ARIA: Adversarially Robust Image Attribution for Content Provenance

Maksym Andriushchenko* EPFL
Xiaoyang Rebecca Li Adobe Research
Geoffrey Oxholm Adobe Research
Thomas Gittings University of Surrey
Tu Bui University of Surrey
Nicolas Flammarion EPFL
John Collomosse Adobe Research

Abstract

Image attribution – matching an image back to a trusted source – is an emerging tool in the fight against online misinformation. Deep visual fingerprinting models have recently been explored for this purpose. However, they are not robust to tiny input perturbations known as adversarial examples. First we illustrate how to generate valid adversarial images that can easily cause incorrect image attribution. Then we describe an approach to prevent imperceptible adversarial attacks on deep visual fingerprinting models, via robust contrastive learning. The proposed training procedure leverages training on ℓ_∞-bounded adversarial examples, it is conceptually simple and incurs only a small computational overhead. The resulting models are substantially more robust, are accurate even on unperturbed images, and perform well even over a database with millions of images. In particular, we achieve 91.6% standard and 85.1% adversarial recall under ℓ_∞-bounded perturbations on manipulated images compared to 80.1% and 0.0% from prior work. We also show that robustness generalizes to other types of imperceptible perturbations unseen during training. Finally, we show how to train an adversarially robust image comparator model for detecting editorial changes in matched images. Project page: https://max-andr.github.io/robust_image_attribution.

1. Introduction

Fake news and misinformation are major societal threats being addressed by new computer vision methods to determine content authenticity. Such methods fall into two camps: detection and attribution. Detection methods automatically identify manipulated or synthetic images through visual artifacts or statistics [59, 60, 65]. Attribution methods match an image to a trusted database of originals [5, 6, 45]. Once matched, any differences may be visualized, and any associated provenance data displayed. Rather than making automated judgments, the goal of image attribution is to enable users to make more informed trust decisions [24].

This paper considers specifically the *image attribution* problem where the goal is to differentiate between 'non-editorial' transformation of content (*e.g.* due to resolution, format or quality change) and editorial change where content is digitally altered to change its meaning. Nguyen et al. [45] use contrastive training to learn a visual hashing function that is invariant to non-editorial, but sensitive to editorial changes. In such 'tamper-sensitive' matching, a manipulated image would not be falsely corroborated by provenance data associated with the original. By contrast, Black et al. [6] learn a 'tamper-invariant' image fingerprint which is insensitive to *both* non-editorial and editorial change, and visually highlight manipulated changes using a separate model.

This paper reports on the novel problem of *adversarial attack and defense for these image attribution approaches*. They rely on deep neural networks to learn visual fingerprints for near-duplicate image matching. However, deep networks are known to be vulnerable to *adversarial attacks* [52] that use subtle image perturbations to cause dramatic changes in the output of the models. Visual search models based on deep networks are not exceptions and recently have been shown to be also vulnerable to adversarial attacks [15].

We make the following technical contributions:

1. Adversarial attack of image attribution models. We present a white-box gradient-based method for crafting adversarial examples to attack both tamper-sensitive and tamper-invariant image attribution models. We show it is possible to closely match the perceptual fingerprints of unrelated images by using small ℓ_∞-adversarial perturbations. For tamper-sensitive models, we show that the original image may be incorrectly matched given a manipulated query. For tamper-invariant models, we additionally show that the image comparison post-process used to visualize areas of image manipulation may also be fooled to show either no manipulation or false areas of manipulation. Thus, we show that trust in both state-of-the-art image attribution approaches may be defeated by adversarial examples.

2. Robust contrastive learning for image attribution. We describe a novel robust contrastive learning algorithm

*Work done as internship at Adobe Research.

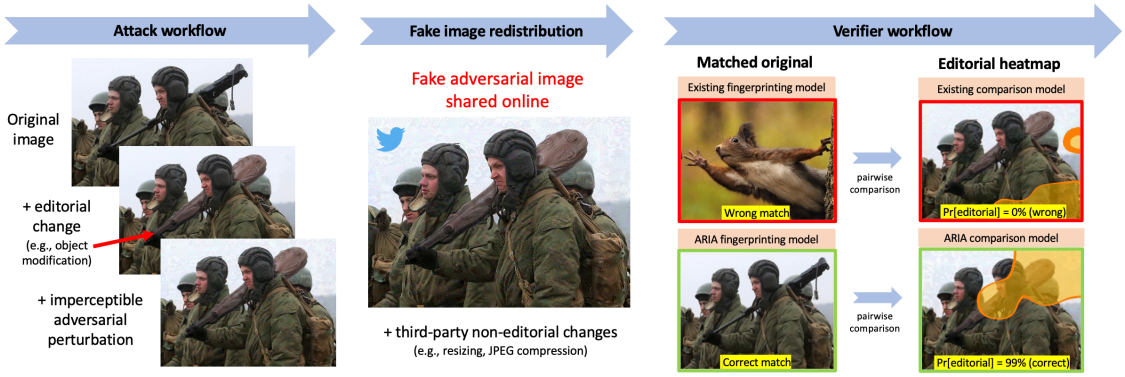

Figure 1. Workflow of image attribution. ARIA applied to the framework of Black et al. [6] enables both robust *matching* and *comparison* of manipulated images, despite the presence of imperceptible adversarial attacks. We consider an attack workflow in which an attacker adds both editorial changes (e.g., an object modification or a face swap) and an adversarial perturbation prior to the distribution of the adversarial image online which may involve additional non-editorial changes (e.g., resizing, JPEG compression). Existing image attribution models can be fooled by such attacks: the image fingerprinting model produces an irrelevant match and the comparison model predicts either no change or false editorial changes. At the same time, our ARIA training produces models which are both accurate and adversarially robust.

to train image fingerprinting models for attribution, to ensure robustness both to non-editorial transformations *and* imperceptible adversarial perturbations, preventing attacks illustrated in Fig. 1. We show that this algorithm improves adversarial robustness of both tamper-sensitive [45] and tamper-invariant [6] image fingerprinting models, and we also discuss how to make the image comparator model [6] robust. The approach is conceptually simple and leads to a relatively small computational overhead ($\approx 2\times$ slowdown) compared to standard contrastive learning. We also show that our adversarially-robust image hashing models have benefits in terms of interpretability: they output perceptually similar images under hash inversion attacks [51].

Our work comprehensively studies adversarial robustness for the growing body of image attribution work, and is timely given the emergence of cross-industry standards reliant, in part, on such attribution. For example, the Coalition for Content Provenance and Authenticity (C2PA) [1] proposes to fight misinformation by embedding secure audit information on content manipulation within image metadata. Many online media distribution channels (such as social media sites) strip away this metadata. C2PA proposes to counter this by the use of perceptual hashes for image attribution, and advocates computing the hashes on the client side due to privacy concerns. This implies, as with our work, that the attacker has *white-box* access to the target model (*i.e.* the attacker is assumed to know all details of the system being attacked). Our work is the first to demonstrate the significance of these attacks on attribution models. Moreover, by offering the first defense against adversarial attacks for such models, we contribute to the protection of provenance systems implementing such standards.

2. Related Work

Image fingerprinting for provenance. Image fingerprinting models robust to non-editorial transformations were proposed in Black et al. [6] and Nguyen et al. [45]. These represent two complementary approaches to applying image retrieval to the attribution problem. Both approaches match query images to a trusted database of originals, invariant to non-editorial changes such as resolution, format or quality change. However, the approaches differ in their consideration of manipulated images. Black et al. [6] train the image retrieval model to match manipulated images to originals successfully, i.e., to bring such images pairs close together in the search embedding. By contrast, Nguyen et al. [45] train the model to separate such image pairs, encouraging matching to fail in the presence of content manipulation. The advantage of [45] is a simpler pipeline as it has only an image retrieval model (with an optional geometric verification step), while [6] relies on a separate image comparator network that analyzes whether a pair of images are identical, different or have been manipulated, visualizing the difference. Whilst robust to non-editorial distortions, both approaches are not robust to adversarial attacks, motivating our work.

Adversarial attacks on deep neural networks for image classification were pioneered by Szegedy et al. [52], who demonstrated that minor perturbations of pixel values are sufficient to induce significant classification mistakes despite little perceptible difference ('covert' approaches). Goodfellow et al. [20] demonstrated linearity of this effect in input space, introducing the fast gradient sign method (FGSM) to quickly compute adversarial perturbations via backpropagation without the need for solving costly optimizations. An iterative form of this method for more robust attacks was later presented [39]. Such attacks have received significant attention in recent years with many variants proposed to covertly attack image classifiers [13, 21, 44]. Adversarial patches take a complementary, 'overt', approach via synthesis of vivid 'stickers' [7] that occupy only a small region yet induce misclassification [7, 18] or misdetection [10, 55].

Recently, adversarial attacks on image retrieval models have been demonstrated via similar means. Tolias et al.

[56] show that image retrieval models are non-robust. They perform targeted attacks in the white-box and semi black-box setting (unknown pooling). Bai et al. [4] and Dolhansky and Ferrer [15] show that image *hashing* models can be fooled as well, including attacks which exactly produce target hashes.

Contrastive learning. Several popular self-supervised learning approaches are based on contrastive learning: SimCLR [12], MoCo [27], and BYOL [25]. Most robust self-supervised approaches focus on robust transfer learning [11, 23, 30, 33, 35, 37, 63] or multi-objective optimization [8, 31, 43] to improve adversarial robustness. The focus of these works differ from our focus on image retrieval. In particular, they do not benchmark image retrieval performance and for training, they rely on augmentations that are optimized for transfer learning and not for image attribution (e.g., extreme crops used in SimCLR: between 8% and 100% of the area as in Chen et al. [12]). E.g., Kim et al. [37] propose a two-stage robust training approach: first generating instance-wise adversarial examples for the SimCLR loss and then combining together the standard SimCLR loss with the robust loss. Tamkin et al. [53] propose to use an image-to-image network that generates adversarial examples which are then projected onto a (relatively large) ℓ_1-ball and they do not cover adversarial robustness. However, neither of these approaches is considered in the image retrieval setup. [49] propose adversarial training in the *latent* space with the goal of improving standard generalization. Closer to image attribution, Panum et al. [46] combine deep metric learning algorithms with adversarial training but they perform only small-scale experiments and evaluate their models via nearest neighbour classification which is different from the retrieval under editorial and non-editorial distortions as we do in the context of image attribution.

3. Vulnerability of Attribution Models

We start from studying adversarial vulnerability within the context of the image attribution approaches of Nguyen et al. [45] and Black et al. [6].

1. Image fingerprinting (IF). We consider an IF model $\phi : \mathbb{R}^d \to \mathbb{R}^D$ which performs the mapping of an image to its D-dimensional feature vector (fingerprint) used to output the most similar image from a database $\mathcal{X}$. We denote by x an original image contained in $\mathcal{X}$, and x_{query} the image x modified by some transformation. For all IF methods, we wish the match to be invariant to a set of non-editorial transformations $\bar{E}(x)$, and in some cases (*e.g.* Black et al. [6]) also editorial manipulations of the image $E(x)$.

In all cases we consider adversarial perturbations of x specific to the model ϕ generated under some perturbation budget ϵ. We call the perturbations *targeted* or *untargeted* perturbations depending on whether an attack targets retrieval of a specific incorrect image, or its objective is simply to prevent retrieval of the correct image. In the case of methods seeking to retrieve x for $x_{query} \in \bar{E}(x)$ only [45], a specific form of targeted attack attempts to fool ϕ to retrieve its original as illustrated in Fig. 2.

2. Image comparison (IC). IC methods visualize pixel regions containing *editorial* changes, performing an 'intelligent differencing' operation between x_{query} and the top retrieval from the IF model, ignoring any visual change due to non-editorial transformations. We consider the approach of [6] where the model outputs both such a visualization (a 7×7 heatmap) and additionally assigns the image pair to three categories: same images with non-editorial changes, same images with editorial changes or different images. The first goal of the attacker is to make the comparator classify an image with editorial changes to the first category. The second goal is to make the comparator output a misleading heatmap describing editorial changes (see Fig. 3).

3.1. Adversarial Attack Scope

We consider attacks within the following scope. First, adversarial perturbations should be **imperceptible** such as perturbations bounded within a small ℓ_∞- or ℓ_2-norm. Imperceptibility is a crucial property as, from an attack perspective, a user should not realize an image has been manipulated in any way. This requirement makes, e.g., patch-based [7], ℓ_1- or ℓ_0-bounded perturbations not relevant in our case since they render tampering visually obvious. Second, we consider a **white-box** knowledge model: all details of the model are assumed known. This prevents the so-called *security-by-obscurity* where the security or robustness of a system relies on such detail (e.g., model architecture or parameters) being secret. In practice, such detail can be leaked or reverse engineered, particularly for models deployed on edge devices or client-side (as is advocated in emerging standards [1]). Third, we focus on **fully realizable attacks**, i.e., after generating adversarial examples, we save them as valid JPEG files and only then evaluate the model on them. This differs from most prior works, where imperceptible adversarial images are generated such that they have arbitrary real values in the range $[0, 1]$ [41, 52]. Instead, after saving them as JPEG files, the pixel values are quantized to 8-bit and the image is compressed introducing further non-editorial changes. We show in Sec. 5 that even standard attacks described below (as opposed to robust attacks [3]) are sufficient for the considered attack scenario as we can successfully reduce many performance metrics to zero.

3.2. Implementation of adversarial attacks

We generate adversarial attacks using projected gradient descent (PGD) [41] under ℓ_∞-norm constraints since the gradients are available[1] in the white-box setting and the ℓ_∞-norm is a useful proxy for the imperceptibility requirement.

Image fingerprinting attack. The goal of an untargeted attack is to make the resulting adversarial example $x_{query}+\delta$ have an IF sufficiently different from the original one so that $x_{query} + \delta$ gets matched to an incorrect image. For this, we can *maximize* the ℓ_2-distance between an IF of $x_{query} + \delta$

[1] We use differentiable image resizing to make sure that the whole image preprocessing pipeline is differentiable.

Figure 2. Attacks on image fingerprinting (IF) models. Visualization of untargeted adversarial examples of size $\varepsilon_\infty = 8/255$ on two IF approaches with complementary goals. Upper: a model seeking to match an original image, invariant to any editorial change in the query [6]. Lower: a model seeking to *avoid* matching an edited query to the original [45]. In both cases it is possible to attack the model to defeat the goal, and in both cases ARIA defends it successfully.

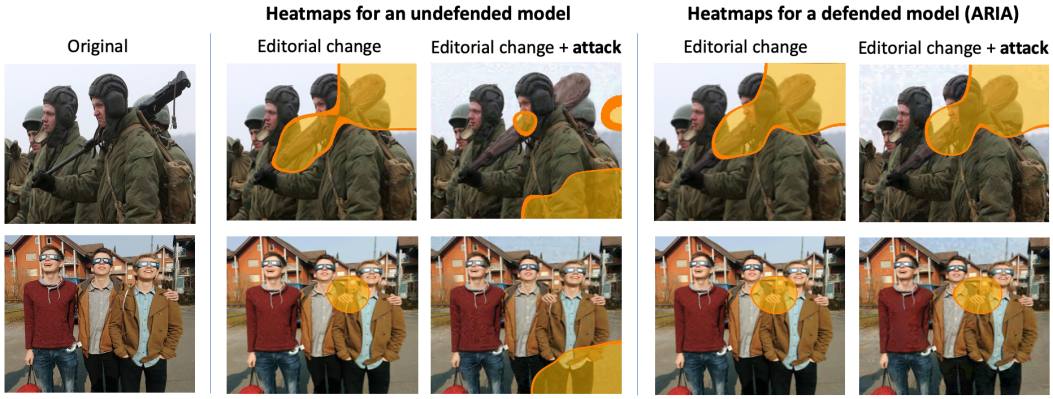

Figure 3. Attack on the image comparator (IC) model of Black et al. [6]. Visualization of the heatmap generated for an undefended and ARIA defended models queried using an adversarial example of budget $\varepsilon_\infty = 8/255$. The attack targets a heatmap prediction within the bottom-right quadrant of the image. Shown for two different images drawn from the PSBattles dataset. In all cases the ARIA defended model predicts a heatmap near-identical to the original heatmap inferred by the undefended model in the absence of the attack.

and the IF of the original x:

$$\begin{aligned} \max \quad & \|\phi(x_{query}+\delta)-\phi(x)\|_2^2 \\ \text{s.t.} \quad & \|\delta\|_\infty \le \varepsilon, \quad 0 \le x_{query}+\delta \le 1. \end{aligned} \tag{1}$$

In Fig. 2 (upper), we show results of such an untargeted attack on Black et al. [6]. The method attempts to match query images exhibiting both editorial and non-editorial change, back to an original. The attack successfully defeats this goal.

By contrast, in Fig. 2 (lower) the model of Nguyen et al. [45] aims to avoid matching an edited image to an original, in order to avoid corroborating a manipulated image with the provenance of its original. We perform a *targeted* attack which attempts to match a query image x_{query} to the original image x:

$$\begin{aligned} \min \quad & \|\phi(x_{query}+\delta)-\phi(x)\|_2^2 \\ \text{s.t.} \quad & \|\delta\|_\infty \le \varepsilon, \quad 0 \le x_{query}+\delta \le 1. \end{aligned} \tag{2}$$

We note here that when producing adversarial examples for the OSCAR-Net of [45], we assume the object detector's output to be fixed (see the details in the sup. mat).

Image comparator attack. To attack the prediction module of an IC model [6] (Fig. 3), we minimize the probability p_C of the ground truth class y (e.g., 'same images with editorial changes') over the perturbation δ added only to x_{query}:

$$\begin{aligned} \min \quad & \log p_C(x_{top-1}, x_{query}+\delta)_y \\ \text{s.t.} \quad & \|\delta\|_\infty \le \varepsilon, \quad 0 \le x_{query}+\delta \le 1. \end{aligned} \tag{3}$$

For the heatmap attack, we minimize the cosine similarity between the predicted heatmap $f_T(x)$ and ground truth $t \in [0,1]^{7\times 7}$ since this is the loss used for training in [6]:

$$\begin{aligned} \min \quad & \cos(f_T(x_{query}+\delta), t) \\ \text{s.t.} \quad & \|\delta\|_\infty \le \varepsilon, \quad 0 \le x_{query}+\delta \le 1. \end{aligned} \tag{4}$$

Moreover, targeted attacks on heatmaps may trick a user into distrusting an *original* image. To generate targeted adversarial examples, we can just flip the minimization to maximization and use some target label or heatmap.

4. Adversarially Robust Image Attribution

In this section, we propose a robust contrastive learning algorithm applicable to both fingerprinting approaches described in Nguyen et al. [45] and Black et al. [6], including the image comparator of the latter.

4.1. Robust contrastive learning

We propose a method that adapts contrastive learning with the SimCLR loss [12] to be robust to imperceptible adversarial examples. Denote by $L(\{x_i\}_{i=1}^{2N})$ the SimCLR loss defined on a batch of paired *positive* examples, where i-th and $(N+i)$-th examples correspond to the same images but with different *transformations*:

$$L(\{x_i\}_{i=1}^{2N}) = \frac{1}{2N}\sum_{i=1}^{N}[\ell(\{x_i\}_{i=1}^{2N})_{i,N+i} + \ell(\{x_i\}_{i=1}^{2N})_{N+i,i}],$$

$$\ell(\{x_i\}_{i=1}^{2N})_{i,j} = -\log \frac{e^{\cos(\phi(x_i),\phi(x_j))/\tau}}{\sum_{k=1}^{2N} \mathbb{1}_{k\neq i} e^{\cos(\phi(x_i),\phi(x_k))/\tau}} \quad (5)$$

These transformations are of two kinds: non-editorial $\bar{E}(x)$ (e.g., affine transformations and ImageNet-C [29]) and editorial $E(x)$ (e.g., available as paired images in the PS-Battles dataset [28]). Black et al. [6] use both of them as *positive* examples while Nguyen et al. [45] only treats images with non-editorial changes as positive.

Then to train adversarially robust IF models, we change the objective following the robust optimization framework [41] by adding an inner loop to maximize the loss on adversarially perturbed images $x_i + \delta_i$:

$$\min_{\theta\in\mathbb{R}^{|\theta|}} \mathbb{E}_{\{x_i\}_{i=1}^{2N}\sim D}\left[\max_{\substack{\|\delta_i\|_\infty\leq\varepsilon \\ 0\leq x_i+\delta_i\leq 1}} L(\{x_i+\delta_i\}_{i=1}^{2N})\right], \quad (6)$$

where θ denotes the model parameters and D the data distribution. We note that although adversarial training [41] is an established technique for image classification, IF models are trained differently. Thus, it is not clear in advance if the findings and pitfalls of robust training of image classification models (e.g., such as catastrophic or robust overfitting [48, 61]) transfer to the IF setting.

To solve the inner maximization problem, we use a few iterations of projected gradient ascent (in practice, up to 3) for the inner maximization problem, where each iteration requires an evaluation of the input gradient $\nabla_{\delta_i} L(\{x_i + \delta_i\}_{i=1}^{2N})$ via backpropagation. Using a few iterations of the attack comes out to be sufficient to prevent the *catastrophic overfitting* problem [2, 61] which also manifests itself in training IF models, as we observe in the experimental part.

The final objective that we use combines the robust version of the SimCLR loss [12] with the hashing term from [45] for large-scale search has the following form:

$$\min_{\theta\in\mathbb{R}^{|\theta|}} \mathbb{E}_{\{x_i\}_{i=1}^{2N}\sim D}\left[\max_{\substack{\|\delta_i\|_\infty\leq\varepsilon \\ 0\leq x_i+\delta_i\leq 1}} L(\{x_i+\delta_i\}_{i=1}^{2N})\right] + \quad (7)$$
$$\alpha\, \mathbb{E}_{x\sim D}\left[\|\phi(x) - \operatorname{sign}(\phi(x))\|^3\right].$$

For [6], which proposes no end-to-end hashing, we mostly report the models trained without the hashing term. In practice, we approximate the expectations using mini-batches, and we apply the hashing term on the same examples as the main loss. We do not use projection layers on top of the target embeddings as in Chen et al. [12] since we found this leads to worse performance.

4.2. Robust image comparator network

Next we discuss how to make the *image comparator model* from Black et al. [6] robust to adversarial attacks. First, we note that the image comparator performs a classification task (both for the prediction and heatmap modules) for which there are well-described solutions in the literature on how to improve their robustness [22, 41, 64]. Thus, we use the multi-task objective of Black et al. [6] with the classification loss $w_c\mathcal{L}_C$ and heatmap loss $w_t\mathcal{L}_T$ (we use the same losses and weights as in Black et al. [6], i.e. the cross-entropy and cosine similarity and $w_c = w_t = 0.5$), but we add an inner maximization operator to it:

$$\min_{\theta\in\mathbb{R}^{|\theta|}} \mathbb{E}_{x_1,x_2,y,t}\left[\max_{\substack{\|\delta\|_\infty\leq\varepsilon \\ 0\leq x_2+\delta\leq 1}} w_c\mathcal{L}_C(x_1, x_2+\delta, y) + w_t\mathcal{L}_T(x_1, x_2+\delta, t)\right], \quad (8)$$

where y is the class label, t is the ground-truth heatmap. Note that we add adversarial perturbations only to the corrupted query image x_2 as we assume that the image comparator operates on the images x_1 from the database which contains non-adversarial original images. As corrupted images, we use either original images under a non-editorial distortion (see the experimental setup below), manipulated images from PSBattles or simply a different image under a non-editorial distortion. We use SGD for the outer loop and a few steps of PGD to approximate the maximization. We note that the image comparator model is fully differentiable, including the geometric alignment RAFT module [54] which is a part of their model. Thus, we both train *and* generate adversarial examples completely end-to-end for this network.

5. Experimental evaluation

In this section, we provide a detailed description of the experimental setup, describe the main results and show ablation studies that give more insights in the proposed method.

5.1. Experimental setup

Performance metrics. For the image retrieval model of Black et al. [6], we similarly consider recall-based metrics assuming that for all queried images, the originals have been indexed: (1) *standard recall*: $\Pr[f(x_{query}) = f(x)]$ which is the probability of retrieving a correct image under some image corruption (non-editorial transformation, manipulation or both), (2) *adversarial recall*: $\Pr[f(x_{query}) + \delta = f(x)]$ which is the probability of retrieving a correct image under non-editorial transformations and an adversarial perturbation

δ. For the image comparator model [6], we use the same metrics as in their paper: *average precision* (AP) for the classification module and the *intersection over union* (IoU) over the images with editorial changes for the heatmap module. We use all images for the former and only images with editorial manipulations for the latter.

We note that unlike [6], OSCAR-Net [45] is trained to distinguish non-editorial transformations from editorial manipulations. Thus, we follow their metrics using standard mAP and top-1 recall (R@1) for the non-editorially transformed query set, also inverse mAP (imAP) and inverse recall (iR@1) for the tamper query set. For all metrics, the higher is better.

Training details. For training the robust image retrieval models we use the ResNet-50 architecture and Behance1M; a subset of 1M images sourced from a public digital art portfolio website (`behance.net`).[2] We use Behance1M for training since it is significantly larger than PSBattles [28] and more diverse, containing not only photos but also graphics. As a starting set of parameters, we use the model from Black et al. [6] pre-trained for 2 epochs with standard contrastive training. We train the model to be robust not only to adversarial examples (after two epochs of standard training) but also to *non-editorial transforms* via data augmentation, for which we use the `beacon_aug` library [40] to apply ImageNet-C corruptions [29] then resizing, rotation, padding, cropping, horizontal flips, and JPEG compression.

Evaluation details. We evaluate robustness on Behance and on PSBattles [28]; we use the 'hard' subset of the latter defined in Black et al. [6]. As in past work [6, 45], we add distractor images from stock photography (Adobe Stock thumbnails). We use the FAISS library [36] for efficient image retrieval. We generate adversarial attacks using ℓ_∞ PGD with 50 iterations using $\varepsilon = {}^8\!/_{255}$, step size ${}^4\!/_{255}$ decayed by a factor of 2 at 25%, 50%, and 75% of iterations. Unless specified otherwise, we evaluate the image retrieval models on full PSBattles with 2M and 100K Adobe Stock distractors, following the settings in [6] and [45], respectively. We apply adversarial attacks in the original pixel space with differentiable resizing to 224×224 and JPEG-90 compression after it (thus, it is a *fully realizable* attack).

5.2. Robust retrieval: Black et al. [6] approach

Large-scale robustness evaluation. The main evaluation results for the robust image retrieval models trained on Behance1M are presented in Table 1 where we measure recall with 2M distractor images using the IVF1024, PQ16 index from FAISS library following Black et al. [6] (we include non-FAISS results also in the sup.mat., omitted as too slow to be practical). We report models trained with different ε_∞ since we want to show models with a different *robustness-accuracy tradeoff* [57]. We measure adversarial recall under perturbations of size $\varepsilon_\infty = {}^8\!/_{255}$ since this is the most commonly used perturbation size in the literature [14, 41]. Compared to the existing IF models trained contrastively on PSBattles [6] and in a supervised way on ImageNet [32, 47, 50], our proposed method allows us to train IF models which are robust to imperceptible adversarial perturbations *in addition* to being highly accurate. For example, our robust model trained with $\varepsilon_\infty = {}^4\!/_{255}$ achieves 91.6% standard and 85.1% adversarial recall compared to 80.1% and 0.0% achieved by the main baseline from [6]. In addition, Table 1 shows that robust contrastive learning with $\varepsilon = {}^4\!/_{255}$ leads to a higher recall for manipulated images compared to the standard model even despite having a worse recall on non-editorial distortions. We note that this shows another interesting case when adversarial training can improve the performance on *real-world distribution shifts* in addition to the benefits of adversarial training known before on, e.g., common corruptions [19, 38, 62] or transfer learning [50, 58]. Finally, the models trained with the hashing term (see Eq. (7)) perform slightly worse than the models producing real-valued IFs, but there can still be interesting use-cases for them such as faster search, lower memory requirements, and a potential storage in key-value data structures.

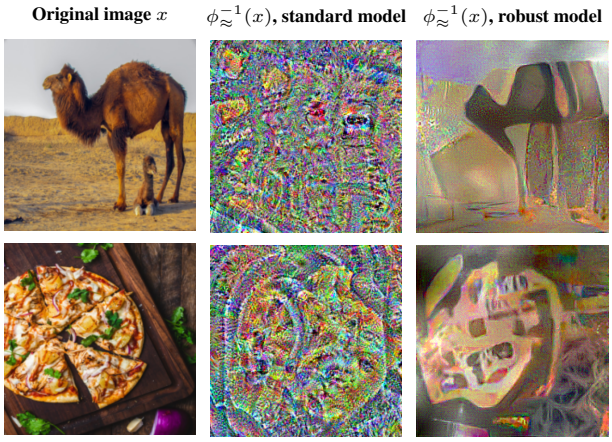

Figure 4. Visualization of the hash inversions $\phi_{\approx}^{-1}(x)$ (see Eq. (9)) for two original images x (**left**) for a standard model (**middle**) and ARIA model with $\varepsilon_\infty = {}^4\!/_{255}$ (**right**), both trained on Behance1M.

Robustness to unseen adversarial perturbation. In Table 2, we show the robustness results for perturbations which were unseen during training such as ℓ_2-bounded perturbations ($\varepsilon_2 = 0.5$) and ℓ_∞-perturbations of a larger radius compared to those used for training ($\varepsilon_\infty \in \{{}^{12}\!/_{255}, {}^{16}\!/_{255}, {}^{32}\!/_{255}\}$). We can see that robustness indeed generalizes to these other types of perturbations: e.g., $\varepsilon_2 = 0.5$ is sufficient to reduce the adversarial recall to 0.0% for the undefended model of Black et al. [6] while all our ℓ_∞-trained robust models achieve 83%+ adversarial recall.

Plausible hash inversions. Here we show that adversarially robust image *hashing* models output plausible images under the *hash inversions attacks*. This type of attacks has recently received a lot of attention as a significant weakness of neural hashing models [51]. The goal of this attack is to compromise the validity of the hashing model by finding some *irrelevant* images that lead to the *same* binary hash as some target image x. The formulation is similar to that of targeted adversarial attacks but without the ℓ_∞-norm constraint on

[2] We will publish the URL list of these images upon acceptance.

Existing models	Non-editorial distortions: No attack R@1	Non-editorial distortions: No attack R@100	Non-editorial distortions: ℓ_∞ adversarial R@1	Non-editorial distortions: ℓ_∞ adversarial R@100	Editorial manipulations: No attack R@1	Editorial manipulations: No attack R@100	Editorial manipulations: ℓ_∞ adversarial R@1	Editorial manipulations: ℓ_∞ adversarial R@100	Editorial + non-editorial: No attack R@1	Editorial + non-editorial: No attack R@100	Editorial + non-editorial: ℓ_∞ adversarial R@1	Editorial + non-editorial: ℓ_∞ adversarial R@100
	Top-1 and top-100 recall for different query sets											
Standard supervised, ImageNet [47]	20.8	39.6	0.0	0.1	87.2	**95.4**	0.0	0.2	15.1	33.2	0.0	0.1
DeepAugment + AugMix supervised, ImageNet [32]	47.9	66.9	0.1	0.3	86.7	95.3	0.0	0.1	36.5	58.0	0.0	0.3
Robust supervised, $\varepsilon_\infty = 4/255$, ImageNet [50]	39.4	53.0	10.2	20.8	86.3	95.0	31.2	56.1	30.8	46.5	5.5	15.2
Undefended contrastive, PSBattles [6]	74.1	93.7	0.0	0.0	80.1	92.9	0.0	0.0	55.3	83.5	0.0	0.0
Our new models												
Undefended contrastive, Behance1M (ours)	**97.8**	**98.8**	1.3	12.4	88.6	92.4	0.4	4.5	84.7	89.7	1.0	9.5
ARIA contrastive + hashing, $\varepsilon_\infty = 4/255$, Behance1M	93.5	96.2	74.7	81.9	87.5	92.8	76.4	86.5	79.2	87.6	56.3	70.7
ARIA contrastive + hashing, $\varepsilon_\infty = 8/255$, Behance1M	89.0	93.5	80.7	83.4	87.0	92.5	80.6	87.9	74.8	84.7	57.9	72.5
ARIA contrastive, $\varepsilon_\infty = 2/255$, Behance1M	97.3	98.2	79.0	82.1	90.8	93.6	81.0	86.4	**87.3**	**91.0**	63.0	71.0
ARIA contrastive, $\varepsilon_\infty = 4/255$, Behance1M	96.4	97.3	83.0	85.7	**91.6**	93.9	85.1	**89.9**	86.7	90.3	**69.7**	**77.0**
ARIA contrastive, $\varepsilon_\infty = 8/255$, Behance1M	94.2	96.0	**83.7**	**87.4**	90.4	93.3	**85.5**	**89.9**	83.1	88.0	69.3	**77.0**

Table 1. Standard and ℓ_∞ adversarial ($\varepsilon_\infty = 8/255$) top-1 and top-100 recall for different ResNet-50 models evaluated on PSBattles [28]. The database contains original images from PSBattles and 2M distractor images from Stock indexed using the IVF1024, PQ16 index from FAISS library following Black et al. [6]. We use three query sets based on PSBattles: (1) non-editorial distortions (ImageNet-C and affine) on original images, (2) editorial manipulations but no distortions, (3) editorial manipulations with non-editorial distortions.

Models	Recall	Adversarial recall ε_2 5.0	Adversarial recall ε_∞ 16/255	Adversarial recall ε_∞ 32/255
Undefended, PSBattles [6]	92.2%	0.0%	0.0%	0.0%
ARIA, $\varepsilon_\infty = 2/255$, Behance1M	**99.6%**	**84.2%**	38.8%	2.8%
ARIA, $\varepsilon_\infty = 4/255$, Behance1M	98.4%	84.0%	50.6%	9.8%
ARIA, $\varepsilon_\infty = 8/255$, Behance1M	97.8%	83.4%	**60.0%**	**15.4%**

Table 2. Standard and adversarial top-1 recall for attacks *unseen* during training. We use evaluation on a query set of non-editorial distortions of Behance1M images (500 distractors).

the perturbation magnitude and starting from some arbitrary $\hat{x}$ (we use a constant gray image). We take the robust model from Table 1 trained to output binary hashes via the sign function for which we use a differentiable approximation ($\texttt{tanh}$ function with a parameter β):

$$\phi_{\approx}^{-1}(x) = \underset{0\leq\hat{x}\leq 1}{\arg\min} \|\tanh(\beta \cdot \phi(\hat{x})) - \text{sign}(\phi(x))\|_2^2 . \quad (9)$$

We solve the formulation using 1'000 iterations of PGD ensuring the exact hash match in each case. We note that similar formulations in the context of image classification have been studied in [16, 17, 42] from the interpretability perspective where they focused on inverting real-valued embeddings of a deep network instead of hashes.

As we can see from Fig. 4, hash inversion attacks on standardly trained hashing models tend to produce obscure high-frequency patterns. At the same time, our robust image hashing models tend to focus more on shapes of objects which are approximately recovered under hash inversions. This behaviour is closely related to the adversarial vulnerability problem: the attacker can use non-robust features [34] to arbitrarily manipulate the model's hash. However, once we fix this problem via robust training, hash inversions start to be more related to the original images.

Hyperparameter importance. Here we analyze the hyperparameters of ARIA training: the number of PGD iterations and the perturbation radius ε_∞. We train multiple models on Behance1M and report their baseline (*i.e.* no attack) and adversarial recall (*i.e.* for an attack with budget $\varepsilon_\infty = 8/255$). Fig. 5 (top) suggests that, similar to image classification, catastrophic overfitting [2, 61] leads to unstable performance for robust contrastive learning, so training with a single iteration of PGD should be avoided. We observe 2-3 iterations of PGD to be sufficient: which leads to the slowdown factor of 1.9× and 2.3×, respectively. Fig. 5 (bottom) shows that the model trained with $\varepsilon = 8/255$ starts to *overfit* in terms of adversarial recall after 7 epochs. This suggests that training with a too large ε_∞ may be problematic.

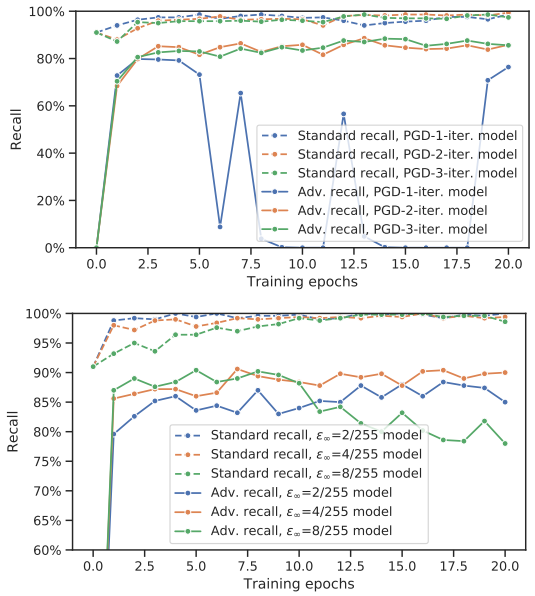

Figure 5. Standard and ℓ_∞ adversarial ($\varepsilon_\infty = 8/255$) top-1 recall at different epochs for models trained with (1) different numbers of PGD iterations (using $\varepsilon_\infty = 4/255$) and (2) different perturbation radii ε_∞ (using 3 iterations of PGD) . We evaluate a query set of non-editorial distortions of Behance1M images (500 distractors).

5.3. Robust retrieval: Nguyen et al. [45] approach

We present our attack and defence results on OSCAR-Net [45] in Table 3. Following the evaluation protocol of the attribution benchmark in [45], we report mAP/R@1 scores for

Models	No attack						ℓ_∞ adversarial			
	mAP	**R@1**	**imAP**	**iR@1**	$\mathbf{F_{mAP}}$	$\mathbf{F_{R@1}}$	**imAP**	**iR@1**	$\mathbf{F_{mAP}}$	$\mathbf{F_{R@1}}$
Undefended [45]	**78.66**	**66.35**	72.83	81.05	**37.82**	**36.48**	15.55	22.87	12.98	17.01
ARIA, $\varepsilon_\infty = 2/255$ (ours)	54.52	39.63	78.64	85.00	32.20	27.03	34.05	42.84	20.96	20.59
ARIA, $\varepsilon_\infty = 4/255$ (ours)	55.86	43.38	**79.62**	85.81	32.83	28.81	35.79	45.17	21.81	22.13
ARIA, $\varepsilon_\infty = 8/255$ (ours)	52.18	38.11	78.89	**85.88**	31.41	26.40	**55.92**	**66.28**	**26.99**	**24.20**

Table 3. Metrics for no attack and ℓ_∞ adversarial ($\varepsilon_\infty = 8/255$) attack for OSCAR-Net models, using queries from PSBattles. For mAP and R@1, queries have only non-editorial transforms applied. For iMAP and iR@1 digitally manipulated images with no distortions are used, both with and without adversarial perturbations. F-scores are calculated based on the appropriate mAP/R@1 following Nguyen et al. [45].

Models	No attack		ℓ_∞ adversarial	
	AP	**IoU**	**AP**	**IoU**
Undefended ICN [6]	**96.4%**	58.1%	0.6%	5.1%
ARIA ICN, $\varepsilon_\infty = 2/255$	**96.4%**	**61.5%**	65.0%	37.9%
ARIA ICN, $\varepsilon_\infty = 4/255$	95.9%	59.3%	83.1%	43.7%
ARIA ICN, $\varepsilon_\infty = 8/255$	95.5%	55.9%	**90.7%**	**44.9%**

Table 4. The average precision (**AP**) and intersection over union (**IoU**) between the predicted and ground truth editorial heatmaps for the **image comparator network** (ICN) with/without adversarial perturbations of radius $\varepsilon_\infty = 8/255$.

the non-editorially transformed query set and imAP/iR@1 scores for the editorially transformed query set, together with the harmonic F score balancing the two terms (F_{mAP} and $F_{R@1}$). Additionally, we perform adversarial attacks on the editorial query set using $\varepsilon_\infty = 8/255$, creating a new query set called ℓ_∞ **adversarial**. We trained OSCAR-Net models with $\varepsilon_\infty \in \{2/255, 4/255, 8/255\}$ to defend against such attacks, using Eq. (2) and report its performance on both the no attack and ℓ_∞ **adversarial** attack scenarios.

Whilst baseline OSCAR-Net works well, it performs poorly on ℓ_∞ **adversarial**, with 57% drop in imAP and iR@1 scores; the baseline model is easily fooled by the attack. In contrast, all 3 defence models outperform OSCAR-Net by significant margins (3× improvement on imAP and iR@1 for the best model). These models also perform better at imAP and iR@1 scores on the standard benchmark, at the cost of performance reduction on the non-editorial set. We note this trade-off is already observed in the original OSCAR-Net [45] and is associated with feature generalization versus discrimination. The trade-off is steered towards boosting the discrimination of editorial changes because the models are trained to defend against adversarial attacks on such changes. The value of ε_∞ can be used to determine the defence strength, with $\varepsilon_\infty = 8/255$ yielding the best performance, closest to performance on the standard benchmark. This is consistent with subsec. 5.2 when defending [6].

5.4. Robust image comparator

We fine-tune the model from Black et al. [6] for 40 epochs using 3 iterations of PGD attack for training using different ℓ_∞ radii ($\varepsilon_\infty \in \{2/255, 4/255, 8/255\}$) and show the results in Table 4. We benchmark robust image comparator models separately since they are trained independently of image fingerprinting models and provide complementary information about the presence of an editorial change and its location.

Classification module. First, we observe that our robust training method substantially improves the classification precision under untargeted adversarial examples. E.g., for the model trained with $\varepsilon_\infty = 2/255$, we preserve the same precision as the model from [6] (96.4%) but substantially improve the adversarial precision (from 0.6% to 65.0%). The most robust model is the one trained with $\varepsilon_\infty = 8/255$ which achieves 90.7% adversarial precision which, however, sacrifices 0.9% of precision. We benchmark targeted attacks and show confusion matrix over classes in the sup. mat.

Heatmap module. Quantitatively, we improve the heatmap performance in terms of the adversarial IoU significantly: from 5.1% to 44.9% for the most robust model (trained with $\varepsilon_\infty = 8/255$) which has a comparable IoU: 55.9% vs. 58.1%. However, if we consider the robust model trained with $\varepsilon_\infty = 4/255$, it has both better baseline IoU (59.3% vs. 58.1%) and adversarial IoU (43.7% vs. 5.1%). Qualitative results for a robust heatmap module are visualized in Fig. 3: a robust comparator correctly highlights the same edited area regardless of whether adversarial perturbations are added. At the same time, the comparator from [6] can be fooled also in a targeted way to highlight *any* area (e.g., right bottom corner) as manipulated.

6. Conclusions

We began by showing that the current state-of-the-art image attribution models (both image fingerprinting and comparison), are not robust to imperceptible adversarial attacks. This is concerning since these attacks are *fully realizable* and can be applied directly in a digital format where the attacker has *white-box* access to the model. To bridge this vulnerability, we proposed a simple and effective training technique for image attribution that significantly improves robustness to various adversarial perturbations including the ones which were unseen during training. We applied this to two fingerprinting approaches [6, 45]: those seeking to match manipulated content and those seeking to avoid matching such content. Finally, we also showed how a recent manipulation localization approach [6] can be trained robustly including both its classification and heatmap modules.

Overall, we think that adversarial vulnerability of image attribution models presents a significant negative societal impact, and that our proposed method provides a *well-motivated* and *practical* way to solve it. Our solution, ARIA, is particularly timely given the emergence of content authenticity standards that advocate for use of image attribution [1]. ARIA has two limitations: the increased training time (however, only $\approx 2\times$) and the trade-off for some models between a significant improvement in robustness and minor loss of accuracy.

References

[1] The coalition for content provenance and authenticity (C2PA) technical specification v1.0. Technical report, C2PA, 2022. URL https://c2pa.org/specifications/specifications/1.0/specs/C2PA_Specification.html.

[2] Maksym Andriushchenko and Nicolas Flammarion. Understanding and improving fast adversarial training. In *Proc. NeurIPS*, 2020.

[3] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *Proc. ICML*, pages 284–293. PMLR, 2018.

[4] Jiawang Bai, Bin Chen, Yiming Li, Dongxian Wu, Weiwei Guo, Shu-tao Xia, and En-hui Yang. Targeted attack for deep hashing based retrieval. In *Proc. ECCV*, pages 618–634. Springer, 2020.

[5] Alexander Black, Tu Bui, Simon Jenni, Vishy Swaminathan, and John Collomosse. VPN: Video provenance network for robust content attribution. In *Proc. CVMP*, 2021.

[6] Alexander Black, Tu Bui, Hailin Jin, Vishy Swaminathan, and John Collomosse. Deep image comparator: Learning to visualize editorial change. In *Proc. Workshop on Media Forensics (WMF) at CVPR*, pages 972–980, June 2021.

[7] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

[8] Anh Bui, Trung Le, He Zhao, Paul Montague, Seyit Camtepe, and Dinh Phung. Understanding and achieving efficient robustness with adversarial supervised contrastive learning. *arXiv preprint arXiv:2101.10027*, 2021.

[9] Pin-Chun Chen, Bo-Han Kung, and Jun-Cheng Chen. Class-aware robust adversarial training for object detection. In *Proc. CVPR*, pages 10420–10429, 2021.

[10] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Chau. ShapeShifter: Robust physical adversarial attack on faster r-CNN object detector. In *ML and Knowledge Discovery in Databases*, pages 52–68. Springer International Publishing, 2019.

[11] Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *Proc. CVPR*, 2020.

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. ICML*, pages 1597–1607. PMLR, 2020.

[13] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.

[14] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Proc. NeurIPS: Datasets and Benchmarks track*, 2021.

[15] Brian Dolhansky and Cristian Canton Ferrer. Adversarial collision attacks on image hashing functions. *Proc. CVPR WS*, 2021.

[16] Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In *Proc. CVPR*, pages 4829–4837, 2016.

[17] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations. *arXiv preprint arXiv:1906.00945*, 2019.

[18] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Dawn Song, Tadayoshi Kohno, Amir Rahmati, Atul Prakash, and Florian Tramer. Note on attacking object detectors with adversarial stickers. *arXiv preprint arXiv:1712.08062*, 2017.

[19] Nicolas Ford, Justin Gilmer, Nicolas Carlini, and Dogus Cubuk. Adversarial examples are a natural consequence of test error in noise. In *ICML*, 2019.

[20] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[21] Sven Gowal, Jonathan Uesato, Chongli Qin, Po-Sen Huang, Timothy Mann, and Pushmeet Kohli. An alternative surrogate loss for pgd-based adversarial testing. *arXiv preprint arXiv:1910.09338*, 2019.

[22] Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020.

[23] Sven Gowal, Po-Sen Huang, Aaron van den Oord, Timothy Mann, and Pushmeet Kohli. Self-supervised adversarial robustness for the low-label, high-data regime. In *Proc. ICLR*, 2021.

[24] Sam Gregory. The adobe content authenticity initiative approach to authenticity infrastructure against media manipulation, 2020. URL https://blog.witness.org/2020/08/adobe-content-authenticity-initiative-approach-authenticity-infrastructure-media-manipulation/.

[25] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

[26] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proc. ICCV*, pages 2961–2969, 2017.

[27] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B Girshick. Momentum contrast for unsupervised visual representation learning. corr abs/1911.05722 (2019). *Proc. CVPR*, 2020.

[28] Silvan Heller, Luca Rossetto, and Heiko Schuldt. The ps-battles dataset-an image collection for image manipulation detection. *arXiv preprint arXiv:1804.04866*, 2018.

[29] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proc. ICLR*, 2019.

[30] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *Proc. ICML*, 2019.

[31] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Proc. NeurIPS*, 2019.

[32] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021.

[33] Chih-Hui Ho and Nuno Vasconcelos. Contrastive learning with adversarial examples. *arXiv preprint arXiv:2010.12050*, 2020.

[34] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Proc. NeurIPS*, 2019.

[35] Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning. In *Proc. NeurIPS*, 2020.

[36] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017.

[37] Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning. *Proc. NeurIPS*, 2020.

[38] Klim Kireev, Maksym Andriushchenko, and Nicolas Flammarion. On the effectiveness of adversarial training against common corruptions. *arXiv preprint arXiv: arXiv:2103.02325*, 2021.

[39] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[40] Xiaoyang Rebecca Li, Yannick Hold-Geoffroy, Oxholm Geoffrey, Krishna Kumar Singh, Zhifei Zhang Zhang, Richard Zhang, Maksym Andriushchenko, et al. Beacon-aug: A cross-library image augmentation toolbox. https://github.com/adobe-research/beacon-aug, 2021. Online; accessed Sep-22-2021.

[41] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[42] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015.

[43] Chengzhi Mao, Ziyuan Zhong, Junfeng Yang, Carl Vondrick, and Baishakhi Ray. Metric learning for adversarial robustness. *NeurIPS*, 2019.

[44] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proc. CVPR*, pages 2574–2582, 2016.

[45] Eric Nguyen, Tu Bui, Vishy Swaminathan, and John Collomosse. Oscar-net: Object-centric scene graph attention for image attribution. In *Proc. ICCV*, pages 14499–14508, 2021.

[46] Thomas Kobber Panum, Zi Wang, Pengyu Kan, Earlence Fernandes, and Somesh Jha. Exploring adversarial robustness of deep metric learning. *arXiv preprint arXiv:2102.07265*, 2021.

[47] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. *Technical Report*, 2017.

[48] Leslie Rice, Eric Wong, and J Zico Kolter. Overfitting in adversarially robust deep learning. In *ICML*, 2020.

[49] Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? *arXiv preprint arXiv:2106.11230*, 2021.

[50] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? *Proc. NeurIPS*, 2020.

[51] Lukas Struppek, Dominik Hintersdorf, Daniel Neiderand, and Kristian Kersting. Learning to break deep perceptual hashing: The use case neuralhash. *arXiv preprint arXiv:2111.06628*, 2021.

[52] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Dumitru Erhan Joan Bruna, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proc. ICLR*, 2013.

[53] Alex Tamkin, Mike Wu, and Noah Goodman. Viewmaker networks: Learning views for unsupervised representation learning. In *Proc. ICLR*, 2021.

[54] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proc. ECCV*, pages 402–419. Springer, 2020.

[55] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. *arXiv preprint arXiv:1904.08653*, 2019.

[56] Giorgos Tolias, Filip Radenovic, and Ondrej Chum. Targeted mismatch adversarial attack: Query with a flower to retrieve the tower. In *Proc. ICCV*, pages 5037–5046, 2019.

[57] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *Proc. ICLR*, 2019.

[58] Francisco Utrera, Evan Kravitz, N Benjamin Erichson, Rajiv Khanna, and Michael W Mahoney. Adversarially-trained deep nets transfer better. *arXiv preprint arXiv:2007.05869*, 2020.

[59] Sheng-Yu Wang, Oliver Wang, Andrew Owens, Richard Zhang, and Alexei A Efros. Detecting photoshopped faces by scripting photoshop. In *Proc. ICCV*, 2019.

[60] Zhi Wang, Yiwen Guo, and Wangmeng Zuo. Deepfake forensics via an adversarial game. *arXiv preprint arXiv:2103.13567*, 2021.

[61] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. *ICLR*, 2020.

[62] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proc. CVPR*, 2020.

[63] Cong Xu and Min Yang. Adversarial momentum-contrastive pre-training. *arXiv preprint arXiv:2012.13154*, 2020.

[64] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proc. ICML*, 2019.

[65] Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. *Proc. ICML*, 2020.

Appendix

Organization of the appendix

In Sec. A we present additional training and evaluation details. In Sec. B, we provide further implementation details for the attacks and defences both for OSCAR-Net and Black et al. [6] models. In Sec. C, we show multiple additional experiments such as accuracy of the retrieval with exact nearest neighbour search, additional hash inversion visualizations, robustness of OSCAR-Net to unseen adversarial perturbations, accuracy over classes and targeted attacks on the heatmaps for ICN models.

A. Training and evaluation details

Training details. For the models trained according to the approach of Black et al. [6], we use the learning rate 0.01, SimCLR temperature 0.1, 3 steps of PGD for training using step sizes $\{1/255, 2/255, 4/255\}$ for $\varepsilon_\infty \in \{2/255, 4/255, 8/255\}$, respectively.

For the OSCAR-Net [45] models, we use the default hyperparameters except the learning rate which is set to $1e-6$ and SimCLR temperature of 0.8. For ARIA training, we use 3 steps of PGD with the step size $0.5\,\varepsilon_\infty$.

For the image comparator models, we use the default training hyperparameters with 3 steps of PGD for training using step sizes $\{1/255, 2/255, 4/255\}$ for $\varepsilon_\infty \in \{2/255, 4/255, 8/255\}$, respectively.

Evaluation details. For the attacks unseen during training, we use 200 iterations of PGD (we increase it from 50 iterations used throughout the paper to account for larger perturbation radii) using the step size of $\varepsilon_\infty = 4/255$ for ℓ_∞-perturbations and $\varepsilon_2 = 0.5$ for ℓ_2-perturbations.

For hash inversions, we use 1000 iterations of PGD with the step size $4/255$, and the approximation parameter $\beta = 1$.

Training time. Standard training of the Black et al. [6] model on Behance1M takes 34.3 hours while ARIA training takes 72.8 hours (i.e., $2.3\times$ factor slowdown) on two NVIDIA V100 GPUs for 20 epochs.

Standard OSCAR-Net training on PSBattles takes 31.6 hours while ARIA training takes 65.1 hours (i.e., $2.1\times$ factor slowdown) on a single NVIDIA GeForce RTX 3090 GPU for 10 epochs.

We note that for both models, ARIA uses 3 steps of PGD for training but the slowdown factor is less than $4\times$ which is due to more effective GPU utilization for robust training.

Examples of non-editorial transformations. In Fig. 6 and Fig. 7, we show images with non-editorial changes from PSBattles which we used for the "*Editorial + non-editorial*" query sets for evaluation of the OSCAR-Net models and models of Black et al. [6].

B. Further details on the attack and defence scope on OSCAR-Net and Black *et al*. models

A model needs to be differentiable with respect to the input image in order to perform an effective adversarial attack (and defence) on it. In other words, our main prerequisite is that we should be able to back-propagate the gradient of the loss to the original input. Despite being complex attribution models, we show that OSCAR-Net [45] and Black *et al*. [6] both can meet this requirement.

OSCAR-Net consists of an object detection module (Mask-RCNN [26]) to decompose an image into a set of objects, followed by 3 sub-networks to learn the global image features, object-level features (including object CNN visual, shape and geometry features) as well as the relation features between objects. These features are pooled via a fully-connected graph transformer network to produce a compact binary embedding. Note that OSCAR-Net does not aim to learn object detection (the Mask-RCNN module weights are not updated during training), and we do the same. Here we focus on attacking and defending the multi-branch feature extraction and aggregation which are learnable in OSCAR-Net. Thus, we apply our perturbations to the full image after the object detection step, i.e. we treat the output of the object detector as constant. We note that there exists adversarial attack and defence approaches on object detection [9] and integrating those on OSCAR-Net could be a topic of future work.

Black *et al*. consists of two distinct models that are trained separately: an image retrieval model insensitive to both editorial and non-editorial changes, followed by an image comparator (IC) model distinguishing editorial from non-editorial transformations. Given a query, the image retrieval model returns top-k candidate images which are brought to the IC model to determine if there exists a 'matched' image among the candidates and whether the query has editorial or non-editorial changes. The IC model also outputs an editorial heatmap if editorial change is predicted on a query-candidate pair. The retrieval model has a simple ResNet-50 architecture and is trained with SimCLR loss [6], hence is fully differentiable. The IC model is more complex with a dewarping unit to align the query with the candidate image, followed by a CNN-based feature extraction module to output the editorial prediction and heatmap. Both sub-modules are differentiable with respect to the input image pair and we have demonstrated that adversarial attacks could be performed on both prediction and heatmap in our main paper, as well as an adversarially robust training method to defend against such attacks.

We refer to [6, 45] for more details on the architecture and training strategies of the two above approaches.

C. Additional experiments

Retrieval with exact nearest neighbour search for Black et al. [6] models. First of all, we note that exact nearest neighbour search reported in Table 5 is not practical

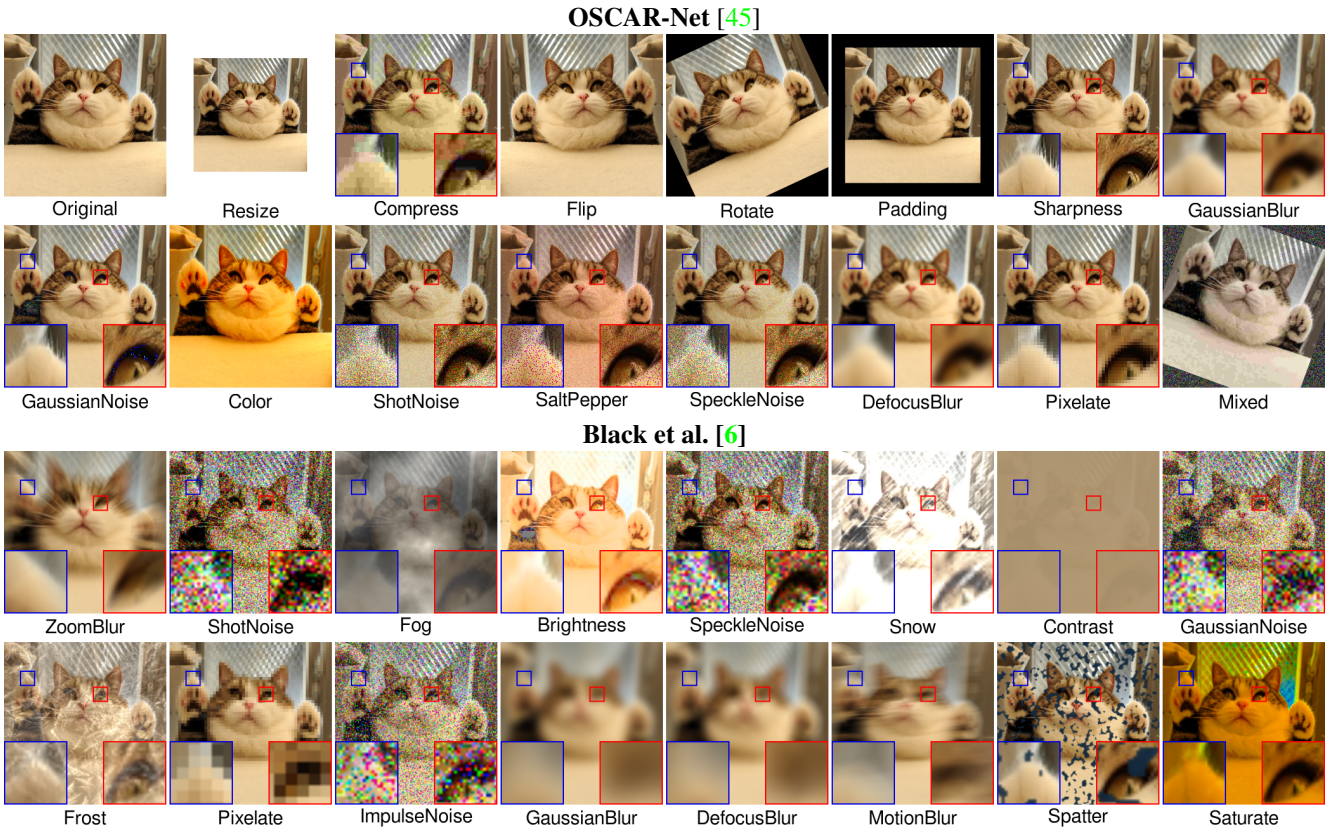

Figure 6. Examples of non-editorial changes applied to the same image from PSBattles according to the query sets used to evaluate the OSCAR-Net [45] and Black *et al*. [6] approaches.

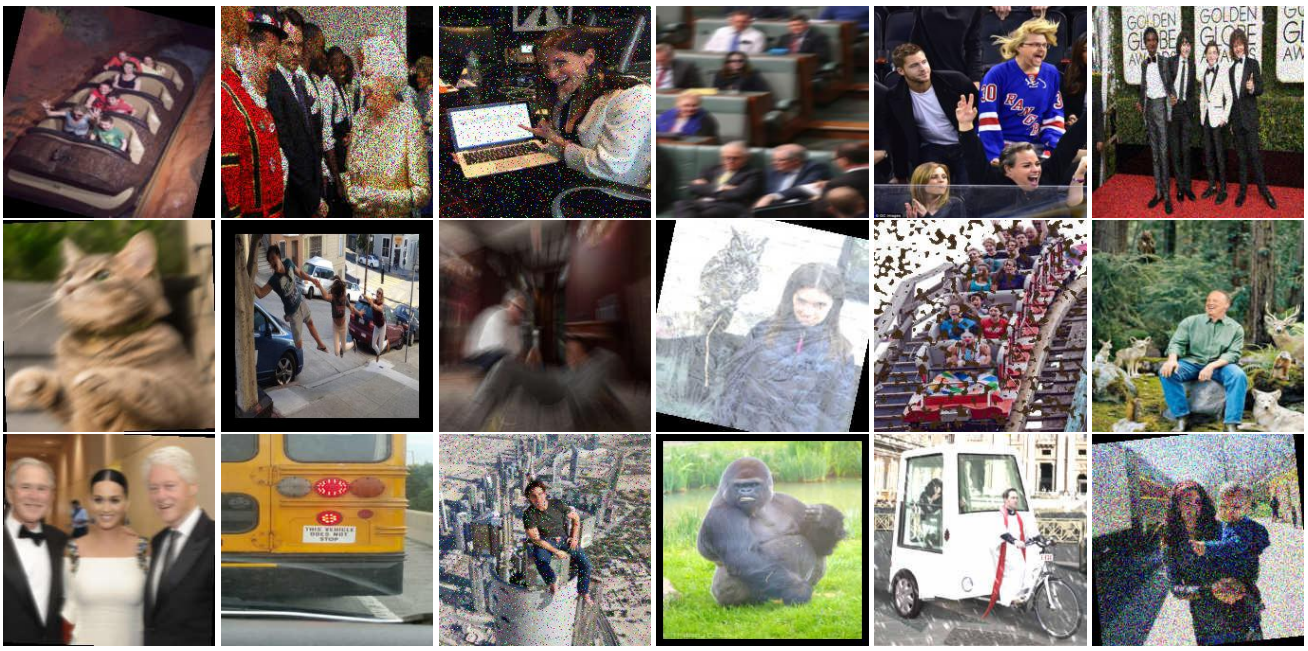

Figure 7. Additional examples of non-editorial changes applied to the images from PSBattles.

	Top-1 and top-100 recall for different query sets											
	Non-editorial distortions				Editorial manipulations				Editorial + non-editorial			
	No attack		ℓ_∞ adversarial		No attack		ℓ_∞ adversarial		No attack		ℓ_∞ adversarial	
Existing models	**R@1**	**R@100**	**R@1**	**R@100**	**R@1**	**R@100**	**R@1**	**R@100**	**R@1**	**R@100**	**R@1**	**R@100**
Standard supervised, ImageNet [47]	45.1	59.3	0.0	0.2	98.3	**99.6**	0.1	0.3	37.3	52.9	0.0	0.3
DeepAugment + AugMix supervised, ImageNet [32]	75.2	84.5	0.2	2.0	**98.5**	**99.6**	0.0	0.6	67.8	80.7	0.0	0.3
Robust supervised, $\varepsilon_\infty = 4/255$, ImageNet [50]	57.3	66.1	30.3	44.0	97.4	99.2	79.7	92.4	51.2	62.0	22.4	38.0
Undefended contrastive, PSBattles [6]	86.2	96.7	0.0	0.0	87.7	95.5	0.0	0.0	70.0	89.5	0.0	0.0
Our new models												
Undefended contrastive, Behance	99.2	99.9	4.8	25.3	94.4	97.6	0.9	9.8	91.9	96.8	2.6	16.1
ARIA contrastive + hashing, $\varepsilon_\infty = 4/255$, Behance	96.8	98.7	83.8	89.3	92.1	96.7	85.2	93.8	87.1	94.5	69.2	82.7
ARIA contrastive + hashing, $\varepsilon_\infty = 8/255$, Behance	93.5	96.5	84.1	90.8	91.4	96.0	87.0	93.9	82.8	91.1	69.7	82.4
ARIA contrastive, $\varepsilon_\infty = 2/255$, Behance	**99.5**	**100.0**	87.7	90.7	96.1	98.6	91.6	96.9	**94.8**	**98.1**	78.6	87.3
ARIA contrastive, $\varepsilon_\infty = 4/255$, Behance	99.4	99.9	90.5	92.7	96.1	98.4	**93.4**	**97.3**	94.7	97.9	**83.3**	90.4
ARIA contrastive, $\varepsilon_\infty = 8/255$, Behance	98.6	99.7	**94.5**	**95.4**	95.5	98.3	93.2	97.1	92.8	97.2	82.9	**90.9**

Table 5. Standard and ℓ_∞ adversarial ($\varepsilon_\infty = 8/255$) top-1 and top-100 recall for different ResNet-50 models evaluated on PSBattles [28]. The database contains original images from PSBattles and 2M distractor images from Stock indexed using the **exact nearest neighbour search** (unlike Table 1 in the main part that used the approximate IVF1024, PQ16 index). We use three query sets based on PSBattles: (1) non-editorial distortions (ImageNet-C and affine) on original images, (2) editorial manipulations but no distortions, (3) editorial manipulations with non-editorial distortions.

	ℓ_∞ **adversarial,** $\epsilon_\infty = 16/255$				ℓ_∞ **adversarial,** $\epsilon_\infty = 32/255$				ℓ_2 **adversarial,** $\epsilon_2 = 5$			
Models	**imAP**	**iR@1**	$\mathbf{F_{mAP}}$	$\mathbf{F_{R@1}}$	**imAP**	**iR@1**	$\mathbf{F_{mAP}}$	$\mathbf{F_{R@1}}$	**imAP**	**iR@1**	$\mathbf{F_{mAP}}$	$\mathbf{F_{R@1}}$
Undefended [45]	7.69	11.08	7.01	9.50	5.84	8.18	5.44	7.28	38.04	45.37	25.64	**26.95**
ARIA, $\varepsilon_\infty = 2/255$ (ours)	22.64	29.93	16.00	17.05	17.09	23.07	13.01	14.58	**54.30**	**61.55**	**27.21**	24.11
ARIA, $\varepsilon_\infty = 4/255$ (ours)	21.04	27.97	15.29	17.01	16.76	22.36	12.89	14.76	47.46	55.44	25.66	24.34
ARIA, $\varepsilon_\infty = 8/255$ (ours)	**41.85**	**49.56**	**23.22**	**21.54**	**40.43**	**47.09**	**22.78**	**21.06**	42.14	51.52	23.31	21.91

Table 6. Performance metrics for attacks *unseen* during training for OSCAR-Net models, using queries from PSBattles. Evaluation is on a query set of digitally manipulated images with no distortions.

	Average precision, no attack				Average precision, ℓ_∞ adversarial attack			
Models	**All classes**	**Non-editorial changes**	**Edit. + non-edit. changes**	**Different images**	**All changes**	**Non-editorial changes**	**Edit. + non-edit. changes**	**Different images**
Undefended ICN [6]	**96.4%**	**98.2%**	91.4%	99.6%	0.6%	0.0%	0.1%	1.6%
ARIA ICN, $\varepsilon_\infty = 2/255$	**96.4%**	91.8%	**97.7%**	**99.7%**	65.0%	21.6%	84.9%	85.6%
ARIA ICN, $\varepsilon_\infty = 4/255$	95.9%	91.6%	97.0%	99.3%	83.1%	67.6%	87.1%	93.9%
ARIA ICN, $\varepsilon_\infty = 8/255$	95.5%	92.2%	95.5%	98.5%	**90.7%**	**86.6%**	**88.7%**	**96.2%**

Table 7. The average precision for the **image comparator network** (ICN) with/without adversarial perturbations of radius $\varepsilon_\infty = 8/255$ over three different classes (depending on the query image that can be either the same image with non-editorial changes, the same image with editorial and non-editorial changes, or a different image).

for databases that contain millions of images and we report it so that we can analyze the performance drop which occurs due to approximate image retrieval. Table 5 suggests that overall the trends and rankings between different methods are the same as in Table 1 from the main part of the paper. At the same time, as expected, the absolute numbers are higher: e.g., standard top-1 recall for the ARIA model trained with $\varepsilon_\infty = 8/255$ is 99.5% compared to 97.3% with the approximate indexing reported in the main part. Such performance drop is uniform over different methods. We can also see that ImageNet-trained models perform well on images with editorial changes. However, we note that the ImageNet models use the embedding dimension of 2048 which is much larger the 256 used by our contrastively trained models and leads to even slower search time.

Robustness of OSCAR-Net models to unseen adversarial perturbations. Table 6 shows the robustness results of OSCAR-Net for perturbations which were unseen during training. These are ℓ_2-bounded perturbations ($\varepsilon_2 = 5$) and ℓ_∞-perturbations of a larger radius compared to those used for training ($\varepsilon_\infty \in \{16/255, 32/255\}$).

The robustness generalises very well to the larger ℓ_∞-perturbations: e.g. with perturbations of size $\varepsilon_\infty = 32/255$ the F_{mAP} score for the undefended model of Nguyen et al. [45] is reduced to 5.44%, but for all our defended models it is at least 12.89%. In the case of our best defended model it is 22.78%. The ℓ_2 perturbations with $\varepsilon_2 = 5$ are not very successful at attacking the OSCAR-Net model, so it is not possible to draw conclusions about robustness in this case. We think that for ℓ_2 perturbations treating the object detector's output as constant can be suboptimal but we leave better attacks tailored to the OSCAR-Net architecture to future work.

Image comparator models: accuracy over classes. We show the results in Table 7 where we report the average precision over three classes depending on the query image that can be either the same image with non-editorial changes, the same image with editorial and non-editorial changes, or

Models	No attack IoU	ℓ_∞ adversarial Targeted IoU
Undefended ICN [6]	58.1%	48.3%
ARIA ICN, $\varepsilon_\infty = 2/255$	**61.5%**	10.0%
ARIA ICN, $\varepsilon_\infty = 4/255$	59.3%	5.4%
ARIA ICN, $\varepsilon_\infty = 8/255$	55.9%	**3.9%**

Table 8. The average intersection over union (**IoU**) between the predicted and ground truth editorial heatmaps for the **image comparator network** (ICN) with/without *targeted* adversarial perturbations of radius $\varepsilon_\infty = 8/255$. Note that unlike other metrics, a lower targeted IoU is better as it implies a smaller overlap of the predicted heatmap with the wrong target heatmap.

a different image. We can see that the standard precision is approximately uniform over different classes but the adversarial precision can be non-uniform. For example, the ARIA ICN model trained with $\varepsilon_\infty = 2/255$ has only 21.6% adversarial precision on the same images with non-editorial changes. However, using a higher ε for ARIA fixes this problem, e.g., for $\varepsilon_\infty = 8/255$ we get 86.6% adversarial precision.

Image comparator models: targeted attacks on heatmaps. We show the results of targeted attacks on the image comparator models in Table 8. For the attack, we target a random cell of a 7×7 heatmap by maximizing the cosine loss. We note that unlike other metrics, a lower targeted intersection over union (IoU) is better as it implies a smaller overlap of the predicted heatmap with the wrong target heatmap. We can observe that ARIA training successfully reduces the success rate of the attack in terms of IoU from 48.3% (undefended ICN) down to 3.9% (ARIA training with $\varepsilon_\infty = 8/255$).

Hash inversion visualizations. Additional hash inversions for randomly chosen images from PSBattles can be found in Fig. 8. We can observe that in many cases hash inversions for the robust model (trained with $\varepsilon_\infty = 4/255$) recover the shapes of original images. This is in contrast with the high-frequency noise which is observed for the standard model.

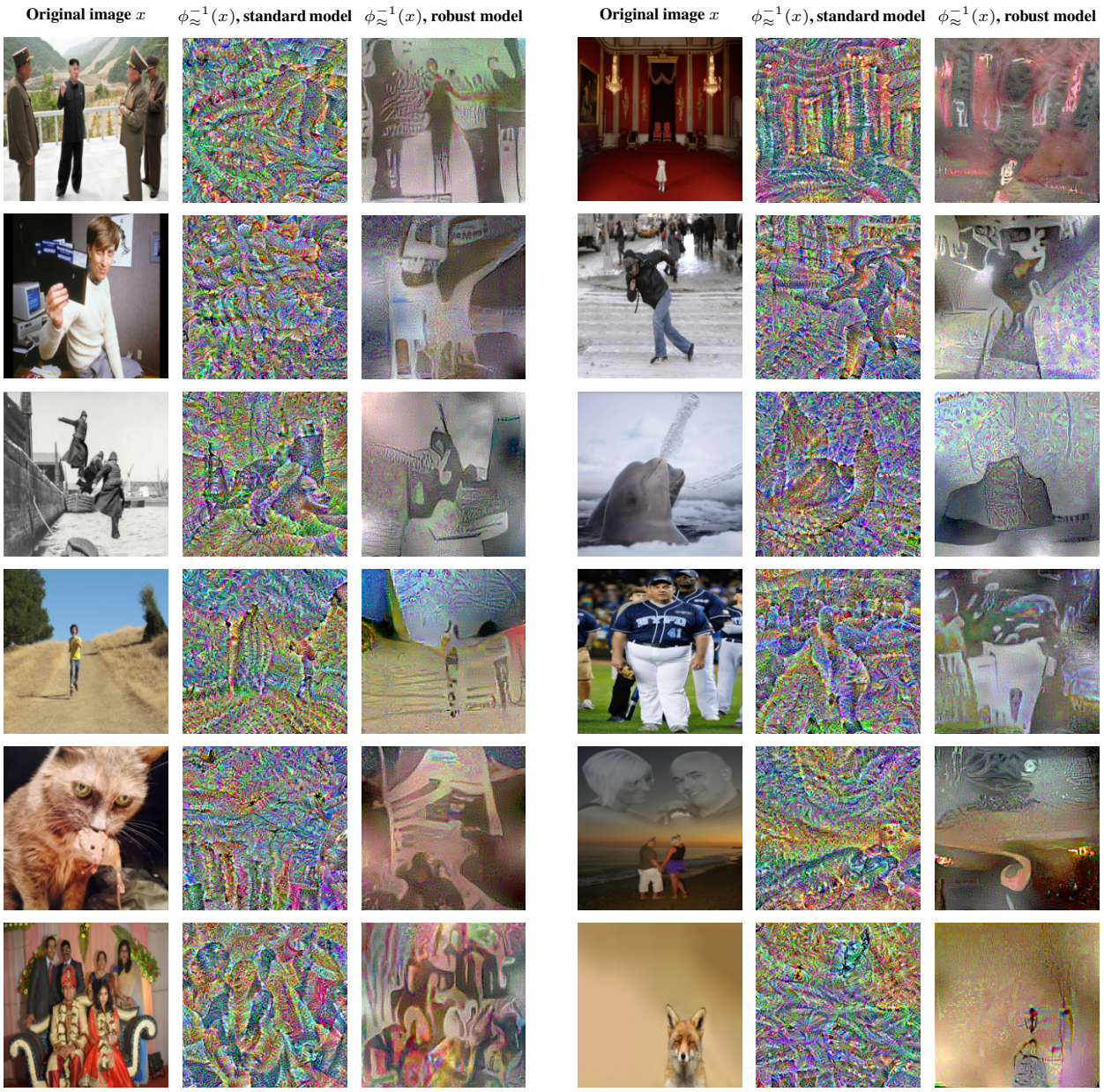

Figure 8. Additional visualizations of the hash inversions $\phi_{\approx}^{-1}(x)$ for twelve original images x (**left**) for a standard model (**middle**) and ARIA model with $\varepsilon_\infty = 4/255$ (**right**), both trained on Behance1M.