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

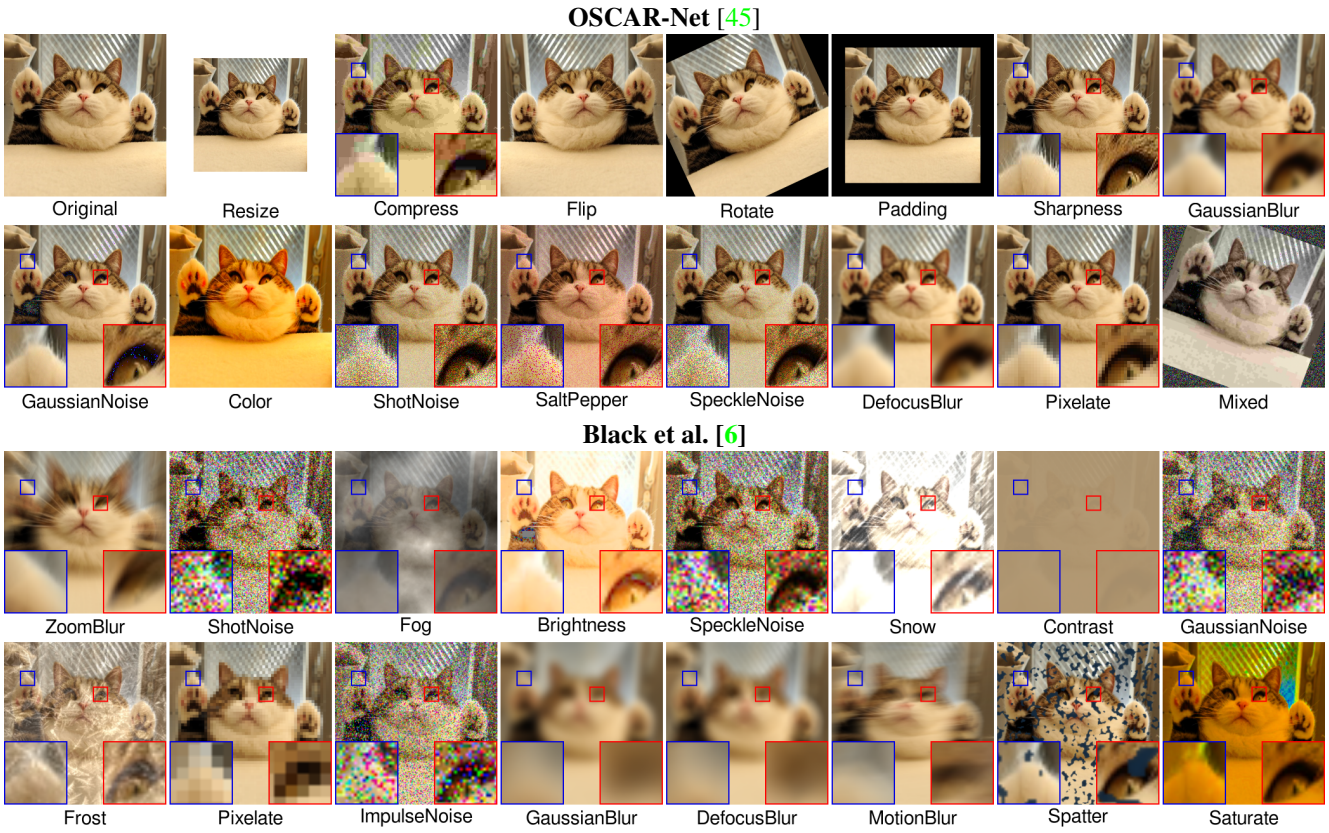

Figure 6. Examples of non-editorial changes applied to the same image from PSBattles according to the query sets used to evaluate the OSCAR-Net [45] and Black *et al*. [6] approaches.

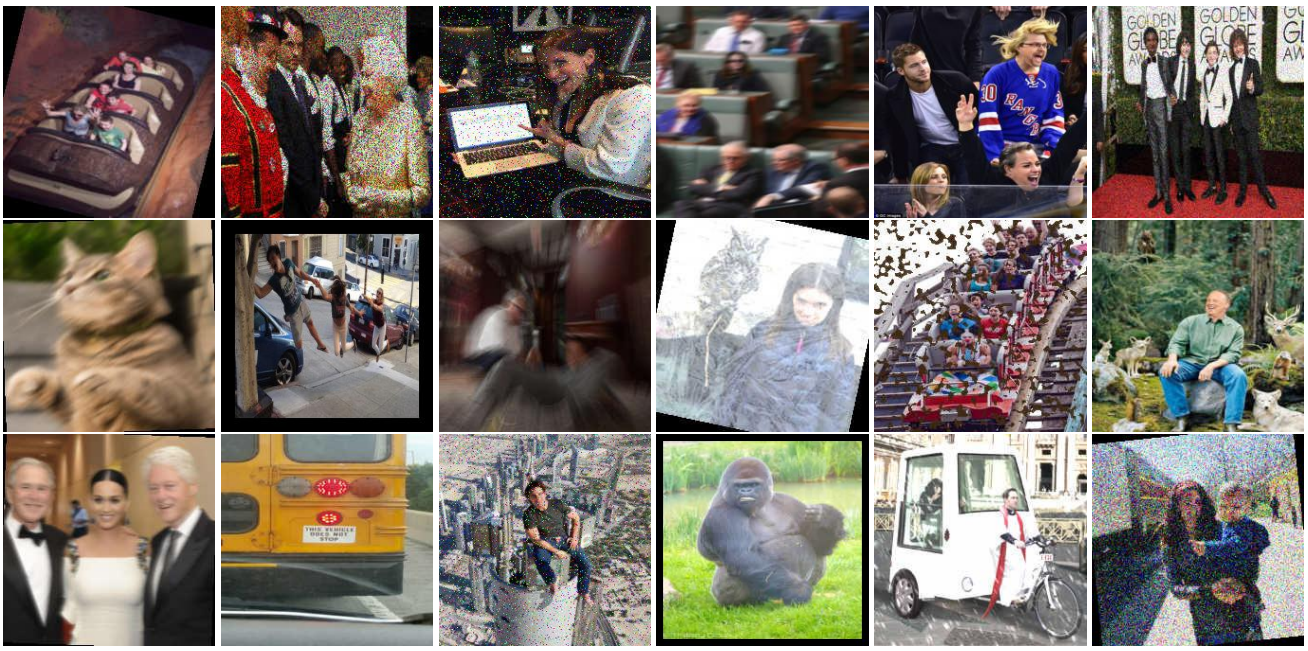

Figure 7. Additional examples of non-editorial changes applied to the images from PSBattles.

	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.

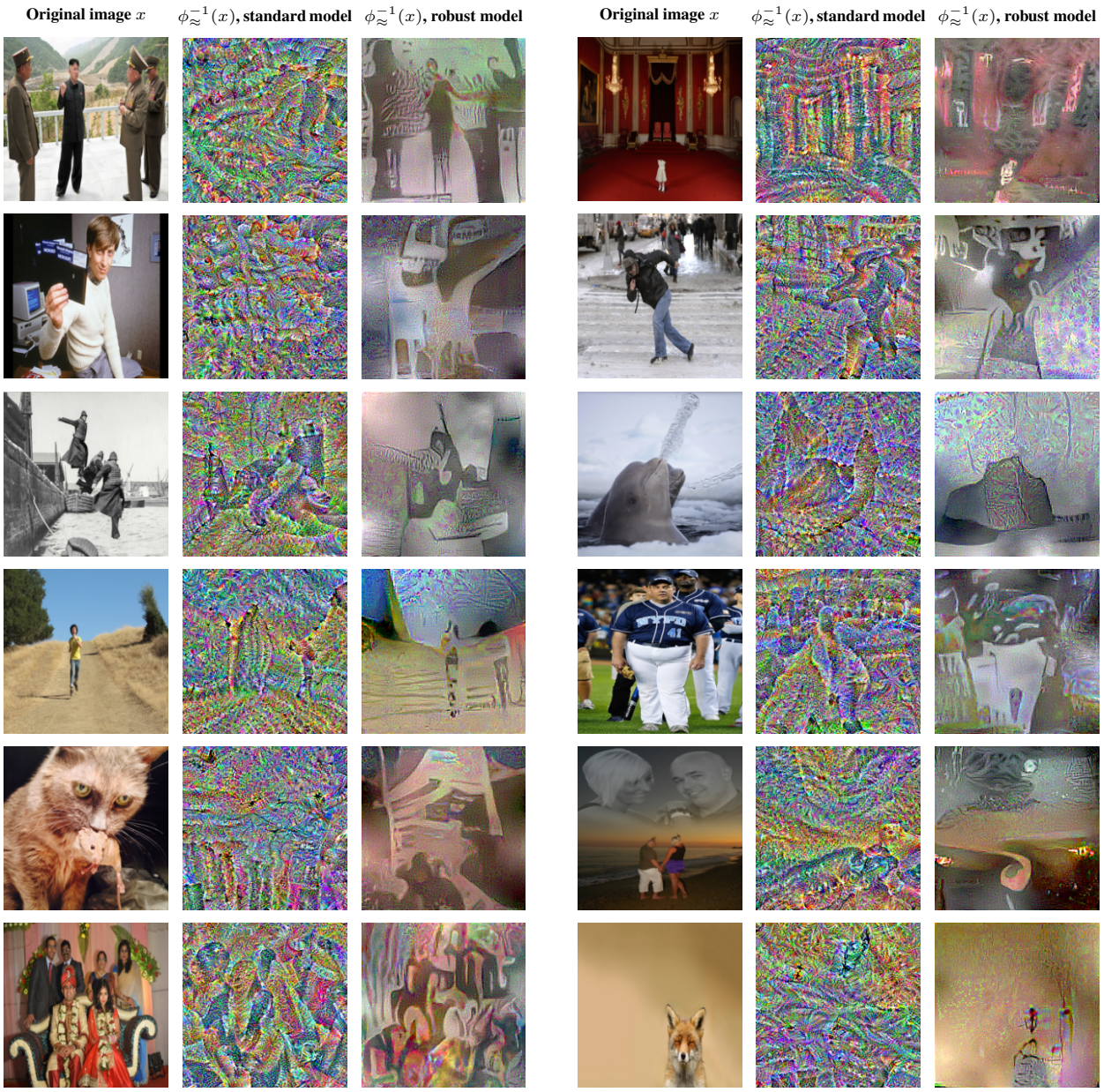

Figure 8. Additional visualizations of the hash inversions $\phi_{\approx}^{-1}(x)$ for twelve original images x (**left**) for a standard model (**middle**) and ARIA model with $\varepsilon_\infty = 4/255$ (**right**), both trained on Behance1M.