wang. Learning to predict saliency on face images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3907–3915, 2015. 10

[52] Jianzhou Yan, Stephen Lin, Sing Bing Kang, and Xiaoou Tang. Learning the change for automatic image cropping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 971–978, 2013. 2

[53] Shoshana Zuboff. *The age of surveillance capitalism: The fight for a human future at the new frontier of power: Barack Obama's books of 2019*. Profile books, 2019. 5

Appendix A. Examples from other platforms such as Facebook and Google feed

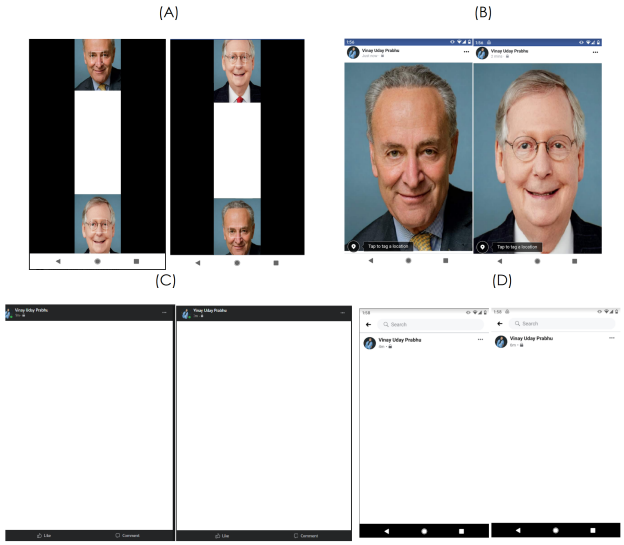


Figure 6: Gallery of images demonstrating the effect of SIC on Facebook across various modalities of access

As stated in the introduction section, SIC is not a platform-specific ill that besieges Twitter alone. The SIC-algorithmic pipeline today permeates other platforms that bank on voracity of consumption of news and social media content that might entail digital images. In this section of the appendix, we present specific examples from two other high-data-volume, high-data-velocity platforms, specifically Facebook and Google-Feed[12].

In Figure 6, we present the vagaries observed in the SIC cropping results on Facebook upon changing the modality of consumption from mobile to desktop and depending on whether we were accessing the platform via the official mobile app or a browser. In sub-figure(A), we see the screenshots of the two input longitudinal 3×1 grid images[13] with their face-detection bounding boxes produced by Facebook. In sub-figure(B), we present the post-cropped image screenshots of the input images when the mode-of-access was via a the Chrome web-browser(version 90.0.4430.210) on a Android 11; Pixel 2 device. As seen, the bottom-face in the input-image-grid seems to be cropped out on both the instances. Sub-figures 6 (C) and (D) present the post-cropped images that intriguingly center-cropped focusing on the middle-blank white space of the 3×1 input image grid when the mode of access was via the desktop-version Version (Chrome 90.0.4430.212 (Official Build) and 64-bit Windows 10.0.1.19041) and the Facebook mobile app (Android - version 319.0.0.39.120).

[12] https://www.blog.google/products/search/feed-your-need-know/

[13] The constituent faces belong to Charles (Chuck) Schumer, an American politician serving as Senate Majority Leader since January 20, 2021 (left in sub-figure 6(B)) and Mitch McConnell, his predecessor (right in sub-figure 6(B))

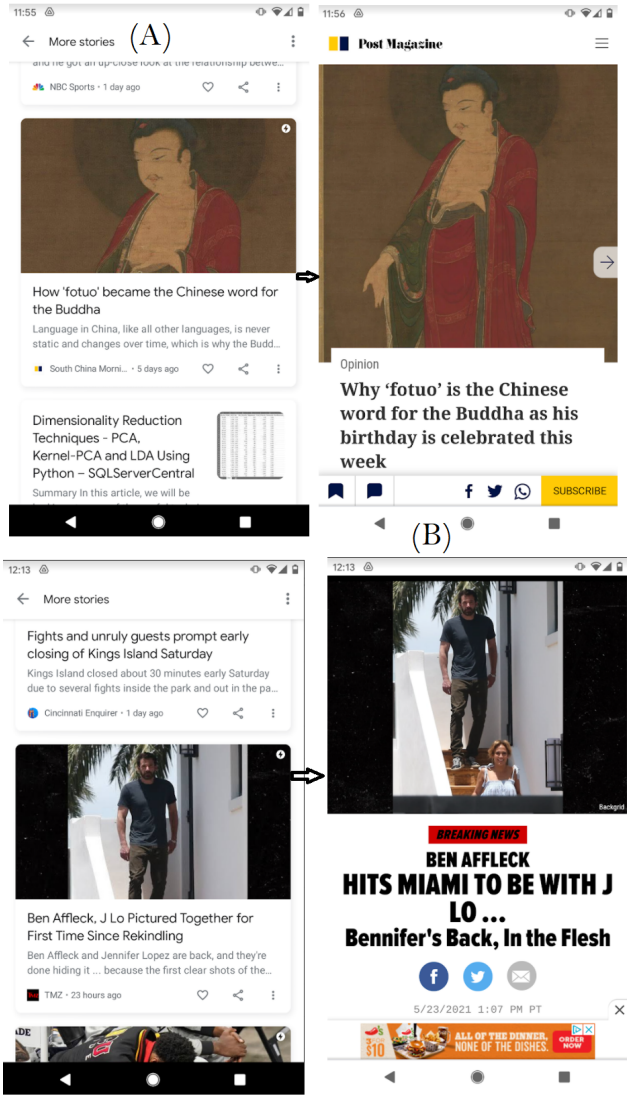


Figure 7: Examples of SIC via Google feed

In figure 7, we present screenshots of images procured while using *Google Feed* [43]. As seen in figure 7(B), we see the same issue of people getting cropped out in the case of *long rectangular-images* much akin to what unravelled on Twitter.

Appendix B. On Typicality, Abnormality and Temporality

Here, we investigate the innate friction between two important constructs that both vie as proxies to saliency: Typicality and Abnormality. In this regard, we present Figure 8

from the recent *flygate* incident[18] that entailed a fly lingering on a participant's head during the televised debate. The saliency model as shown in the image on the right emphasizes on the Eye-Nose-Mouth facial landmarks in tune with its 'expectations' (See [35, 51, 34, 16]), thus centering the *typicality* narrative. This coincides with little or no saliency associated with the anomalous fly on the head which was in fact the center-of-attention, hence quite literally the most saliency region of the image.

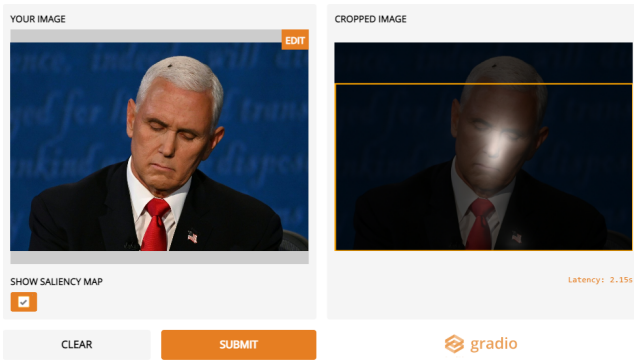


Figure 8: Anomaly or Typicality? An example image from the $\#flygate$ scandal

This observation brings us to the crucial issue of temporality. We observed that most real-world deployed SIC technologies subvert away the temporal aspect of attention variation, that pertains to the fact that even for a single volunteer, different regions of the image elicit different levels of attention based on the time of exposure of the image. In recent work on multi-duration saliency [10], the researchers introduced the `CodeCharts1K` dataset that consists of 1000 images from legacy datasets along with the associated viewing patterns at 0.5s, 3s, and 5 second duration. Fascinatingly, they observed what they termed as a dominant *"boomerang" pattern* in 33% of the images (with faces) where the volunteers fixated on faces at the 0.5 second mark only for their gaze to shift elsewhere at the 3 second mark, and making a comeback to the faces at the 5 second mark. What accentuates things further is that the datasets on which the models under scrutiny were trained also seem to not take into account phenomena related to the *boomerang effect* such as *attentional boost, attentional bounce* [28] and *attentional push*[12]. In [12], the authors introduce this *attentional push effect* as : *the power of the scene actors to direct and manipulate the attention allocation of the viewer*. Their working assumption, as stated is that: *"the attentional locus, i.e. the gaze location of each actor in a scene compels the viewers to direct their attention to that region, even if it has low salience by the scene actor's gaze*. This typically happens in natural imagery where there's an image component that is the center of attention of one or more onlookers contained in that very image. During saliency estimation, unless the model explicitly combines both the saliency map and the attentional push map, the onlooker-pixels receive high saliency scores thereby resulting in cropping out of the very object that all the onlooker(s) are fixating on in the image! An example of this is seen in the GSM saliency map displayed in Figure 16(B) in response to the famous historical image of the Apollo-11 take-off event from 1969 shown in Figure 16(A). The final cropping box (in dark orange in Figure 16(B)) is very similar to twitter's output crop seen in Figure 16(C). With both these models, one sees the attentional push effect vulnerability as the salient onlooker object in the image, that is the rocket being launched, is completely or partially cropped out.

Appendix C. Hand-crafted demonstration of brittleness and discussions

In figure 9, we present an exploration of the now-infamous *Obama-McConnell* image from September 2020 [13]. It is a 3×1 image grid consisting of the face image of the current (as of November 2020) United States senate majority leader Mitch McConnell, followed by blank white image and the face image of the 44^{th} president of the United States, Barack Obama. The reason why this image went viral was that when it was uploaded to *Twitter*, the SIC model entirely cropped out Obama's image only retaining Mitch McConnell's face (See sub-figure9(B)). When we downloaded this image and passed it as input into our GSM, we observed the replication of the Obama-cropped-out result as observed in sub-figures 9(C1) and 9(C2). While the saliency estimation part of the pipeline did a *decent* job of assigning saliency to both the faces (sub-figure9(C1)), marginal differences in the sum of saliency of the individual pixels in the cropping-policy-sliding-windows meant that Obama's image was cropped out, much akin to the actual blackbox algorithm implemented by Twitter (sub-figures 9(C2)). In order to demonstrate the brittleness of the technology and it's vulnerability to simple hand-crafted adversarial perturbations, we used the off-the-shelf *mild pixelation* edit-option provided by the TOAST-UI-Image editor (sub-figure 9(D)). The resultant image (see sub-figure(D1)), which looks nearly indistinguishable to the original image in sub-figure 9(A), when passed through the GSM, results in a saliency distribution as shown in sub-figure 9(E) and the eventually cropped image (now consisting of Barack Obama's face!) in sub-figure9(F). The same mildly pixelated image in sub-figure 9(D1) when now fed into twitter's cropping algorithm resulted in a slightly modified image (as seen in sub-figure 9(G)), albeit still retaining McConnell's face.

This is a fascinating example as it provides insights into many aspects of the brittleness of SIC that we now explore. **The inter-model variation:** Firstly, we note the stark difference between the response of the two models, that is

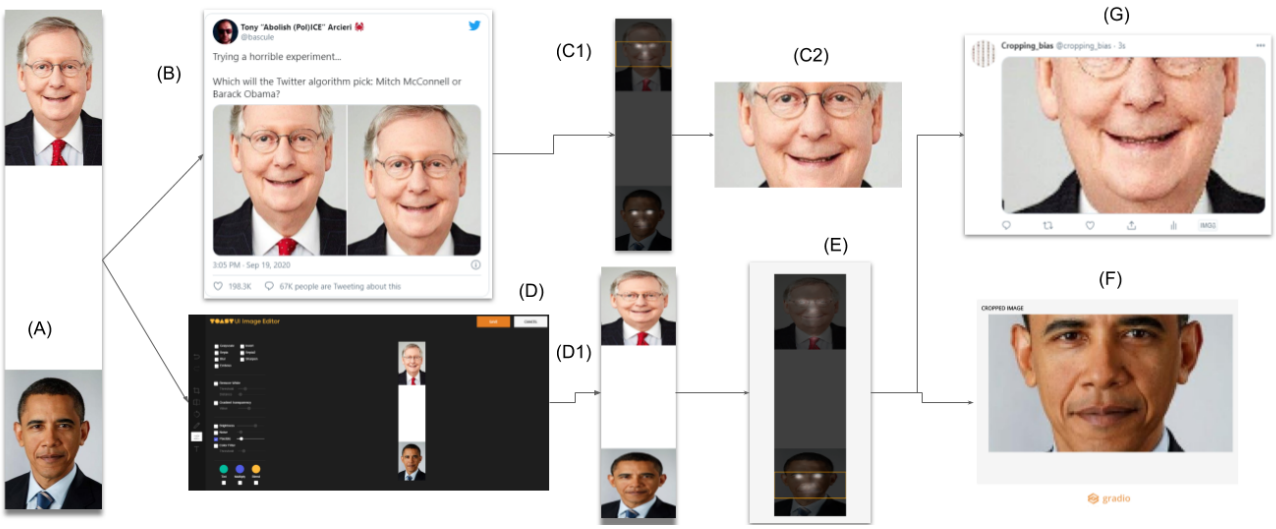


Figure 9: Figure demonstrating the brittleness of the saliency based image cropping technology via the hand-crafted perturbation approach

Twitter's and GSM's to the same image (in Sub-figure9D1) in spite of both having similarly *high* scores on the MIT's saliency benchmark and showing comparatively similar behavior on many of the previous images tried. This highlights the inter-model brittleness aspect. That is, two models trained with *similar* engineering objectives attaining *high* on a community accepted benchmark having drastically different results on the same image (or the same set of images). It is to be noted that the model deployed by Twitter [44] is a Fisher pruned model while ours is a non-pruned model. Model compression is a tricky endeavor and often, a compressed model that attains similar validation accuracy might come at the expense of enhanced vulnerability to OOD inputs [32]. Figure 10 further highlights this phenomenon.

The intra-model variation: Secondly, in Figure 9, we observed that slight perturbations to the input image dramatically changed the output for the GSM model. The clean input image in sub-figure 9(A) resulted in McConnell's face being chosen at the output (sub-figure 9(C2)) whereas it's adversarially perturbed version resulted in Obama's face as the cropped output sub-figure 9(D1)).

Appendix D. Adversarial vulnerability and the tribulations of an ethics oracle

With regards to this specific 3×1 image modality alone, should any platform stick to the current SIC modules that end up choosing only one constituent face image at the output, we argue that a *bias-free* SIC module is a pipe-dream, not just from some hand-wavy philosophical viewpoint, but also as mandated by the technical aspects as well. Hence, in order to explore the despondency of an *Ethics Oracle* who has been entrusted with the task of *fixing the bias*, we turn to Figure 11. Now, let us define the prototypical input image to be $I = [F_i, W, F_j]^T$. Here W represents the white blank image inserted in the middle, $F_i \in$A and $F_j \in B$ represent equi-sized images of faces of individuals belonging to two categories say A and B, controlled for all factors such as *saturation, size, resolution, lighting conditions, facial expressions, clothing and eye gaze* (Considered in [22]). Now, let $\xi(I)$ represent the **deterministic** transformation brought about the SIC module. Without loss of generality, let us now assume that based on some deontological (or consequentionist) framework, that $F_i \in A$ is the *ethically correct choice*[14] of face-selection for the input image $I = [F_i, W, F_j]^T$. Now, let $\zeta()$ represent the image editing process used by either a software agent (or a tinkering user) that results in small putatively imperceptible changes to the face images, with the goal that the *output* face image is now F_j instead of F_i. That is, the edited input image is $I^\dagger = [F_i^\dagger, W, F_j^\dagger]^T$, where the constituent face images are:

$$F_i^\dagger = \zeta(F_i), F_j^\dagger = \zeta(F_j) \tag{1}$$

such that the output cropped image is now $\xi(I^\dagger) = F_j$, while ensuring that $I^\dagger$ is still within the *semantic neighborhood*[15] of the original image.

[14] At this juncture, it is important to emphasize that we do not believe that there exists such a *correct* universal choice. Rather, our goal is to demonstrate that **irrespective** of what this correct choice might be based on a user's value-system, the user is **guaranteed** be let down and **will** experience bias

[15] For some small $\epsilon > 0$, we have $\delta(I, I^\dagger) \leqslant \varepsilon$ where $\delta()$ is the semantic distance metric

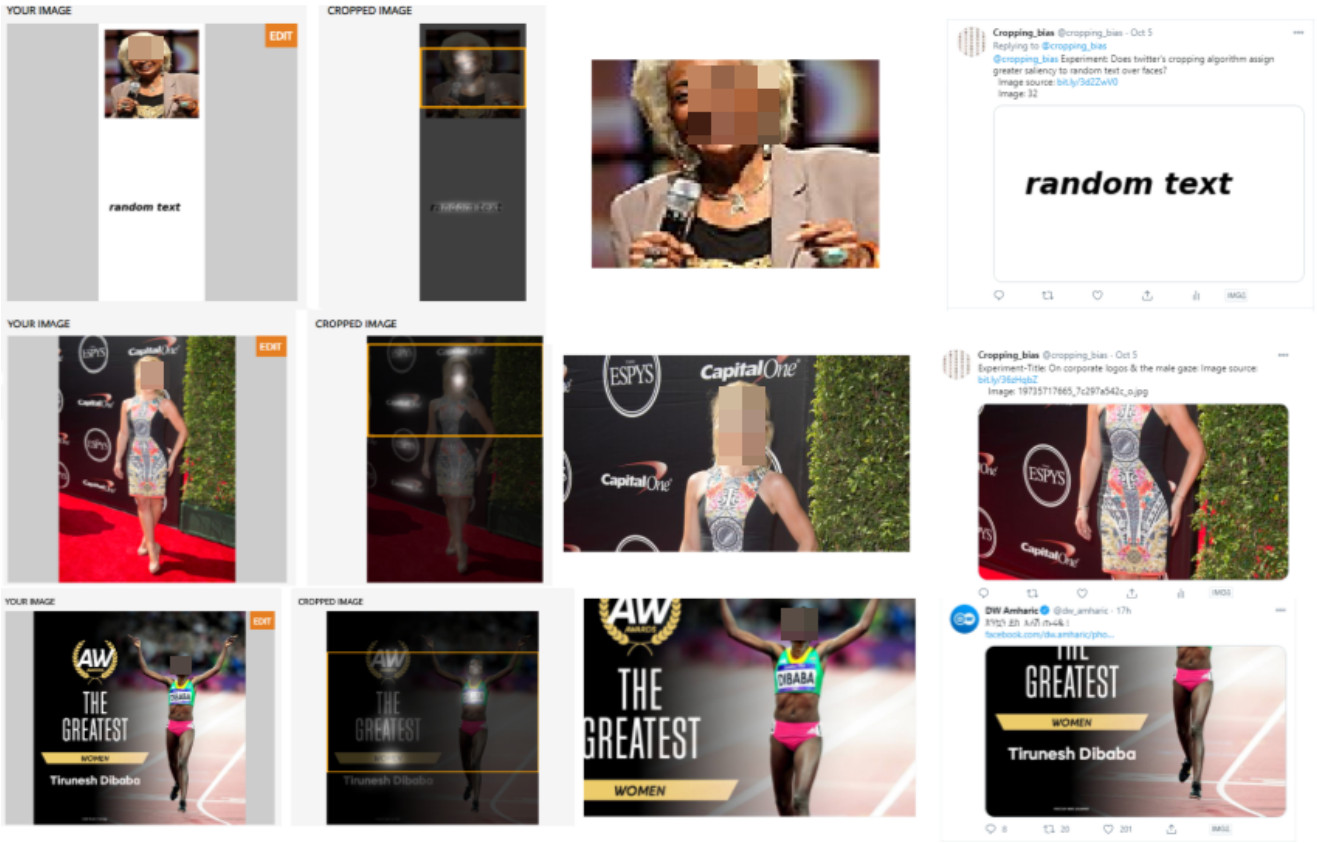


Figure 10: Example images where the GSM using [17] gave different results compared to twitter's SIC

Now, the arduous task in front of the *Ethics Oracle* is to ensure that for every conceivable editing function $\xi \in \Im$ and for every image $I^{\dagger} = [F_i^{\dagger}, W, F_j^{\dagger}]^T$ generated from a given input image I such that $\delta\left(I, I^{\dagger}\right) \leqslant \varepsilon$, the cropped output should remain $\xi\left(I^{\dagger}\right) = F_i$. Or,

$$\forall \xi \in \Im \& \forall \left(I, I^{\dagger}\right) : \delta\left(I, I^{\dagger}\right) \leqslant \varepsilon, \xi\left(I\right) = F_i, \xi\left(I^{\dagger}\right) = F_i \tag{2}$$

Defining $\rho\left(F_i\right)$ to be a resize-and-normalize-the-pixels function, with regards to Eq 1, the top-image ($I^{\dagger} = [F_i^{\dagger}, W, F_j^{\dagger}]^T$) was generated as $F_i^{\dagger} = \zeta\left(F_i\right) = \rho\left(F_i\right)$. The bottom image $F_j^{\dagger}$ was generated as a linear combination of a universal adversarial image and the normed-resized original image as per the following equation: $F_j^{\dagger} = \alpha_{vgg} v_{vgg} + (1 - \alpha_{vgg})\rho\left(F_j\right)$, where v_{vgg} is the *universal adversarial image* for the VGG-19 architecture[16] from [24]. In Figure 12(A), we see an example image generated according to eq.D, where the top and bottom face-images are `CFD-BM-002-013-N.jpg` and `CFD-WM-002-009-N.jpg` respectively (drawn from the CFD dataset) and $\alpha_{vgg} = 0.02496$. The result of feeding this image to twitter's SIC algorithm was that only `CFD-WM-002-009-N.jpg` *survived*. Now, when we increased α_{vgg} by a mere 0.00001, we were able to flip the image that appears n the post-cropping result to `CFD-BM-002-013-N.jpg`!

[16]Sourced from https://github.com/LTS4/universal/blob/master/precomputed/VGG-19.mat

D.1. Debiasing as tech-solutionist pipedream

Any engineer or a team of engineers entrusted with the arduous task of actually training and deploying this *ethics oracle* (even in this severely restricted toy-problem scenario), has to now grapple with recent results produced by the Machine Learning (ML) community that hint that adversarial vulnerability of ML models might well be inescapable and is a consequence of the high-dimensionality of the latent space of the input data (Ex: Real world images like the ones users upload on OSNs). To this end, we present two recent theorems from ML literature, from [8] and [38] respectively, that both derive their inspiration from isoperimteric inequality lemmas.

Theorem 1 (Existence of Adversarial Examples : [38]). *Consider a classification problem with m object classes, each distributed over the unit sphere $\mathbb{S}^{n-1} \subset \mathbb{R}^n$ with density functions $\{\rho_c\}_{c=1}^m$. Choose a classifier function*

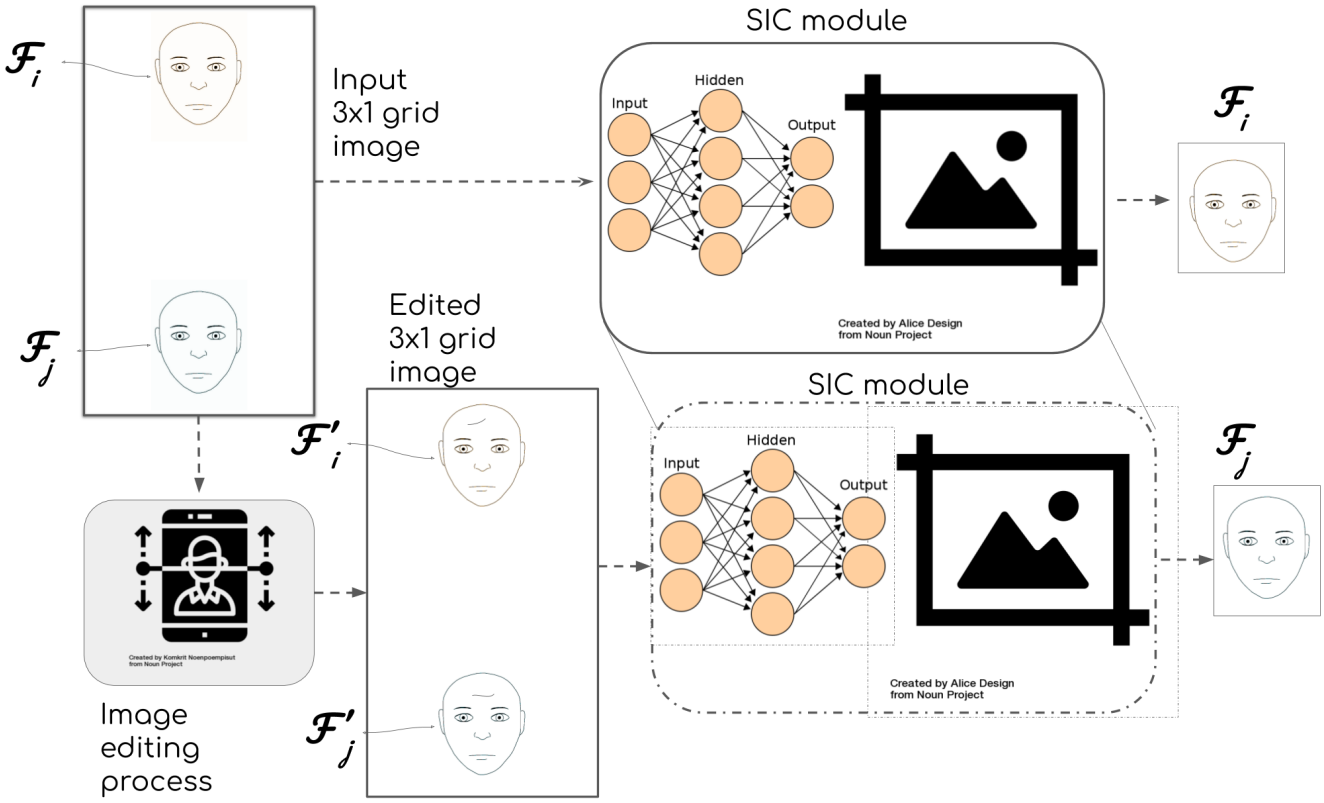


Figure 11: Motivating why fixing the 3×1 will remain a pipedream

$\mathcal{C} : \mathbb{S}^{n-1} \to \{1, 2, \ldots, m\}$ that partitions the sphere into disjoint measurable subsets. Define the following scalar constants:

- *Let V_c denote the magnitude of the supremum of ρ_c relative to the uniform density. This can be written $V_c := s_{n-1} \cdot \sup_x \rho_c(x)$.*
- *Let $f_c = \mu_1\{x|\mathcal{C}(x) = c\}$ be the fraction of the sphere labeled as c by classifier $\mathcal{C}$.*

Choose some class c with $f_c \leq \frac{1}{2}$. Sample a random data point x from ρ_c. Then with probability at least

$$1 - V_c \left(\frac{\pi}{8}\right)^{\frac{1}{2}} \exp\left(-\frac{n-1}{2}\epsilon^2\right) \tag{3}$$

one of the following conditions holds:

1. *x is misclassified by $\mathcal{C}$, or*
2. *x admits an ϵ-adversarial example in the geodesic distance.*

Theorem 2 (Existence of Adversarial Examples-2: [38])**.** *Let $f : \mathbb{R}^m \to \{1, \ldots, K\}$ be an arbitrary classification function defined on the image space. Then, the fraction of datapoints having robustness less than η satisfies:*

$$\mathbb{P}\left(r_{in}(x) \leq \eta\right) \geq \sum_{i=1}^{K} (\Phi(a_{\neq i} + \omega^{-1}(\eta)) - \Phi(a_{\neq i})) \,, \tag{4}$$

where Φ is the cdf of $\mathcal{N}(0, 1)$, and $a_{\neq i} = \Phi^{-1}\left(\mathbb{P}\left(\bigcup_{j \neq i} C_j\right)\right)$.

In particular, if for all i, $\mathbb{P}(C_i) \leq \frac{1}{2}$ (the classes are not too unbalanced), we have

$$\mathbb{P}\left(r_{in}(x) \leq \eta\right) \geq 1 - \sqrt{\frac{\pi}{2}} e^{-\omega^{-1}(\eta)^2/2} \,. \tag{5}$$

To see the dependence on the number of classes more explicitly, consider the setting where the classes are equiprobable, i.e., $\mathbb{P}(C_i) = \frac{1}{K}$ for all i, $K \geq 5$, then

$$\mathbb{P}\left(r_{in}(x) \leq \eta\right) \geq 1 - \sqrt{\frac{\pi}{2}} e^{-\omega^{-1}(\eta)^2/2} e^{-\eta\sqrt{\log\left(\frac{K^2}{4\pi \log(K)}\right)}} \,. \tag{6}$$

(This theorem is a consequence of the Gaussian isoperimetric inequality first proved in [5] and [40].) It is important to note that this *ethics oracle* has to be robust in the adversarial perturbation regime as well as the unrestricted adversarial-change regime as the tinkering users who seek to expose the fragility of the cropping system can be safely assumed to have all the time and image editing software at their disposal. Further, we posit that even if we were to restrict

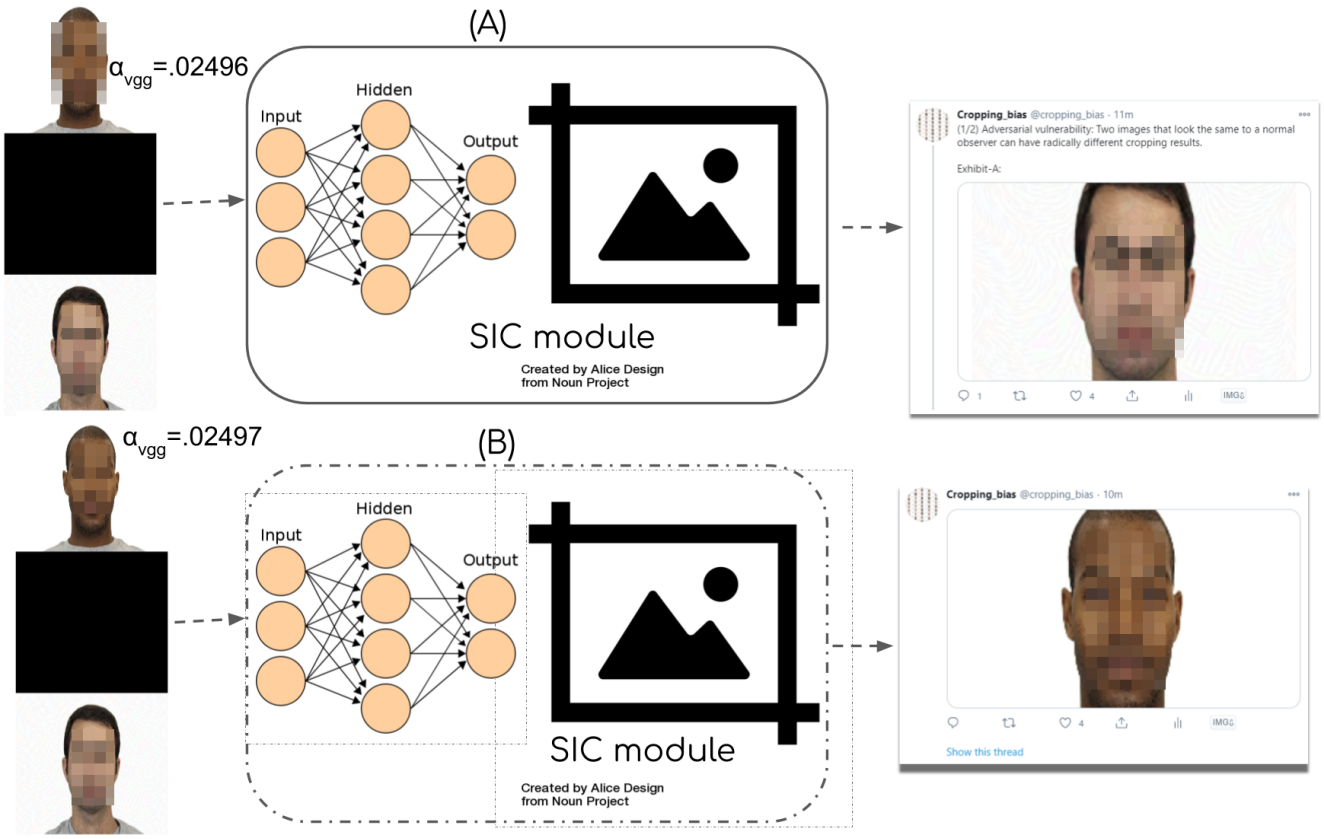


Figure 12: Demonstration of real-world adversarial vulnerability with 3×1 image grid framework

the possible set of editing transformations to be hand-crafted using only the `ToastUI` editor we have incorporated in our UI experiment (that is $\Im = TOAST - UI$), the probability of this adversarial susceptibility is close to 100%. We'd also like to insist that a tech-solutionist approach where cherry-picking an idiosyncratic academic dataset (such as CFD) and demonstrating images of face belonging to sets A and B are being cropped out at equal rates would only an escapism and will likely enrage the users of the platform further.

Besides these, we also catalogue SIC failures in images with only textual content (See Figure14), images with the *onlooker effect* or *attentional push* (Figure 16) and tweets with two co-uploaded images (Figure15).

Appendix E. Python implementation of the functions used

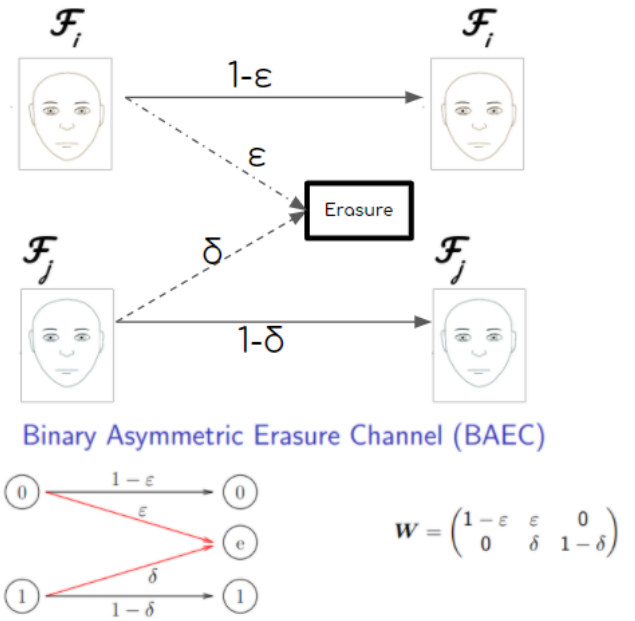
In this section, we share the details of the implementation of the image-generation processes used to test the fragility of SIC on twitter. In the following sub-sections, we share the Python implementations for these functions:

1. `send_tweet()` : Function to send out a tweet
2. `gen_3x1()` : Function to generate images for the 3x1 grid (with CFD images)
3. `gen_3x1_cfd_adv()`: Function to generate images for the 3x1 grid experiment using CFD images with VGG-19 universal adversarial image effect.
4. `gen_3x1_txt()`: Function to generate images for the 3x1 grid (Face + blank + text)

E.1. `send_tweet()`

```
##########################################
import tweepy
# 1: Function to send out a tweet using tweepy
def send_tweet(file_name_input,tweet_txt):
    """
  Function to send out a tweet
    """
    twitter_auth_keys = {
        "consumer_key"        : "xxxxxxxxxxxxx",
        "consumer_secret"     : "xxxxxxxxxxxxx",
        "access_token"        : "xxxxxxxxxxxxxxx
    ",
        "access_token_secret" : "xxxxxxxxxxxxx"
    } # Insert your credentials here

    auth = tweepy.OAuthHandler(
            twitter_auth_keys['consumer_key'],
```



The capacity-achieving distribution is determined by finding the solution x^* of

$$\varepsilon \log \varepsilon - \delta \log \delta = (1-\delta)\log(\delta + \varepsilon x) - (1-\varepsilon)\log(\varepsilon + \delta/x)$$

and setting

$$\frac{p_0^*}{p_1^*} = x^*, \quad p_0^* + p_1^* = 1.$$

Figure 13: SIC as an asymmetric erasure channel

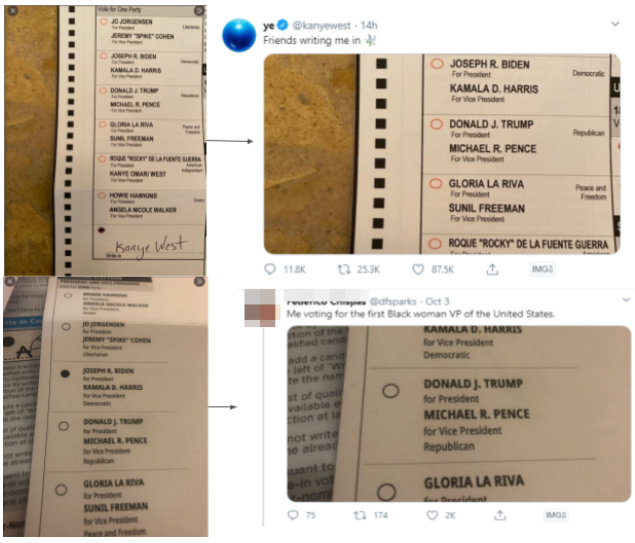


Figure 14: SIC failures in images with only textual content

```
            twitter_auth_keys['consumer_secret']
            )
    auth.set_access_token(
            twitter_auth_keys['access_token'],
```

Figure 15: SIC failures in tweets with 2 co-uploaded images

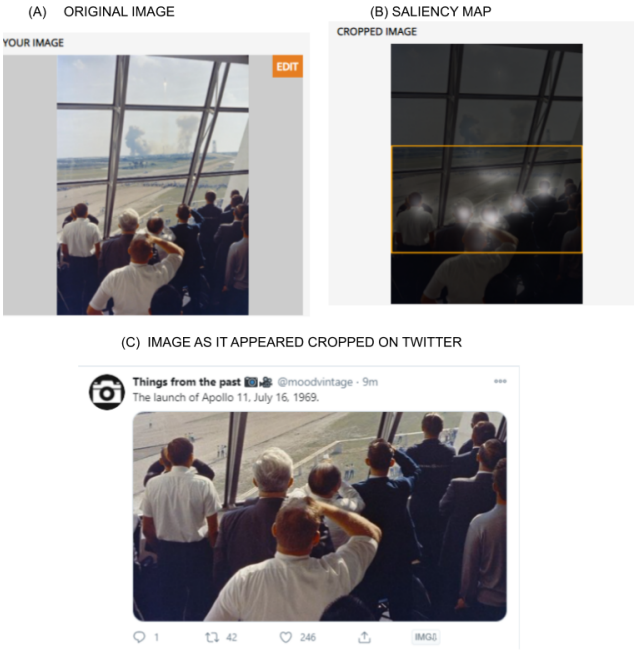


Figure 16: An example image demonstrating the *attentional push* vulnerability

```
            twitter_auth_keys['
    access_token_secret']
            )
    api = tweepy.API(auth, wait_on_rate_limit=
    True, wait_on_rate_limit_notify=True)


    # Upload image
    media = api.media_upload(file_name_input)

    # Post tweet with image
    tweet = tweet_txt
    post_result = api.update_status(status=tweet,
     media_ids=[media.media_id])
    return post_result
########################################
```

Listing 1: `send_tweet()` using *tweepy*

E.2. Function: `gen_3x1()`

```
import numpy as np
```

```
import cv2 as cv
import matplotlib.pyplot as plt
########################################
def gen_3x1(ind_test,
            path_top,
            path_bottom,
            dir_output,
            tweet_send=False):
    """
    ind_test: Index of the image in the test list
    path_top: List of top-image paths
    path_bottom:List of bottom-image paths
    dir_output: Directory for the ouput image
    tweet_send: Do you want to tweet out the
    image grid?
    """
    top_img=cv.imread(path_top)
    w_img=cv.imread(path_bottom)
    plt.figure(figsize=(2.6,2.24*3),dpi=100)
    ############################
    plt.subplot(3,1,1)
    plt.imshow(top_img.astype(int)[...,::-1],
    aspect='auto')
    plt.axis('Off')
    plt.subplot(3,1,2)
    plt.imshow(np.zeros((224,224,3),dtype='int'),
    aspect='auto')
    plt.axis('Off')
    plt.subplot(3,1,3)
    plt.imshow(w_img.astype(int)[...,::-1],aspect
    ='auto')
    plt.axis('Off')
    plt.tight_layout()
    plt.subplots_adjust(wspace=0, hspace=0)
    file_name_ind=f'./{dir_output}/bw_{ind_test}.
    png'
    plt.savefig(file_name_ind)
    plt.close()
    ############################
    if(tweet_send):
        tweet_txt_ind="Top: "+path_top.split('/')
    [-1] + "  Bottom:  " + path_bottom.split('/')
    [-1]
        _=send_tweet(file_name_ind,tweet_txt_ind)
    return None
```

Listing 2: Code for the `gen_3x1()` function

E.3. Function: `gen_3x1_cfd_adv()`

```
import numpy as np
import cv2 as cv
import matplotlib.pyplot as plt
#################
# Fetch the VGG-19 universal adv. image
!wget https://github.com/LTS4/universal/raw/
    master/precomputed/VGG-19.mat
import scipy.io as sio
vgg_uni=sio.loadmat('VGG-19.mat')
vgg_mat=vgg_uni['r']
plt.imshow(vgg_mat)
###############
def gen_3x1_cfd_adv(ind_test,
            path_top,
            path_bottom,
            dir_output,
      vgg_mat,
            tweet_send=False):
    """
    ind_test: Index of the image in the test list
    path_top: List of top-image paths
    path_bottom:List of bottom-image paths
    dir_output: Directory for the ouput image
    tweet_send: Do you want to tweet out the
    image grid? (True/False)
  vgg_mat: Universal adversarial image for VGG-19
     arch,

    """
    top_img=cv.imread(path_top)
    w_img=cv.imread(path_bottom)
    plt.figure(figsize=(2.6,2.24*3),dpi=100)
    ############################
    plt.subplot(3,1,1)
    plt.imshow(top_img.astype(int)[...,::-1],
    aspect='auto')
    plt.axis('Off')
    plt.subplot(3,1,2)
    plt.imshow(255*np.zeros((224,224,3),dtype='
    int'),aspect='auto')
    plt.axis('Off')
    plt.subplot(3,1,3)
    ################
  v_norm=(vgg_mat+10)/20
  img = cv2.imread(path_bottom, cv.
    IMREAD_UNCHANGED)
  x = cv.resize(img, (224,224), interpolation =
    cv.INTER_AREA).astype(int)[...,::-1]
  x_norm=x/x.max()
  x_out=(alpha)*v_norm+(1-alpha)*x_norm
  w_img_background=x_out/x_out.max()
  #################
  plt.imshow(w_img_background,aspect='auto')
    plt.axis('Off')
    plt.tight_layout()
    plt.subplots_adjust(wspace=0, hspace=0)
    file_name_ind=f'./{dir_output}/bw_{ind_test}.
    png'
    plt.savefig(file_name_ind)
    plt.close()
    ############################
    if(tweet_send):
        tweet_txt_ind="Top: "+path_top.split('/')
    [-1] + "  Bottom:  " + path_bottom.split('/')
    [-1]
        _=send_tweet(file_name_ind,tweet_txt_ind)
    return None
```

Listing 3: Code for the `gen_3x1_cfd_adv()` function

E.4. Function: `gen_3x1_txt()`

```
def gen_3x1_txt(ind_test,
            path_top,
            dir_output,
      s='random text',
            tweet_send):
    """
    ind_test: Index of the image in the test list
    path_top: List of top-image paths
    path_bottom:List of bottom-image paths
    dir_output: Directory for the ouput image
  s: String - Textual content into the bottom
    image
```

```
    tweet_send: Do you want to tweet out the
    image grid?

    """
    top_img=cv.imread(path_top)
    plt.figure(figsize=(2.6,2.24*3),dpi=100)
    ############################
    plt.subplot(3,1,1)
    plt.imshow(top_img.astype(int)[...,::-1],
    aspect='auto')
    plt.axis('Off')
    plt.subplot(3,1,2)
    plt.imshow(255*np.zeros((224,224,3),dtype='
    int'),aspect='auto')
    plt.axis('Off')
    plt.subplot(3,1,3)
    ################
    plt.imshow(np.ones((224,224,3), dtype=np.
    float),aspect='auto',cmap='gray',vmin=0,vmax
    =1)
#     fig = plt.gcf()
#     size = fig.get_size_inches()*fig.dpi # size
     in pixels
    plt.text(10,10,s, fontsize=18)
    plt.axis('Off')
    plt.subplots_adjust(wspace=0, hspace=0)
    file_name_ind=f'./{dir_output}/bw_{ind_test}.
    png'
    plt.savefig(file_name_ind)
    plt.close()
    ############################
    if(tweet_send):
        tweet_txt_ind="Top: "+path_top.split('/')
    [-1] + "  Bottom:  " + path_bottom.split('/')
    [-1]
        _=send_tweet(file_name_ind,tweet_txt_ind)
    return None
```

Listing 4: Code for the `gen_3x1_txt()` function