# Saliency-Aware Privacy Protection in Augmented Reality Systems

Gautham Ramajayam
Stony Brook University
gramajayam@cs.stonybrook.edu

Tao Sun
Stony Brook University
tao@cs.stonybrook.edu

Chiu C. Tan
Temple University
cctan@temple.edu

Lannan Luo
George Mason University
lluo4@gmu.edu

Haibin Ling
Stony Brook University
hling@cs.stonybrook.edu

## ABSTRACT

The augmented reality (AR) Metaverse environment combines the physical and virtual world together. Privacy is a major concern in AR since the cameras use to capture the physical world can also capture other images that may potentially violate user or by-stander privacy. Advances in deep learning to process images and videos have exacerbated such privacy risks. This paper presents a new technique to protect privacy in AR systems by combining the idea of visual saliency together with privacy-sensitive object detection. We show that our technique is able to provide additional context to a given image to better balance between privacy and overall usability of the system.

## CCS CONCEPTS

• **Security and privacy → Human and societal aspects of security and privacy**; • **Computing methodologies → Artificial intelligence**.

## KEYWORDS

privacy protection, deep learning, visual saliency

**ACM Reference Format:**
Gautham Ramajayam, Tao Sun, Chiu C. Tan, Lannan Luo, and Haibin Ling. 2023. Saliency-Aware Privacy Protection in Augmented Reality Systems . In *First Workshop on Metaverse Systems and Applications (MetaSys '23), June 18–22, 2023, Helsinki, Finland.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3597063.3597358


*MetaSys '23, June 18–22, 2023, Helsinki, Finland*

ACM ISBN 979-8-4007-0213-6/23/06…$15.00
https://doi.org/10.1145/3597063.3597358

## 1 INTRODUCTION

Augmented Reality (AR) is a unique component of the metaverse that combines the physical world together with the virtual world [17, 21]. The commonly available AR systems are *mobile AR* (MAR) [5] in the form of a head-mounted device like the HoloLens, or a portable device in the form of a tablet or smartphone. This form of AR is expected to continue to grow in the coming years [11, 27].

A typical AR system consists of a camera that captures the physical world and a display that allows the projection of virtual objects overlaid on top of the physical objects. Algorithms are deployed for 3D scene geometry estimation, scene semantics understanding, and virtual scene rendering. While the data pipeling can be implemented entirely on the AR device, most AR systems rely on the computational resources of a backend cloud service to perform many of the operations.

The privacy implications of AR systems is a major concern [1, 4, 8, 9]. In particular, users are concerned that the camera (which is an integral part of an AR system) may capture information about the user or bystanders, which may reveal private information [6, 12, 29].

### 1.1 Current privacy protection and its limitations

A key privacy technique used to protect privacy in AR systems [14, 15, 34], and earlier smartphone camera apps [2, 22, 30] is *obfuscation*. This is where specific objects (e.g. faces, license plates, computer screens, etc.), or regions (e.g. entire background) are first identified, and then distorted so as to make them unintelligible (e.g. blurring the license plate) to the viewer of the video recordings [25]. In this paper, we will use the general term "blurring" to refer to this distortion process, though in practice, other methods such as blacked outs can also be used. The typical process of obfuscation is as following. The system developer will first identify a list of objects/regions that are privacy sensitive and then design computer vision algorithms to automatically identify these

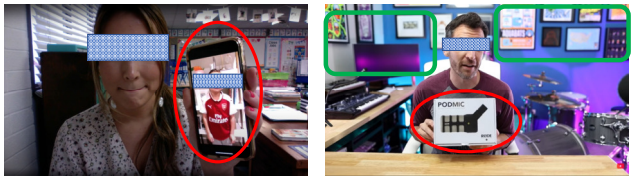

**Figure 1: Example of limitations of existing privacy scheme. Green box indicates correct application of privacy rules, and red box indicates incorrect application. Faces are intentionally masked for privacy reasons.**

objects in the video and then blur then out. End users may be able to configure the system to selectively blur certain objects (e.g. blur logos but not license plates), depending on the application requirements.

A limitation with the current approach of obfuscation is that the concept of privacy cannot be easily reduced to a set of objects or simple rules. We illustrate this using two images extracted from Youtube to represent what an AR system might capture in Figure 1. The example on the left shows the result of a common privacy rule to blur third-party faces that is supposed to protect the privacy of bystanders. It is clear from the context that the third-party face shown on the smartphone (red bounding box) should not be blurred, since it is apparent the user wants to show this image to the camera. The image on the right illustrates the result of another common privacy rule to blur all text information. This type of rule is used to prevent the camera from potentially capturing documents on the background, diplomas hanging on the wall, emails on computer screens, and so on (green bounding box). However, from the context of the image, the area bounded by the red box should **not** have been blurred, since the user is intentionally showing it to the camera.

### 1.2 Our contributions

Our approach to this problem combines modern deep learning-based saliency prediction and object detection algorithms. Different from previous methods that detect a list of predefined privacy-sensitive objects, our method only detect a single class of privacy regions. We utilize a key concept called *Visual Saliency* [38], meaning visually distinctive objects or regions in an image, with a prior knowledge that privacy regions are less likely to be salient and vice versa. This mutually exclusive relationship makes it possible to detect privacy-sensitive objects while considering their specific context. We first formulate a two-stage method that refines results from a privacy object detector using saliency scores. Then we propose a hybrid model that combines object detector and saliency detector together and can be trained in an end-to-end way.

To validate the proposed methods, we collected a dataset of video call scenario and annotated the privacy sensitive objects in the images of the dataset. Then, the proposed methods are tested on the dataset along with standard object detection algorithms. The effectiveness of the proposed approach is clearly validated by the experimental results.

## 2 RELATED WORK

Security and privacy are important concerns of metaverse [10, 39]. The privacy of AR systems are of particular concern because of the camera used to capture the physical surroundings can be combined with machine learning to extract privacy sensitive data [23, 35, 37, 40]. Work by [18] was among the first to explore the potential for an adversary to hide malicious code in an AR app to extract privacy sensitive information from the camera feed.

Closely related to the AR privacy are research on *visual privacy* camera systems such as those found on smart glasses, smartphones, and body cameras. Research on visual privacy protection techniques‘to detect and blur specific objects [3, 24, 25], as well as gestures and other mechanisms to express privacy preferences [16, 36] can be applied to AR systems as well. This has led to research in OS support for AR apps [7, 14, 19, 26] to support these techniques within the requirements of AR systems.

Object detection is one of the fundamental tasks in computer vision. The goal is to recognize and locate predefined objects in an image. There are two major categories of deep learning based object detection methods, one-stage methods and two-stage methods. YOLO [31] is one of the most representative one-stage methods. It frames object detection as a regression problem, and predicts the bounding boxes and class probabilities directly from full images without post-processing. YOLO is extremely fast due to its unified architecture. YOLOv3 [32] improves over YOLO by predicting across 3 different scales. This makes it more accurate to detect objects of different sizes. Faster R-CNN [33], different from YOLO, needs to generate potential bounding boxes first and then classify these proposed boxes. The results are refined with post-processing. Faster R-CNN represents two-stage methods that are accurate but slow and hard to optimize.

Saliency detection has different motivation compared to generic object detection. Salient object detection aims to find most visually distinctive objects or regions in an image. Whether one object is salient depends on its context. For example, the face of a person could be a salient one when he/she is talking to another one in a video conference, but not when the video is in presentation mode. Saliency detection can be implemented using deep neural networks with pixel-level classification losses. A recent survey can be found in [38]. Among many salient object or saliency detection algorithms, two of them are mostly related to our study: Hou et

al. [13] builds upon VGGNet, and fuses classification losses at 6 different scales. Qin et al. [28] proposes a new architecture named U$^2$-Net that does not rely on pretrained backbones. It adopts a two-level nested U-structure and novel Residual U-blocks. Saliency detection has an inborn connection with privacy detection. It is less likely that a salient region contains privacy-sensitive objects.

## 3 SALIENCY-AWARE PRIVACY DETECTION AND PROTECTION

One key step in protecting privacy is to detect privacy regions in a camera-captured image. This privacy detection, or more generally image detection, is one of the fundamental tasks in computer vision. Deep learning based methods have greatly improved the detection accuracy on benchmarks like MS-COCO [20]. Besides their good performance, a significant advantage over conventional computer vision methods is that they can be easily transferred to new datasets. For example, we can take a YOLOv3 model for generic objection detection pretrained on MS-COCO and fine-tuned it on the privacy protection images for privacy detection. The powerful representation ability of deep learning enables context-aware privacy detection where the the spatial information of surroundings to the persons and temporal information of current activities are considered.

Usually, saliency regions in an image are less likely to be privacy sensitive. With this prior knowledge, we can use a saliency detection model to refine the results from privacy detector. Depending on the way to use this information, we investigate two types of privacy detection approaches: a two-stage method and an end-to-end one.

### 3.1 Two-stage method

In a two-stage method, the privacy region detector and saliency detector are trained independently. To make a prediction on an image, the privacy detector first detects candidate privacy regions. These candidates are then filtered based on the saliency detection results. For the first stage, we experimented with two representative image detectors, YOLOv3 [32] and Faster-RCNN [33], as the privacy region detector, and selected the *Deeply Supervised Salient Object Detection with Short Connections* method [28] described in the related work section as the saliency detector. The detailed procedure of saliency filtering is outlined in the evaluation section.

### 3.2 End-to-end method

A potential drawback of the previous two-stage method is that the generation of privacy regions does not consider the saliency information, which can be sub-optimal. Rather than manually thresholding with saliency map, an alternative way is to take the saliency map as additional features to the privacy region detector and let model automatically learn the relationship between privacy region and saliency region. Figure 2 shows the architecture of the end-to-end method using YOLOv3 as privacy region detector and $U^2$-Net as saliency detector. YOLOv3 and U$^2$-Net takes the same original image as the network input. Then the output saliency map from U$^2$-Net are viewed as additional feature presentations and added to the latent feature of YOLOv3 before generating privacy regions. Since YOLOv3 works at three different resolutions, the saliency map is rescaled to fit the corresponding ones. In this hybrid model, the generation of privacy regions not only relies on object semantics but also on contextual saliency.

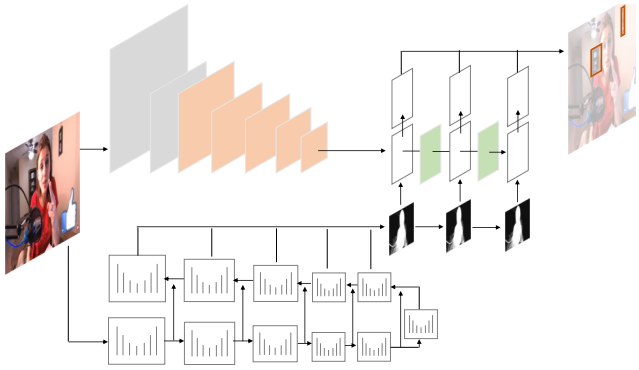

**Figure 2: Architecture of hybrid YOLOv3+$U^2$-Net. The lower subnet is the $U^2$-Net [28] for saliency detection and the upper subnet is the YOLOv3 [32] for privacy object detection. Residual connections and module meanings are omitted for clarity. Please refer to original paper for details.**

## 4 EVALUATION

### 4.1 Data collection

We created our own dataset for evaluation since there are no AR datasets on privacy. Since there are many potential AR applications, no single dataset is likely to capture all scenarios. Instead, we focused on a common scenario where the AR device is used in a conversation type environment, i.e. a user wearing a head-mounted device talking to someone else, or a user having a teleconference using an AR enabled device. Such a scenario will capture objects of interest that are part of the conversation, as well as objects that are may potentially be privacy sensitive.

To create our dataset, we collect YouTube videos that are about video call, since this is similar to what an AR device will capture, and extract a total of 1,000 images from them. In each image, the privacy regions, e.g., containing faces and text depending on the context, are manually labeled.

We name this dataset as *YouTube Dataset*. Figure 3 shows exemplar images and their ground-truth notations of privacy regions.

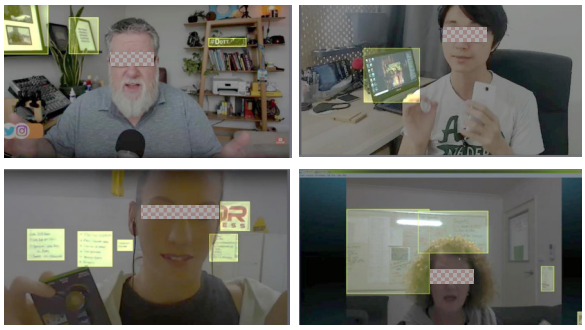

**Figure 3: Exemplar images with notations from the *YouTube Dataset*. Eye regions are blocked on purpose for visualization (same for Figure 3 as well).**

## 4.2 Experiment details

We follow the standard evaluation procedure and split the *YouTube Dataset* by 80% for training and 20% for testing. Each model is trained on the training set, and then evaluated on the test set using mean average precision (mAP). This metric measures how well the predicted privacy regions are consistent with the ground-truth privacy regions.

*4.2.1 YOLOv3 with Manual Thresholding (MT) using saliency.* A YOLOv3 model pretrained on MS-COCO [20] is fine-tuned on the training images of the collected dataset to detect privacy regions. Meanwhile, a pretrained saliency detector is fine-tuned on the training images to detect salient regions. In the Manual Thresholding (MT) step, the predicted privacy regions from the YOLOv3 model are refined with the saliency map. If the average saliency score of a detected privacy region is above 0.5, it will be rejected and not considered as a privacy region.

*4.2.2 Faster R-CNN with Manual Thresholding using Saliency.* This model replaces the YOLOv3 in previous method with the pretrained Faster R-CNN model. It adopts the same procedure to refine detected privacy regions with saliency score thresholding.

*4.2.3 Hybrid YOLOv3+$U^2$-Net.* The hybrid model combines YOLOv3 and $U^2$-Net into one integrated model and train it in an end-to-end way. The saliency map from $U^2$-Net are input as additional feature to the YOLOv3 model. The whole model is trained on the training images for 100 epochs. To further analyze the effect of saliency map, we also add another step of manual threshold as we do in previous methods.

## 4.3 Quantitative results

The evaluation results are listed in Table 1. As can be seen from the table, using saliency in visual privacy detection clearly improves the mAP compared to not using saliency. YOLOv3 performs comparable to Faster R-CNN, though its mAP is slightly higher. Hybrid YOLOv3+$U^2$-Net improves over YOLOv3 only, showing the effect of combining saliency into privacy detection procedure. An interesting point is that the post-processing of manually thresholding of saliency works better than end-to-end training. The reasons may be that our current dataset is not large enough to implicitly learn the mutual exclusive relationship between privacy region and saliency region. This can be improved by shifting the region classification probability of the privacy detector, rather than using saliency map as latent feature only. We leave this for future work.

**Table 1: Comparison of different privacy detection methods on the test set. Note: MT is short for "manual threshold", E2D for "end to end".**

| Object detector | Saliency detector | Combination strategy | mAP (%) |
|---|---|---|---|
| Faster R-CNN | None | N/A | 25.5 |
| Faster R-CNN | $U^2$-Net | MT | 33.6 |
| YOLOv3 | None | N/A | 27.4 |
| YOLOv3 | $U^2$-Net | MT | 35.1 |
| YOLOv3+$U^2$-Net | $U^2$-Net | E2E | 29.8 |
| YOLOv3+$U^2$-Net | $U^2$-Net | E2E + MT | 34.7 |

## 4.4 Qualitative results

In this section, we provide some qualitative results. Figure 4 shows two test images and their privacy region detection results. As can been seen, the saliency map indicates the most prominent objects in the images, i.e., persons. The hybrid model detects privacy regions in the images, but there are false positive results. After manually thresholding with saliency map, these false positive results are removed.

# 5 CONCLUSION AND FUTURE WORK

In this paper, we present the deep learning approach to protect privacy in camera-based critical applications. We point out the limitations of current approaches in improper predetermined privacy-sensitive objects and under-exploitation of context of privacy situation. We utilize the mutual exclusive relationship beween privacy regions and saliency regions, and propose a deep learning-based privacy-sensitive object detection approach. Our experiments on the collected video call dataset demonstrate its effectiveness.

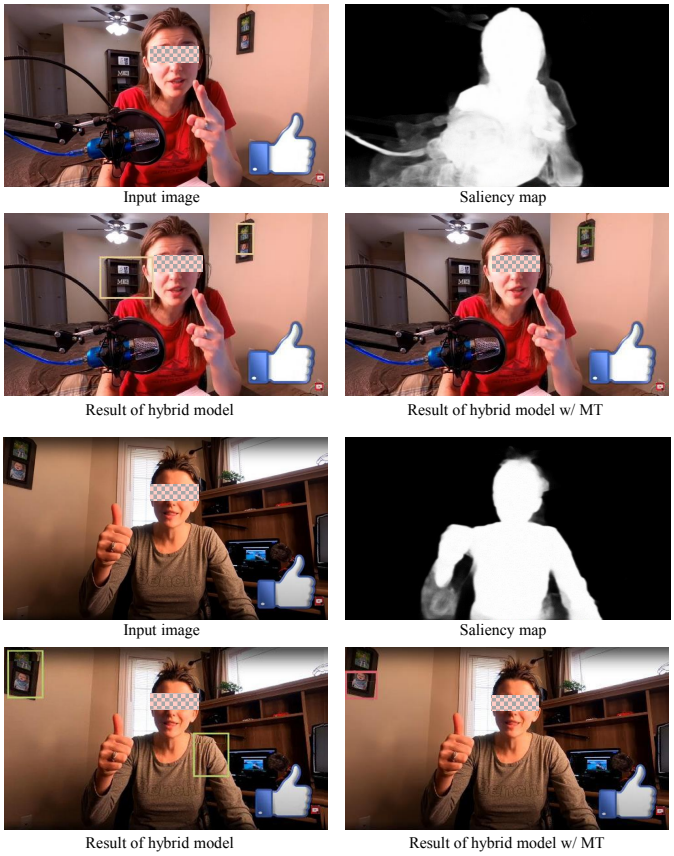


**Figure 4: Privacy region detection results with hybrid model and manual thresholding.**

Our results are very preliminary and there are several future works to explore. Currently, end-to-end training is not as effective as manually thresholding. It is worthwhile to study how to bring saliency map into the decision procedure of the privacy object detector. For example, the classification score as privacy object may be shifted based on the saliency score. In video data, the temporal information from adjacent frames could be used to better infer current activates. This is critical as whether an object is privacy sensitivity is highly related to the person's intention. Moreover, more effective ways, such as earlier feature fusion, for integrating the saliency information into privacy sensitivity detection can potentially improve the detection accuracy. Finally, another direction is to inpaint privacy-sensitive regions to make it natural rather than simply obfuscating the regions.

## REFERENCES

[1] Alessandro Acquisti, Ralph Gross, and Frederic D Stutzman. Face recognition and privacy in the age of augmented reality. *Journal of Privacy and Confidentiality*, 6(2):1, 2014.

[2] Paarijaat Aditya, Rijurekha Sen, Peter Druschel, Seong Joon Oh, Rodrigo Benenson, Mario Fritz, Bernt Schiele, Bobby Bhattacharjee, and Tong Tong Wu. I-pic: A platform for privacy-compliant image capture. In *Proceedings of the 14th annual international conference on mobile systems, applications, and services*, pages 235–248, 2016.

[3] Mohammed Eunus Ali, Anika Anwar, Ishrat Ahmed, Tanzima Hashem, Lars Kulik, and Egemen Tanin. Protecting mobile users from visual privacy attacks. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 1–4, 2014.

[4] Kent Bye, Diane Hosfelt, Sam Chase, Matt Miesnieks, and Taylor Beck. The ethical and privacy implications of mixed reality. In *ACM SIGGRAPH 2019 Panels*, pages 1–2. 2019.

[5] Dimitris Chatzopoulos, Carlos Bermejo, Zhanpeng Huang, and Pan Hui. Mobile augmented reality survey: From where we are to where we go. *Ieee Access*, 5:6917–6950, 2017.

[6] Scott G Dacko. Enabling smart retail settings via mobile augmented reality shopping apps. *Technological forecasting and social change*, 124:243–256, 2017.

[7] Loris D'Antoni, Alan Dunn, Suman Jana, Tadayoshi Kohno, Benjamin Livshits, David Molnar, Alexander Moshchuk, Eyal Ofek, Franziska Roesner, Scott Saponas, et al. Operating system support for augmented reality applications. In *Presented as part of the 14th Workshop on Hot Topics in Operating Systems*, 2013.

[8] Jaybie A De Guzman, Kanchana Thilakarathna, and Aruna Seneviratne. Security and privacy approaches in mixed reality: A literature survey. *ACM Computing Surveys (CSUR)*, 52(6):1–37, 2019.

[9] Tamara Denning, Zakariya Dehlawi, and Tadayoshi Kohno. In situ with bystanders of augmented reality glasses: Perspectives on recording and privacy-mediating technologies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2377–2386, 2014.

[10] Roberto Di Pietro and Stefano Cresci. Metaverse: security and privacy issues. In *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, pages 281–288. IEEE, 2021.

[11] Tegegne Tesfaye Haile and Mincheol Kang. Mobile augmented reality in electronic commerce: investigating user perception and purchase intent amongst educated young adults. *Sustainability*, 12(21):9185, 2020.

[12] David Harborth. Augmented reality in information systems research: a systematic literature review. 2017.

[13] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3203–3212, 2017.

[14] Jinhan Hu, Andrei Iosifescu, and Robert LiKamWa. Lenscap: split-process framework for fine-grained visual privacy control for augmented reality apps. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, pages 14–27, 2021.

[15] Yoonsang Kim, Saeed Boorboor, Amir Rahmati, and A Kaufman. Design of privacy preservation system in augmented reality. In *IEEE Symposium on Visualization for Cyber Security VIZSEC*, 2021.

[16] Marion Koelle, Swamy Ananthanarayan, Simon Czupalla, Wilko Heuten, and Susanne Boll. Your smart glasses' camera bothers me! exploring opt-in and opt-out gestures for privacy mediation. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction*, pages 473–481, 2018.

[17] Lik-Hang Lee, Tristan Braud, Pengyuan Zhou, Lin Wang, Dianlei Xu, Zijun Lin, Abhishek Kumar, Carlos Bermejo, and Pan Hui. All one needs to know about metaverse: A complete survey on technological

singularity, virtual ecosystem, and research agenda. *arXiv preprint arXiv:2110.05352*, 2021.

[18] Sarah M Lehman, Abrar S Alrumayh, Kunal Kolhe, Haibin Ling, and Chiu C Tan. Hidden in plain sight: Exploring privacy risks of mobile augmented reality applications. *ACM Transactions on Privacy and Security*, 25(4):1–35, 2022.

[19] Sarah M Lehman and Chiu C Tan. Privacymanager: An access control framework for mobile augmented reality applications. In *2017 IEEE Conference on Communications and Network Security (CNS)*, pages 1–9. IEEE, 2017.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[21] Huansheng Ning, Hang Wang, Yujia Lin, Wenxi Wang, Sahraoui Dhelim, Fadi Farha, Jianguo Ding, and Mahmoud Daneshmand. A survey on metaverse: the state-of-the-art, technologies, applications, and challenges. *arXiv preprint arXiv:2111.09673*, 2021.

[22] Katarzyna Olejnik, Italo Dacosta, Joana Soares Machado, Kévin Huguenin, Mohammad Emtiyaz Khan, and Jean-Pierre Hubaux. Smarper: Context-aware and automatic runtime-permissions for mobile devices. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 1058–1076. IEEE, 2017.

[23] Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8466–8475, 2018.

[24] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *Proceedings of the IEEE international conference on computer vision*, pages 3686–3695, 2017.

[25] José Ramón Padilla-López, Alexandros Andre Chaaraoui, and Francisco Flórez-Revuelta. Visual privacy protection methods: A survey. *Expert Systems with Applications*, 42(9):4177–4195, 2015.

[26] Nuno Pereira, Anthony Rowe, Michael W Farb, Ivan Liang, Edward Lu, and Eric Riebling. Arena: The augmented reality edge networking architecture. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 479–488. IEEE, 2021.

[27] Hong Qin, Babajide Osatuyi, and Lu Xu. How mobile augmented reality applications affect continuous use and purchase intentions: A cognition-affect-conation perspective. *Journal of Retailing and Consumer Services*, 63:102680, 2021.

[28] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 106:107404, 2020.

[29] Philipp A Rauschnabel, Jun He, and Young K Ro. Antecedents to the adoption of augmented reality smart glasses: A closer look at privacy risks. *Journal of Business Research*, 92:374–384, 2018.

[30] Nisarg Raval, Animesh Srivastava, Ali Razeen, Kiron Lebeck, Ashwin Machanavajjhala, and Lanodn P Cox. What you mark is what apps see. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, pages 249–261, 2016.

[31] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[32] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[34] Franziska Roesner, David Molnar, Alexander Moshchuk, Tadayoshi Kohno, and Helen J Wang. World-driven access control for continuous sensing. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1169–1181, 2014.

[35] Zhiqi Shen, Shaojing Fan, Yongkang Wong, Tian-Tsong Ng, and Mohan Kankanhalli. Human-imperceptible privacy protection against machines. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1119–1128, 2019.

[36] Jiayu Shu, Rui Zheng, and Pan Hui. Cardea: Context-aware visual privacy protection from pervasive cameras. *arXiv preprint arXiv:1610.00889*, 2016.

[37] Ashwini Tonge and Cornelia Caragea. Image privacy prediction using deep neural networks. *ACM Transactions on the Web (TWEB)*, 14(2):1–32, 2020.

[38] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, Haibin Ling, and Ruigang Yang. Salient object detection in the deep learning era: An in-depth survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, in press.

[39] Yuntao Wang, Zhou Su, Ning Zhang, Rui Xing, Dongxiao Liu, Tom H Luan, and Xuemin Shen. A survey on metaverse: Fundamentals, security, and privacy. *IEEE Communications Surveys & Tutorials*, 2022.

[40] Jun Yu, Baopeng Zhang, Zhengzhong Kuang, Dan Lin, and Jianping Fan. iprivacy: image privacy protection by identifying sensitive objects via deep multi-task learning. *IEEE Transactions on Information Forensics and Security*, 12(5):1005–1016, 2016.