Estimating (and *fixing*) the Effect of Face Obfuscation in Video Recognition

Matteo Tomei, Lorenzo Baraldi, Simone Bronzin[†], Rita Cucchiara
University of Modena and Reggio Emilia, [†]Metaliquid Srl
{name.surname}@unimore.it, [†]simone.bronzin@meta-liquid.com

Abstract

Recent research has shown that faces can be obfuscated in large-scale datasets with a minimal performance impact on image classification and downstream tasks like object recognition. In this paper, we investigate the role of face obfuscation in video classification datasets and quantify a more significant reduction in performance caused by face blurring. To reduce such performance effects, we propose a generalized distillation approach in which a privacy-preserving action recognition network is trained with privileged information given by face identities. We show, through experiments performed on Kinetics-400, that the proposed approach can fully close the performance gap caused by face anonymization.

1. Introduction

As the role of Deep Learning is becoming increasingly important in the development of Artificial Intelligence solutions [3, 24, 12, 8], and as AI algorithms are becoming more and more pervasive, preserving privacy becomes fundamental. Nevertheless, most visual understanding models are trained to work on unaltered images and videos which could contain private information, such as people faces from which identities can be revealed.

With the objective of protecting users' privacy, trained models should ideally work on images and videos in which all private information is obfuscated. In practice, this is hardly feasible, and it may hurt the utility of data samples if too much information is removed. Obfuscating faces is, nevertheless, a reasonable first step towards privacy-preserving visual understanding models.

Recent literature [31] has investigated the role of obfuscating faces in the training set of image classification architectures and demonstrated that this has minimal impact on the accuracy of recognition models. As a further step in the same direction, in this paper we focus on video recognition architectures, and investigate the role of face anonymization in such networks.

Further, we develop a training strategy for privacy-preserving video prediction architectures, which can work on anonymized videos while still guaranteeing high accuracy levels. Our proposal is based on a knowledge distillation schema in which a network is trained to work on anonymized videos while reading knowledge from a network that works on full, non-anonymized, videos – and which therefore can exploit privileged information. Experimentally, we validate the proposed solution on the Kinetics-400 dataset, employing three different video backbones.

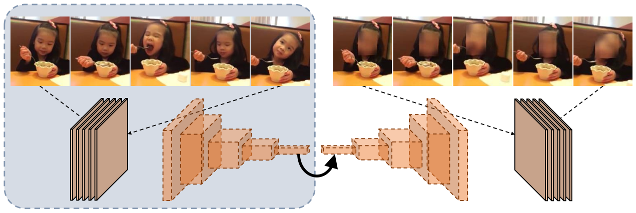

Figure 1. We propose a training schema that can prevent the performance gap caused by face anonymization in video classification networks. A network is trained to work on anonymized images by reading knowledge from a network that works on full images.

Contributions To sum up, our contributions are as follows:

- We investigate and quantify the performance gap in video classification networks when face obfuscation is applied to input clips. Through experiments conducted on Kinetics-400, we find that the anonymization performance gap is considerably more severe than the one observed in image classification datasets.
- We propose a solution to close the performance gap which is caused by face anonymization on video networks. Our solution is based on a Generalized Distillation approach which combines knowledge transfer and access to privileged information. Further, we employ a relational criterion that encourages the transfer of mutual relationships between samples.
- Experiments, conducted on Kinetics-400 and on different video backbones, demonstrate the effectiveness of our proposal. The final privacy-compliant network achieves *better* results than the original backbone while working only on anonymized faces.

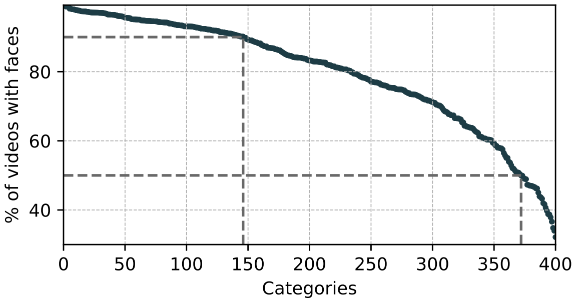

Figure 2. Per-class percentage of videos containing at least one face, on the Kinetics-400 training set.

2. Related Works

Action Recognition Models. Several efforts in designing new network architectures for action analysis have been recently made, largely increasing performance on standard video action recognition datasets [9, 10, 21]. Several approaches have been proposed, exploiting 3D convolutional kernels [3, 25], hence modeling spatio-temporal information together, or combining 2D spatial and 1D temporal kernels [18, 26]. It is also common to use two paths for appearance and motion analysis [3, 5, 20, 29]. Attention-like operators have also been explored in recent works on action recognition [1, 15, 23].

Distilling the knowledge. The technique of distilling knowledge from a teacher network to a student one was originally proposed in [30]. This technique has been then extended to several tasks, and new strategies for transferring knowledge have been proposed [30]. Several approaches extend or propose new KD methodologies for image-related tasks [7, 32, 33, 35], while only a few consider videos [2]. Taking inspiration from privileged information [11, 13, 28], we propose to exploit the knowledge of a teacher which has access to additional information compared to the student (*i.e*. people identity). Other works propose instead to transfer relations between samples, instead of their representations directly [16, 22].

Privacy in Computer Vision. Privacy-related issues are extremely important for vision tasks involving people identities. Uittenbogaard *et al*. [27] propose to remove and inpaint moving objects (like pedestrian) in street-view imagery. Piergiovanni *et al*. [17] collect a dataset of action videos with anonymized face identities. A study on the effect of face obfuscation on ImageNet [4] has been presented in [31]. In [19], the authors propose an adversarial approach with a video anonymizer that tries to remove sensitive information while still solving action detection, and a discriminator that is trained to recover privacy-sensitive information from the modified videos.

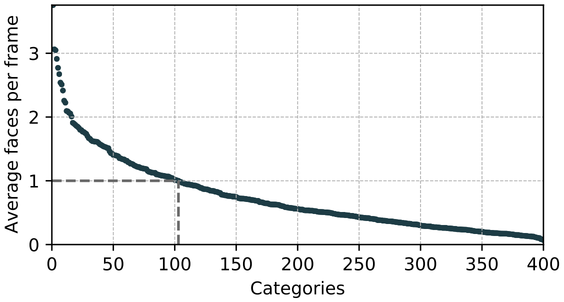

Figure 3. Average number of per-frame faces, for each category of the Kinetics-400 training set.

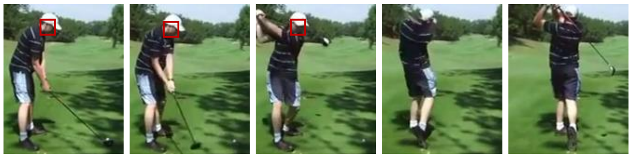

Figure 4. A sample video clip from Kinetics-400, annotated with face bounding boxes.

3. Analyzing faces in action videos

In this section, we first introduce a simple way for preserving privacy in video clips, by assuming that the actor's identity can be hidden by masking its face. We then conduct analyses on Kinetics-400 [9], one of the largest publicly available video classification databases, to quantify the presence of faces in video classification datasets and motivate the appropriateness of building privacy-preserving video networks.

Detecting and anonymizing faces in videos. Given a target video classification dataset $\mathcal{D}$, we independently apply a pre-trained frame-level face detector [12] to all the video frames in the dataset, obtaining face coordinates for all the involved actors. We then build an anonymized version of the initial dataset, $\mathcal{D}^b$, by blurring all the faces found in $\mathcal{D}$ with a normalized box filter. In our initial experiments, we found that applying a frame-level detection strategy on Kinetics-400 [9] is enough to guarantee complete anonymization of all face instances in most of the videos, without sorting to the usage of tracking techniques.

Faces in the Kinetics-400. Kinetics-400 [9] consists of approximately 240k training videos and almost 20k validation videos. Each video lasts around 10s and belongs to one of 400 different human action categories. Since video clips mainly come from YouTube, there are no constraints on camera motion, illumination, and viewpoint.

To assess the bearing of faces in the dataset, and hence its possible influence in action recognition, we measure the amount of bounding boxes found in $\mathcal{D}$. Figure 2 reports the per-class percentage of training videos with faces, *i.e*. that have at least one frame containing a bounding box predicted by the face detector model [12]. As it can be seen, in the majority of the classes (372 out of 400), more than 50% of the available videos contain faces. For 146 categories, instead, more than 90% of videos show at least one person identity, and there are no classes with less than 30% of videos containing faces. This highlights the importance of building privacy-preserving video classification techniques, especially when handling user-generated videos.

We further investigate the number of different identities in Kinetics-400 videos. In Figure 3, indeed, we report the average number of faces per-frame, for all categories. Out of 400 classes, 103 have more than one face per frame, on average. The majority of classes have between zero and one face per frame, underlining that most of the actions involve a single actor. It must be also noted, although, that for some actions it is common that the actor's face is visible only for a limited period of time, while being occluded or not visible in the remaining frames (like in the *golf driving* sample reported in Figure 4).

4. Privacy-preserving action recognition

Motivated by the prevalence of face identities in video classification datasets, we aim at building a privacy-preserving video model which can deal with obfuscated faces while still maintaining high accuracy levels. To this end, we consider the incorporation of an "intelligent teacher" which has access to privileged, non-anonimyzed, training data. We consider our training data as formed by a collection of triplets

$$\{(x_1, x_1^*, y_1), ..., (x_n, x_n^*, y_n)\}, \tag{1}$$

where each pair (x_i, y_i) is a video-label pair drawn from the anonymized dataset $\mathcal{D}^b$, and x_i^* is the non-anonymized version of the video from $\mathcal{D}$, which discloses the face identity and thus personal information. In our setting, x_i^* provides additional information which can be used at training time to support the learning process, while the final model will not have access to x_i^* at test time.

4.1. Exploiting Knowledge Distillation

To exploit the privileged information given by non-anonymized videos, we propose to employ knowledge distillation. In short, our proposal is as follows: (1) we first learn a teacher network, which is trained on privileged data drawn from $\mathcal{D}$, *i.e.* using pairs $\{x_i^*, y_i\}$. (2) Once the teacher model is trained, we compute its soft labels to distill the model into a student model with the same architecture as the teacher. (3) We train the student model by employing both softly-labeled data from the teacher and hard-labeled data from the anonymized dataset $\mathcal{D}^b$, *i.e.* pairs $\{x_i, y_i\}$.

Teacher training with privileged information. Given a video backbone architecture, we first build a teacher model f_t by training the architecture on the non-anonymized dataset $\mathcal{D}$. Considering that the network will need to solve a classification problem, we learn the representation

$$\arg\min_{\boldsymbol{W}} \frac{1}{n} \sum_{i=1}^{n} \ell(\sigma(f_t(x_i^*)), y_i), \tag{2}$$

where x^* is an input video clip belonging to $\mathbb{R}^{T\times H\times W\times 3}$ and $\boldsymbol{W}$ indicates the set of weights of the teacher. Denoting with Δ^c the set of c-dimensional probability vectors, σ is a softmax operator $\sigma : \mathbb{R}^c \rightarrow \Delta^c$ defined as

$$\sigma(z)_k = \frac{e^{z_k}}{\sum_{j=1}^{c} e^{z_j}} \tag{3}$$

for all $k \in \{1...c\}$. Further, $\ell : \Delta^c \times \Delta^c \rightarrow \mathbb{R}$ is the standard cross-entropy loss function, *i.e.*

$$\ell(\hat{y}, y) = -\sum_{k=1}^{c} y_k \log \hat{y}_k. \tag{4}$$

Student training. We then train a student model which will only employ anonymized video clips x_i drawn from $\mathcal{D}^b$, and which will distill knowledge from a teacher trained using privileged, non-anonymized, information. For the sake of simplicity, we build a model from the same backbone architecture employed for the teacher. During training, we distill [8] the representation learned in the teacher into

$$\arg\min_{\boldsymbol{W}} \frac{1}{n} \sum_{i=1}^{n} \left[\ell(\sigma(f_s(x_i)), y_i) + \lambda \ell_d(\hat{s_i}, s_i)\right], \tag{5}$$

where, here, $\boldsymbol{W}$ indicates weights of the student model, ℓ_d a knowledge distillation loss, while $\hat{s_i}$ and s_i indicate soft-predictions from the student and teacher model, respectively. These are computed as

$$\begin{aligned} \hat{s_i} &= \sigma(f_s(x_i)/T) \in \Delta^c, \\ s_i &= \sigma(f_t(x_i^*)/T) \in \Delta^c. \end{aligned} \tag{6}$$

The temperature parameter $T > 0$ controls the smoothening of the probability predictions from the teacher, and $\lambda > 0$ controls the balance between learning from anonymized data sampled from $\mathcal{D}^b$ and transferring knowledge from f_t.

As it can be noticed, although there is knowledge transfer from the teacher to the student, this happens only at the logit level, thus ensuring a potentially high obfuscation level of sensible data. In other words, while f_t has access to non-anonymized training data, only high-level features without spatial support are actually considered during knowledge transfer.

As knowledge distillation loss, we employ the Kullback-Leibler divergence between the output probabilities of f_s and f_t. This is defined as

$$\ell_d(\hat{s}, s) = -\sum_{k=1}^{c} s_k \log\left(\frac{\hat{s_k}}{s_k}\right). \tag{7}$$

Relational Knowledge Distillation. To further increase the transfer of privileged information, we also employ a relational knowledge distillation criterion [16], which aims to transfer mutual relations between data samples in the mini-batch, beyond activations from individual samples.

Specifically, given the logits computed by one of the networks on two video clips, we compare them by computing their normalized Euclidean distance:

$$\psi_D(p_i, p_j) = \frac{1}{\mu} \left\| p_i - p_j \right\|_2 , \tag{8}$$

where μ is equal to the average Euclidean distance between pairs of representations inside the mini-batch.

The relational knowledge distillation loss matches pairwise distances generated by the teacher and student models, as follows:

$$\ell_r = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} l_\delta(\psi_D(\hat{p}_i, \hat{p}_j), \psi_D(p_i, p_j)) \tag{9}$$

where $\hat{p_i}$ and $\hat{p_j}$ are the logits computed by the student network from two video clips x_i and x_j, while p_i and p_j are the logits computed by the teacher network on the corresponding x_i^* and x_j^*. Finally, l_δ is the Huber loss:

$$l_\delta(x, y) = \begin{cases} \frac{1}{2}(x-y)^2 & \text{for } |x-y| \leq 1 \\ |x-y| - \frac{1}{2} & \text{otherwise.} \end{cases} \tag{10}$$

The relational KD criterion forces the distance between the representations of (x_i^*, x_j^*) as computed by the teacher to be as close as possible to the distance between the representations of (x_i, x_j) as computed by the student. When using also the relational knowledge distillation term, the final representation which is learned is:

$$\begin{aligned} \arg\min_{\boldsymbol{W}} \frac{1}{n} \sum_{i=1}^{n} [\ell(\sigma(f_s(x_i)), y_i) + \\ + \lambda \ell_d(\hat{s}_i, s_i) + \beta \ell_r(f_s(x_i), f_t(x_i^*))] , \end{aligned} \tag{11}$$

where β controls the relational knowledge distillation term.

5. Experiments

5.1. Implementation details

We employ a ResNet(2+1)D-18 backbone for our analyses and experiments, and further adopt a ResNet3D-18 [26] and a SlowFast-R50 [5] backbone to showcase the applicability of our proposal to other networks.

R(2+1)D and R3D. We adopt the 18-layers version of ResNet(2+1)D and ResNet3D and train them following the original implementation presented in [26]: during training, frames are resized to 128 × 171 and randomly cropped to 112 × 112. Each input clip consists of 16 consecutive frames. During each epoch, three clips per video are randomly sampled for temporal jittering. Training is performed with synchronized SGD on multiple NVIDIA V100 GPUs, for a total mini-batch size of 1536 and 45 epochs. The learning rate is set to 0.96 with a linear warm-up in the first 10 epochs and then divided by 10 at epochs 20, 30 and 40. At inference time, 112 × 112 center crops are extracted from 10 uniformly sampled clips from the video and their predictions are averaged.

SlowFast-R50. We adopt the ResNet50 [6] based SlowFast implementation. SlowFast networks consist of a Slow pathway and a Fast pathway, and have been originally proposed in [5]. During training, the shorter side of the input clip is resized to a random value between 256 and 320, maintaining the aspect ratio. Random crop is then applied, to reduce the clip to a size of 224 × 224. From a 64-frames clip, the Slow network samples 4 frames (one every 16 frames), while the Fast network samples 32 frames (one every 2 frames). Again, 3 clips are randomly sampled from a video in each epoch, and SGD is adopted with a total mini-batch size of 2048. The base learning rate is set to 1.6 with linear warm-up during the first 5 epochs and cosine annealing afterward. Training lasts for 45 epochs. During inference, we extract 256 × 256 center crops from 10 uniformly sampled clips and average their predictions.

Additional details. In all experiments, the λ value is set to 12.500, while the β value in Eq. 11 is set to 10. The temperature parameter T is set to 5.

5.2. Impact of Face Obfuscation

First, we analyze the performance of a video understanding model when working on anonymized faces. Considering a ResNet(2+1)D-18 [26] backbone trained on the standard Kinetics-400 dataset (with non-anonymized faces), its top-1 accuracy reaches 69.2 when tested on the original, non-anonymized test set. When the same network is instead tested on the anonymized version of the test set (*i.e.*, when blurring faces), its top-1 accuracy drops to 67.2, which corresponds to a -2.9% relative drop. This highlights that removing identities has a significantly negative impact on action recognition (Table 1, first two rows).

To verify to which extent the aforementioned accuracy drop is due to the lack of information caused by anonymization, or to the different distribution between the training and testing sets, we also train the same network on blurred faces (Table 1, third row). In this setting, the gap is significantly reduced, and the top-1 accuracy increases to 68.6. Still, there is a non-negligible loss in accuracy with respect to a standard action recognition model when employed on non-anonymized videos.

Model	Train set	Test set	top-1	top-5	$\Delta_{\text{top-1}}$
R(2+1)D [26]	$\mathcal{D}_{train}$	$\mathcal{D}_{test}$	**69.2**	**88.1**	–
R(2+1)D [26]	$\mathcal{D}_{train}$	$\mathcal{D}^b_{test}$	67.2	86.3	-2.9 %
R(2+1)D [26]	$\mathcal{D}^b_{train}$	$\mathcal{D}^b_{test}$	68.6	87.7	-0.9 %

Table 1. Performances of the R(2+1)D model [26] for video classification when trained and tested on anonymized and non-anonymized videos. $\mathcal{D}$ indicates the original, non-anonymized dataset; $\mathcal{D}^b$ its anonymized (blurred) version.

5.3. Main results

In the following, we assess the effectiveness of the proposed training schema in reducing the performance gap between a standard action classification network and a privacy-preserving action network. Since the proposed approach involves distillation, and it is well known that self-distillation alone helps in achieving higher performances [14, 34, 35], we first assess the impact of self-distillation in our experimental setting, without using face anonymization.

In Table 2 we report the top-1 and top-5 accuracy of a ResNet(2+1)D-18 [26] trained using knowledge distillation, and using another ResNet(2+1)D-18 as teacher. We empirically find that the model trained on the original, non-anonymized Kinetics-400 dataset, using the knowledge distillation loss (Eq. 5), reaches **70.6** top-1 accuracy on the Kinetics-400 validation set (Table 2, first row), compared to a 69.2 top-1 accuracy of the model trained with cross-entropy only (Table 1, first row). Further, when adding the relational KD criterion the model reaches a 70.4 top-1 accuracy, highlighting that the relational KD brings no advantage in this setting.

Having quantified the role of self-distillation, we now turn to the evaluation of the proposed schema when training privacy-preserving networks. In the last two rows of Table 2, we show the results obtained when using the KL loss with and without the relational KD criterion, and training the student on the anonymized version of Kinetics-400. As it can be seen, when using the KL divergence loss the model achieves a 70.3 top-1 accuracy. When adding the relational KD criterion, instead, the model reaches a 70.4 top-1 accuracy, thus almost closing the accuracy gap with the non-anonymized model (70.4 vs 70.6).

Finally, in Table 3 we show the results of applying the proposed training schema to other video backbones. When using ResNet3D-18 [26] or SlowFast-R50 [5], removing face information and using knowledge distillation still improves the performances compared to the original non-privacy-preserving models. In Table 3, for each backbone, the first row represents the base model trained with cross-entropy on the original videos, while the second row shows results of the model trained with knowledge distillation and the relational KD criterion on videos with obfuscated faces, using soft targets from the base model.

Model	Train set	Test set	KL	RKD	top-1	top-5
R(2+1)D [26]	$\mathcal{D}_{train}$	$\mathcal{D}_{test}$	✓	✗	70.6	89.1
R(2+1)D [26]	$\mathcal{D}_{train}$	$\mathcal{D}_{test}$	✓	✓	70.4	88.8
R(2+1)D [26]	$\mathcal{D}^b_{train}$	$\mathcal{D}^b_{test}$	✓	✗	70.3	88.8
R(2+1)D [26]	$\mathcal{D}^b_{train}$	$\mathcal{D}^b_{test}$	✓	✓	70.4	88.7

Table 2. Performances of the R(2+1)D model [26] for video classification when trained using Knowledge Distillation (KL) and/or the relational KD criterion (RKD) with privileged information.

Model	Priv.-preserv.	Train set	Test set	top-1	top-5
R3D [26]	✗	$\mathcal{D}_{train}$	$\mathcal{D}_{test}$	65.3	85.5
R3D [26]	✓	$\mathcal{D}^b_{train}$	$\mathcal{D}^b_{test}$	**66.4**	**86.4**
SlowFast [5]	✗	$\mathcal{D}_{train}$	$\mathcal{D}_{test}$	73.5	**91.3**
SlowFast [5]	✓	$\mathcal{D}^b_{train}$	$\mathcal{D}^b_{test}$	**73.9**	**91.3**

Table 3. Performances of the R3D model [26] and SlowFast [5] and of their privacy-preserving counterparts, trained using Knowledge Distillation (KL) and the relational KD criterion (RKD) with privileged information.

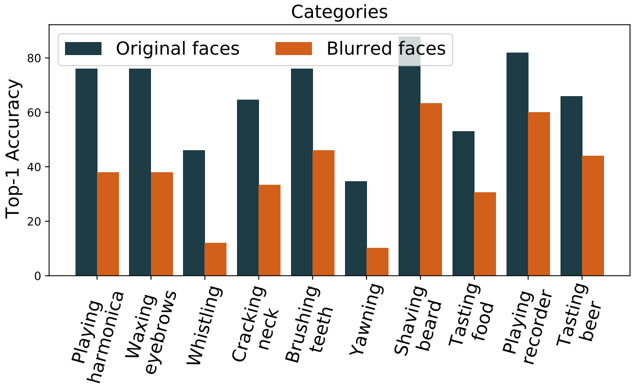

Figure 5. Per-class impact of anonymizing faces, on a ResNet(2+1)D-18 [26] trained on the original Kinetics-400.

5.4. Per-class performance

Finally, we study the per-class impact of face obfuscation on the Kinetics-400 validation set, when using a privacy-preserving model and when using a classical video prediction model. Figure 5 shows the ten classes with the highest gap in top-1 accuracy when testing a base ResNet(2+1)D-18 [26] on the original validation set (blue bars, 69.2 overall top-1 accuracy) and on its anonymized counterpart (orange bars, 67.2 overall top-1 accuracy). In both cases, the model has been trained on the original, non-anonymized videos. As expected, the actions showing the highest accuracy loss involve the actor's face and/or objects which are blurred because they overlap with the face itself. Among the considered ten classes, the ones with the max-

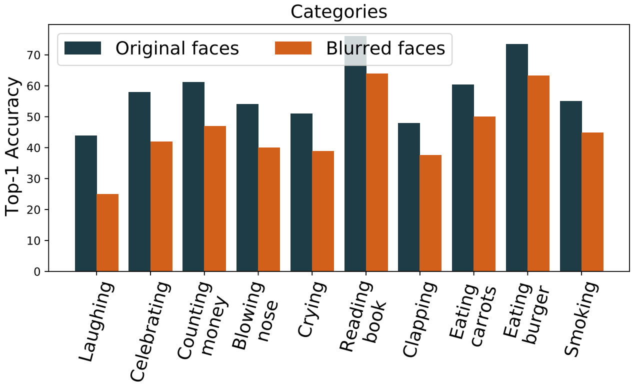

Figure 6. Per-class performances of ResNet(2+1)D-18 [26] models trained with our strategy, one on anonymized and one on non-anonymized data.

imum and minimum accuracy loss are *playing harmonica* and *tasting beer*, with 38% and 22% absolute drop in accuracy, respectively.

In Figure 2 we instead report the same comparison for two ResNet(2+1)D-18 [26], both trained with our knowledge distillation strategy. In the first case (blue bars, overall 70.6 top-1 accuracy) the model is trained and tested on non-anonymized videos, while in the latter (orange bars, overall 70.3 top-1 accuracy) the model is trained and tested on anonymized videos. As it can be observed, the classes with the highest gap are now different from those in Figure 5, even if they still correspond to actions involving the actor's face. The maximum and minimum gap for the considered ten classes are 18.8% and 10.2% for *laughing* and *smoking*, respectively.

6. Conclusion

In this work, we investigated the role of face obfuscation for developing privacy-preserving video prediction networks. To this end, we presented a pipeline to let a video understanding network maintain or even increase its performance when removing people identity information from its input clips. This is achieved by exploiting privileged information from a teacher model, which has access to people faces, and by distilling its logits to a student network, which does not have access to identity-revealing visual details. The generalization of the approach has been demonstrated through experiments with three different backbones, on the Kinetics-400 dataset.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021.

[2] Shweta Bhardwaj, Mukundhan Srinivasan, and Mitesh M Khapra. Efficient video classification using fewer frames. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009.

[5] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the International Conference on Computer Vision*, 2019.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[7] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the International Conference on Computer Vision*, 2019.

[8] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *Advances in Neural Information Processing Systems Workshops*, 2015.

[9] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[10] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision*, 2011.

[11] Wonkyung Lee, Junghyup Lee, Dohyung Kim, and Bumsub Ham. Learning with privileged information for efficient image super-resolution. In *European Conference on Computer Vision*. Springer, 2020.

[12] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: Dual shot face detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[13] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *Proceedings of the International Conference on Learning Representations*, 2016.

[14] Hossein Mobahi, Mehrdad Farajtabar, and Peter L Bartlett. Self-distillation amplifies regularization in hilbert space. *Advances in Neural Information Processing Systems*, 2020.

[15] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. *arXiv preprint arXiv:2102.00719*, 2021.

[16] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[17] AJ Piergiovanni and Michael S Ryoo. Avid dataset: Anonymized videos from diverse countries. *Advances in Neural Information Processing Systems*, 2020.

[18] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *Proceedings of the International Conference on Computer Vision*, 2017.

[19] Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. Learning to anonymize faces for privacy preserving action detection. In *Proceedings of the European Conference on Computer Vision*, 2018.

[20] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, 2014.

[21] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[22] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *Proceedings of the International Conference on Learning Representations*, 2020.

[23] Matteo Tomei, Lorenzo Baraldi, Simone Calderara, Simone Bronzin, and Rita Cucchiara. Video action detection by learning graph-based spatio-temporal interactions. *Computer Vision and Image Understanding*, 206, 2021.

[24] Matteo Tomei, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Image-to-image translation to unfold the reality of artworks: an empirical analysis. In *International Conference on Image Analysis and Processing*, 2019.

[25] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the International Conference on Computer Vision*, 2015.

[26] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[27] Ries Uittenbogaard, Clint Sebastian, Julien Vijverberg, Bas Boom, Dariu M Gavrila, et al. Privacy protection in street-view panoramas using depth and multi-view imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[28] Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: similarity control and knowledge transfer. *J. Mach. Learn. Res.*, 2015.

[29] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Proceedings of the European Conference on Computer Vision*, 2016.

[30] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[31] Kaiyu Yang, Jacqueline Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A study of face obfuscation in imagenet. *arXiv preprint arXiv:2103.06191*, 2021.

[32] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[33] Kaiyu Yue, Jiangfan Deng, and Feng Zhou. Matching guided distillation. In *Proceedings of the European Conference on Computer Vision*, 2020.

[34] Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. Regularizing class-wise predictions via self-knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[35] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the International Conference on Computer Vision*, 2019.