Denoise to Protect: A Method to Robustify Visual Recommenders from Adversaries

Felice Antonio Merra*
felmerra@amazon.com
Amazon, Berlin, Germany

Vito Walter Anelli
vitowalter.anelli@poliba.it
Politecnico di Bari, Italy

Tommaso Di Noia
tommaso.dinoia@poliba.it
Politecnico di Bari, Italy

Daniele Malitesta
daniele.malitesta@poliba.it
Politecnico di Bari, Italy

Alberto Carlo Maria Mancino
alberto.mancino@poliba.it
Politecnico di Bari, Italy

ABSTRACT

While the integration of product images enhances the recommendation performance of visual-based recommender systems (VRSs), this can make the model vulnerable to adversaries that can produce noised images capable to alter the recommendation behavior. Recently, stronger and stronger adversarial attacks have emerged to raise awareness of these risks; however, effective defense methods are still an urgent open challenge. In this work, we propose "Adversarial Image Denoiser" (AiD), a novel defense method that cleans up the item images by malicious perturbations. In particular, we design a training strategy whose denoising objective is to minimize both the visual differences between clean and adversarial images and preserve the ranking performance in authentic settings. We perform experiments to evaluate the efficacy of AiD using three state-of-the-art adversarial attacks mounted against standard VRSs. Code and datasets at https://github.com/sisinflab/Denoise-to-protect-VRS.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; **Adversarial retrieval**; • **Security and privacy** → *Web application security*.

KEYWORDS

Recommender Systems; Adversarial Machine Learning; Multimedia Recommendation

ACM Reference Format:
Felice Antonio Merra, Vito Walter Anelli, Tommaso Di Noia, Daniele Malitesta, and Alberto Carlo Maria Mancino. 2023. Denoise to Protect: A Method to Robustify Visual Recommenders from Adversaries. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3591971

1 INTRODUCTION

Recommender systems (RSs) learn to uncover the users' preferences for supporting their decision-making process on the huge catalogs of e-commerce (e.g., Amazon, Zalando), media stream (e.g., Netflix, Spotify), and social networks (e.g., Instagram, Pinterest) websites. While the core of recommendation algorithms is to exploit collaborative filtering (CF) signals, recently, users' visual preferences have been demonstrated to enhance recommendation performance in fashion [17], food [10], and social [6] domains.

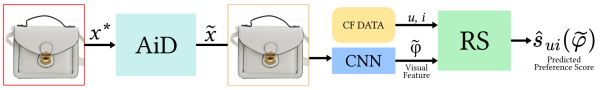


Figure 1: Visual-based RS protected by the Adversarial Image Denoiser (AiD) in the presence of an Adversarial Image (x^*) of the item i. x^* is cleaned by AiD to product $\tilde{x}$ from which the CNN extracts the feature $\tilde{\Phi}$ used by the recommender to measure the relevance score $\hat{s}_{ui}$ of i for the user u.

The economic gain associated with RSs and the performance enhancement proved by their visually-aware variants have made them the target of adversaries [3, 8]. For instance, an adversary can be a seller willing to boost her sales by manipulating items and recommenders with adversarial perturbations [1, 2, 5, 7, 9, 19, 22, 24]. Tang et al. [22] proposed the first adversarial attack procedures for reducing the accuracy of VRSs by directly **altering the image** features via gradient-based perturbation method [15]. Subsequent works focused on adversaries that perform their malicious goals (i.e., pushing an item or a set of items in high positions of the recommendation lists) by directly uploading adversarially perturbed product images. For instance, [19] and [7] have proposed attacks that perturb product images by crafting perturbations that maximize the preference scores predicted by a VRS. Liu and Larson [19] have built the Insider Attack (`WB-INSA`) perturbations by directly employing the gradients measured when maximizing the predicted preference score, while Cohen et al. [7] have used the Sign of the Gradient (`WB-SIGN`) to speed up the perturbation process when still optimizing to increase the recommendability of the target items.

While the literature on proposing attack strategies is rich, only a few works exist on finding solutions to defend visual recommenders. To the best of our knowledge, Adversarial Multimedia Recommendation (AMR) [22] is the only state-of-the-art defensive solution. In this model, Tang et al. [22] integrated VBPR with the adversarial personalized training procedure proposed by He et al. [15] to robustify the model against features perturbation. However, while AMR has been proven to be effective against the adversarial perturbations of the visual features [22], recent attacks by Liu and Larson [19] have tested its limits against adversarial perturbations of product images. Indeed, a small perturbation of a product image

*Corresponding author. Work performed while at Politecnico di Bari, Italy.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3591971

can cause a big variation in the image features against which AMR is not trained on.

Motivated by the lack of adequate defenses, in this work, we propose a novel solution named Adversarial Image Denoiser (AiD). The main intuition of our proposal is to train a model capable to remove the noise from the adversarial images. Technically, we accomplish this by using a U-Net-based denoiser auto-encoder [18] trained on a high- and recommendation- levels guided loss function. The architectural schema of a VRS protected by AID is shown in Figure 1.

To summarize, our main contributions are:

- the proposal of a novel defense solution, named Adversarial Image Denoiser (AiD), that protects VRSs against adversaries that can upload adversarial images on the recommendation platform;
- the verification of AiD protective capabilities against stronger versions of the explored attacks by varying the number of pixels modifiable by the adversary (perturbation budget) and the number of iterations that the adversary can perform to build the malicious noise;
- the validation of AiD robustness by experimentally proving on three real-world visual recommendation datasets that it outperforms AMR on both reducing the variations of the preference scores and the preservation of authentic ranking performance on the target items under two state-of-the-art adversarial attack methods, i.e., `WB-SIGN` and `WB-INSA`.

2 ADVERSARIAL IMAGE DENOISER

To protect a VRS, we propose to remove the noise from x_i, the image of an item i, via a convolutional version of a denoising auto-encoder (DAE) [23] upgraded with a U-net [21] architecture, named DUNET [18]. We define $d_\Omega : x^* \rightarrow \widetilde{x}$ as the denoising function where Ω are the AiD parameters, and x^* is an adversarially perturbed item image (we omit i for readability reasons). DUNET learns how to reconstruct the adversarial noise (δ) to be removed from the adversarial sample such that $\widetilde{x} = x^* - d\widetilde{x}$, where the denoised image $\widetilde{x}$ should be equal or similar to the clean one x while $d\widetilde{x}$, the AiD's learned adversarial noise, should be equal to δ (i.e., the adversarial perturbation added to x to make x^*). AiD is composed of a feedforward (encoder) and a feedback (decoder) path connected with lateral links (fuse operation) going from the encoder layers to their corresponding decoder ones. The input and output shapes are both 224x224x3 which are the input dimensions of ResNet50 [12], the CNN used in our experiments. Architectural details are in [18].

To protect the recommendation performance while preserving the quality of images, we train AiD with a loss function ($\mathcal{L}_{AiD}$) composed of two parts, $\mathcal{L}_{HGD}$ and $\mathcal{L}_{RGD}$. The former is a high-level guided denoiser (HGD) loss function that, differently from standard pixel-level guided denoiser loss, is robust against the amplification of adversarial noise along with the last layers of CNNs (the layers used in VRSs) [18].

Definition 1 (High-level Guided Loss). *Let φ be the item visual features of a clean image x, let $\widetilde{\varphi}$ be the features extracted from the denoised image version $\widetilde{x}$, then the high-level guided denoiser (HGD) loss function is defined as*

$$\mathcal{L}_{HGD} = ||\varphi - \widetilde{\varphi}|| \tag{1}$$

Algorithm 1: Training of AiD

Input: CF data **S**, Adv. Images $\mathcal{D}_T^*$ (training), $\mathcal{D}_V^*$ (validation).
Initial Parameters: Θ and Φ (fixed), Ω (trainable)
Output: Ω for AiD
for *epoch* $= 1, ..., N_{ep}$ **do**
  $ValidLoss \leftarrow -\infty$, $\Omega_{BEST} \leftarrow \Omega$
  for $x^* \in \mathcal{D}_T^*$ **do**
    // Compute AiD Loss
    $x \leftarrow x^*$ corresponding clean image
    $\widetilde{x} \leftarrow d(x^*)$
    $\widetilde{\varphi}, \varphi \leftarrow f(\widetilde{x}), f(x)$
    $\mathcal{L}_{AiD} \leftarrow ||\varphi - \widetilde{\varphi}|| + \eta \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \left(\hat{s}_{ui}(\varphi) - \hat{s}_{ui}(\widetilde{\varphi}, u, i)\right)^2$
    // Compute Ω Gradients and Perform SGD-updates
    $g_\Omega \leftarrow \partial \mathcal{L}_{AiD}(\Omega)/\partial\Omega$
    $\Omega \leftarrow \Omega + \mu g_\Omega$
  end
  // Compute Validation Loss on $\mathcal{D}_V^*$
  $EpValidLoss \leftarrow 0$
  for $x^* \in \mathcal{D}_T^*$ **do**
    $EpValidLoss \leftarrow EpValidLoss + \mathcal{L}_{AiD}(x^*)$
  end
  $EpValidLoss \leftarrow EpValidLoss/|\mathcal{D}_V^*|$
  if $\mathcal{L}_{AiD}(\mathcal{D}_V^*) \leq ValidLoss$ **then**
    ValidLoss $\leftarrow EpValidLoss$
    $\Omega_{BEST} \leftarrow \Omega$
  end
end
$\Omega \leftarrow \Omega_{BEST}$

, where the denoiser is explicitly trained to reconstruct the original visual feature (φ) lately used in the VRS.

The latter component of $\mathcal{L}_{AiD}$ is introduced to make AiD sensitive in preserving the recommendation behavior in authentic settings.

Definition 2 (Recommendation-level Guided Loss). *Let i be the attacked item with (x, x^*)-pair of clean and perturbed images, let $\widetilde{x}$ be the image denoised by AiD, and $\hat{s}_{ui}$ be the predicted score for the user $u \in \mathcal{U}$ on the item $i \in \mathcal{I}$, where $\mathcal{U}$ and $\mathcal{I}$ are the sets of users and items, then RGD loss is:*

$$\mathcal{L}_{RGD} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \left(\hat{s}_{ui}(\varphi) - \hat{s}_{ui}(\widetilde{\varphi}, u, i)\right)^2 \tag{2}$$

After having introduced the two components of the AiD loss, the defense is trained by optimizing the following problem:

$$\arg\min_{\Omega} \mathcal{L}_{AiD} = \arg\min_{\Omega} \left(\mathcal{L}_{HGD} + \eta \mathcal{L}_{RGD}\right) \tag{3}$$

where η is a coefficient to control the impact of $\mathcal{L}_{RGD}$. Note that being our defense solution applied on pre-trained visual recommenders, the parameters of the CNN, and the parameters of the VRS, are fixed, while Ω, the AiD parameters, need to be learned to perform the cleaning of the adversarially perturbed item images. Algorithm 1 shows the pseudocode used for the training of AiD.

3 EXPERIMENTAL SETUP

Datasets. We test the recommendation performance for our defensive method on three datasets. The first two datasets are `Amazon Boys & Girls` [14, 20]) and `Amazon Men` [13, 14, 20]. These are fashion datasets containing feedback on clothing articles. Following the methodology in [13, 14], we filter them with the 5-core technique by removing the users, as well as, the items with less than five feedbacks. The first dataset has 1425 users, 4507 items, and 9213 feedbacks, the second has 16278 users, 31750 items, and 113106 feedbacks. The third dataset is `Pinterest` [11, 16]. We apply 5-core filtering on users, producing a version of 30375 users, 19976 items, and 395418 feedback. We split each dataset into the train, validation, and test sets by adopting the temporal *leave-one-out protocol* for `Amazon Boys & Girls` and `Amazon Men`, and random *leave-one-out protocol* for `Pinterest` since it does not have temporal information.

Since AiD is a model that has to learn to remove noise from adversarially perturbed images, we build a dataset of cleaned and noised images. We select 200 random items in the catalog and run `WB-SIGN` and `WB-INSA` with $T \in 1, 4, 8$ and $\epsilon \in \text{rnd}([1, 16])$, where rnd$(\cdot)$ uniformly samples one integer inside the defined range. Note that bigger T and ϵ result in stronger attacks [7, 19]. Then, we split the target items in training ($\mathcal{D}_T^*$), validation ($\mathcal{D}_V^*$), and test ($\mathcal{D}_\tau^*$) sets using the 8:1:1 proportion.

Recommendation Model. We test the defense solution on **VBPR** (Visual Bayesian Personalized Ranking from Implicit Feedback) [14], the standard visual-based MF recommender widely adopted in to test the robustness of multimedia recommenders in adversarial settings [7, 19].

Defense Baseline. We test the only existing baseline for the protection of visual recommenders that is **AMR** (Adversarial Multimedia Recommendation) [22], an extension of VBPR that integrates the adversarial training procedure proposed by [15].

Attacks. We used two state-of-the-art attacks, i.e., `WB-SIGN` and `WB-INSA`. `WB-SIGN` [7] builds an attack by computing the sign of the gradient of the recommendation score function $\hat{s}(\cdot)$ with respect to all the pixels p_x in the product image x. In particular, the authors apply the chain rule to evaluate the gradient direction. `WB-INSA` [19] adds an adversarial perturbation on the item images through an iterative methodology to maximize the predicted scores over the users in the platform.

Attack Evaluation. We start by analyzing the adversary's capacity in increasing the predicted preference score measuring the mean variation of the preference scores across all the attacked (target) items defined by Burke et al. [4] as $\text{PS} = \frac{1}{|\mathcal{D}_\tau^*|} \sum_{j \in \mathcal{D}_\tau^*} \left(\hat{s}_{uj}(x^*) - \hat{s}_{uj}(x)\right)$ where $\hat{s}_{uj}(x)$ is the score predicted on the authentic image associated with the item j against which the adversary has performed an attack — whose altered predicted score is $\hat{s}_{uj}(x^*)$. Then, to evaluate the adversarial effects on robustifying the recommendation ranking, we start by introducing the Attack Hit Ratio (aHR@K) following the definition in [7]. In particular, let $\text{attack}_{\text{hit}}@\text{K}(j, u)$ be a hit function that is 1 when the target item is in the top-K list of the user u, 0 otherwise, then $\text{aHR@K} := \frac{1}{|\mathcal{D}_\tau^*|} \sum_{j \in \mathcal{D}_\tau^*} \frac{1}{|\mathcal{U}|} \sum_{u \in |\mathcal{U}|} \text{attack}_{\text{hit}}@\text{K}(j, u)$, where, $j \in \mathcal{D}_\tau^*$ indicates that $\text{attack}_{\text{hit}}@\text{K}$ is measured on a target item whose image has been adversarially perturbed. To measure whether the defense has been optimal in preserving the original behavior of the recommender, we introduce a novel measure, i.e., Ranking Robustness, defined in Definition 3.

Table 1: PS measured on $(\epsilon = 4, T = 1)$-attacks. We bold values with effective defenses.

Dataset	Attack	No Def. PS^{VBPR}	Base Def. PS^{AMR}	Our PS^{AiD}	Our $\text{PS}^{AMR+AiD}$
Amazon B&G	WB-INSA	0.8250	1.0432	**0.1410**	**0.2193**
	WB-SIGN	1.8466	1.3349	**1.2668**	**1.1183**
Amazon Men	WB-INSA	2.2217	2.2418	**0.5560**	**0.6057**
	WB-SIGN	2.2413	2.5066	**1.0005**	**1.0969**
Pinterest	WB-INSA	1.9113	**1.3108**	**0.4931**	**0.2205**
	WB-SIGN	1.8929	**1.2817**	**0.6434**	**0.3345**

Definition 3 (Ranking Robustness at K (RR@K)). *Let* aHR@K_{bef} *and* aHR@K_{aft} *be the attack hit ratios measured before and after the attack, respectively, and let* $\Delta\text{aHR@K} = \frac{\text{aHR@K}_{aft} - \text{aHR@K}_{bef}}{\text{aHR@K}_{bef}}$ *be the difference ratio, then* RR@K *is defined as follows:*

$$\text{RR@K} = \left| \frac{\Delta\text{aHR@K}^{w}}{\Delta\text{aHR@K}^{wo}} \right| \tag{4}$$

where $\Delta\text{aHR@K}^{w}$ *and* $\Delta\text{aHR@K}^{wo}$ *are measured when the VRS is protected with and without AiD.*

RR@K $\simeq 0$ means that AiD has reached optimal performance, RR@K $\simeq 1$ is the scenario where AiD does not impact the attacks' efficacy, and RR@K $>> 1$ is the awful situation where the AiD could have considerably impacted the presence of target items in the top-K lists. Note that ΔaHR@K can be negative when the hit ratio after the attack is smaller than before.

Reproducibility Details. We train each visual recommender by varying the learning rate in $\{0.0001, 0.001, 0.01\}$ and the regularization coefficients in $\{0.00001, 0.001\}$, and fixing the number of training epochs to 100, the batch size to 256, and the number of latent factors to 128. The adversarial epochs used for training AMR are 50 (performed after the initial 50 epochs with standard VBPR training) with the adversarial regularization coefficient and ϵ set to 1. Then, we train the proposed AiD for 100 epochs. We set $\eta = 0$ for the first 50 epochs to allow the denoiser to focus on the high-level guided reconstruction. Then, we train AiD for additional 50 epochs, fixing $\eta = 1$, to learn how to preserve the recommendation-level quality. We set the batch size to 16, and, following the experimental protocol by [18], we train the denoiser with the Adam optimizer with $\mu = 0.001$ (the learning rate of the denoiser), validating AiD at the end of each epoch, and conducting the experiment with the checkpoint with the smallest validation loss.

4 RESULTS AND DISCUSSION

In this section, we perform, analyze and discuss the experimental results answering three research questions.

[RQ1] ***Can AiD preserve the original predicted preference scores?*** We start by comparing *PS* with $\epsilon = 4$ and $T = 1$ whose results are reported in Table 1. It can be seen that the use of AiD has been effective in reducing the average prediction shifts for all combinations of black-box and white-box attacks performed against VBPR.

Table 2: Ranking Robustness at 50. Note that AMR is the baseline defense strategy. We bold the most effective defense and we italic the second-to-best defenses.

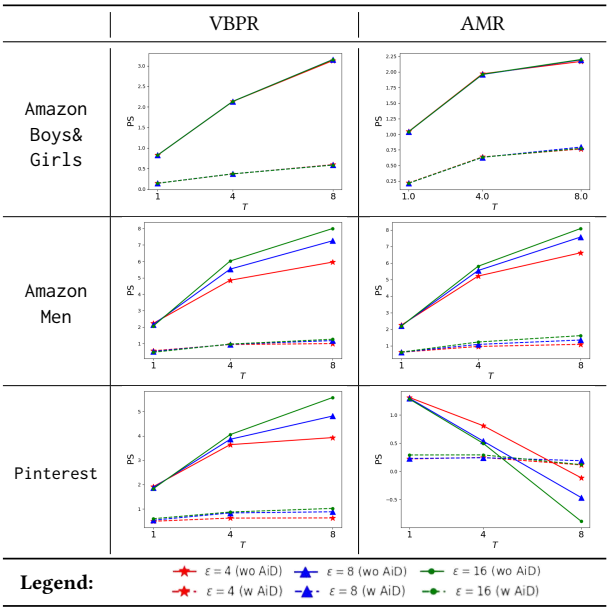
Dataset	Attack	Base Def.	Our	
		RR^{AMR}	RR^{AiD}	$RR^{AMR+AiD}$
Amazon B&G	WB-INSA	0.5138	**0.1509**	*0.2627*
	WB-SIGN	0.2972	*0.6854*	**0.2088**
Amazon Men	WB-INSA	0.4270	**0.0930**	*0.1145*
	WB-SIGN	0.4828	**0.0771**	*0.2557*
Pinterest	WB-INSA	3.2402	*1.1232*	**0.0423**
	WB-SIGN	11.9042	**3.3317**	*3.5820*

For instance, PS^{VBPR} is always reduced by more than three times for each WB-INSA attack independently of the datasets (e.g., 0.1410 < 0.8250; 0.5560 < 2.2217; and 0.4931 < 1.9113 from the top of the table to the bottom). In addition, we can see that protecting VBPR with AiD is more effective than protecting it with only AMR. Indeed, PS^{AiD} are steadily closer to 0 than PS^{AMR} (e.g., 0.4931 vs. 1.3108 for WB-INSA in Pinterest). In addition, we can see from Table 1 that the integration of AiD with AMR can make the defense even stronger in some cases. Indeed, $PS^{AMR+AiD}$ is 0.2205 for the Pinterest example shown below. These results endorse that *AiD outperforms AMR in reducing the impact of the attack on the original preference scores of target items, and it can be empowered when AiD is combined with an adversarially trained recommender (AMR).*

[RQ2] ***Can AiD preserve the original predicted ranking lists?*** To verify if the defense is effective when protecting the recommender when the adversary wants to increase the average position of target items in the recommendation lists, we report in Table 2 the results of the proposed RR measured on top-50 recommendation lists. We can see that applying the proposed denoising approach has been adequate in most of the tested scenarios. Indeed, the fact that the RR values are mostly smaller than 1 in any attack scenario demonstrates that the presence of the AiD has reduced the adversaries' capability to push the target items in higher recommendation positions. Additionally, it is interesting to observe that the only three scenarios in which the RR is higher than one are related to cases where the adversarial attacks were not very powerful in the not-defended setting. In these contexts, we note that the best and second-to-best RR values are related to the usage of AiD either alone or in combination with AMR. We can summarize that *AiD effectively reduces both the adversaries' impact in varying the predicted preference scores and, as shown in this paragraph, the changes of target items' positions in the recommendation lists.*

[RQ3] ***Is AiD effective with stronger and stronger attacks?*** Table 3 presents six plots that show PS^{wo} and PS^{w} when varying the number of steps $T \in \{1, 4, 8\}$ and the perturbations budget $\epsilon \in \{4, 8, 16\}$ for the WB-INSA attack performed against both recommenders being the WB-attack with the lowest PS^{w} values. First, analyzing the continuous lines, we get evidence that *the adversary is more and more effective in the absence of the denoiser with bigger T and ϵ*. Then, the application of the denoiser (dotted lines) intercepted the attempts of stronger adversaries by always showing very low prediction shifts. For instance, it can be noted that while PS^{wo} increases from values close to 1 to higher than 3 for

Table 3: Prediction Shift (PS) of the WB-INSA attack by varying the budget ($\epsilon \in \{4, 8, 16\}$) and the iterations ($T \in \{1, 4, 8\}$).



VBPR trained on Amazon Boys & Girls, PS^{w} always remains less than 1. The same efficient behavior can also be noted on the other plots, and, above all, in AMR for Pinterest, we can observe that the prediction shifts have been maintained close to 0 even when the attack becomes very strong ($\epsilon = 8$). We can conclude that *AiD guarantees low variations of the predicted scores, even with stronger and stronger adversarial attacks.*

5 CONCLUSION

This work has proposed a novel defense to protect visual-based recommender systems (VRS) under adversarial attacks. The defense solution, named Adversarial Image Denoiser (AiD), is a novel model component trained to clean up perturbed images by minimizing an image- and recommendation-aware reconstruction loss. We have investigated the defense performance on three real-world datasets and two standard visual recommender models in adversarial settings under two attack strategies. Experiments have confirmed AiD as a practical solution since it reduced the attacks' power in varying the predicted preference scores and the positions of attacked products outperforming AMR, the state-of-the-art defensive solution. We plan to empower AiD against possible novel and stronger adversarial attacks that might break it. Finally, we want to adapt AiD in multi-modal settings.

ACKNOWLEDGMENTS

The authors acknowledge partial support of the projects PON ARS01_00876 BIO-D, Casa delle Tecnologie Emergenti della Città di Matera, Fincons Smart Rights Management Platform, PIA Servizi Locali 2.0, H2020 Passepartout - Grant n. 101016956, PIA ERP4.0, IPZS-PRJ4_IA_NORMATIVO, Secure Safe Apulia. This work has been carried out while A. C. M. Mancino was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome in collaboration with Polytechnic University of Bari.

REFERENCES

[1] Vito Walter Anelli, Yashar Deldjoo, Tommaso Di Noia, Daniele Malitesta, and Felice Antonio Merra. 2021. A Study of Defensive Methods to Protect Visual Recommendation Against Adversarial Manipulation of Images. In *SIGIR*. ACM, 1094–1103.

[2] Vito Walter Anelli, Yashar Deldjoo, Tommaso Di Noia, and Felice Antonio Merra. 2021. A Formal Analysis of Recommendation Quality of Adversarially-trained Recommenders. In *CIKM*. ACM, 2852–2856.

[3] Vito Walter Anelli, Yashar Deldjoo, Tommaso Di Noia, and Felice Antonio Merra. 2022. Adversarial Recommender Systems: Attack, Defense, and Advances. In *Recommender Systems Handbook*. Springer US, 335–379.

[4] Robin Burke, Michael P. O'Mahony, and Neil J. Hurley. 2015. Robust Collaborative Recommendation. In *Recommender Systems Handbook*. Springer, 961–995.

[5] Huiyuan Chen, Kaixiong Zhou, Kwei-Herng Lai, Xia Hu, Fei Wang, and Hao Yang. 2022. Adversarial Graph Perturbations for Recommendations at Scale. In *SIGIR*. ACM, 1854–1858.

[6] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive Collaborative Filtering: Multimedia Recommendation with Item- and Component-Level Attention. In *SIGIR*. ACM.

[7] Rami Cohen, Oren Sar Shalom, Dietmar Jannach, and Amihood Amir. 2021. A Black-Box Attack Model for Visually-Aware Recommender Systems. In *WSDM*. ACM, 94–102.

[8] Yashar Deldjoo, Tommaso Di Noia, and Felice Antonio Merra. 2021. A Survey on Adversarial Recommender Systems: From Attack/Defense Strategies to Generative Adversarial Networks. *ACM Comput. Surv.* 54, 2 (2021), 35:1–35:38.

[9] Tommaso Di Noia, Daniele Malitesta, and Felice Antonio Merra. 2020. TAaMR: Targeted Adversarial Attack against Multimedia Recommender Systems. In *DSN Workshops*. IEEE, 1–8.

[10] David Elsweiler, Christoph Trattner, and Morgan Harvey. 2017. Exploiting Food Choice Biases for Healthier Recipe Recommendation. In *SIGIR*. ACM, 575–584.

[11] Xue Geng, Hanwang Zhang, Jingwen Bian, and Tat-Seng Chua. 2015. Learning Image and User Features for Recommendation in Social Networks. In *ICCV*. IEEE Computer Society, 4274–4282.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. IEEE Computer Society, 770–778.

[13] Ruining He and Julian J. McAuley. 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In *WWW 2016*.

[14] Ruining He and Julian J. McAuley. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *AAAI*. AAAI Press, 144–150.

[15] Xiangnan He, Zhankui He, Xiaoyu Du, and Tat-Seng Chua. 2018. Adversarial Personalized Ranking for Recommendation. In *SIGIR*. ACM, 355–364.

[16] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *WWW*. ACM, 173–182.

[17] Yang Hu, Xi Yi, and Larry S. Davis. 2015. Collaborative Fashion Recommendation: A Functional Tensor Factorization Approach. In *ACM Multimedia*. ACM, 129–138.

[18] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. 2018. Defense Against Adversarial Attacks Using High-Level Representation Guided Denoiser. In *CVPR*. Computer Vision Foundation / IEEE Computer Society, 1778–1787.

[19] Zhuoran Liu and Martha A. Larson. 2021. Adversarial Item Promotion: Vulnerabilities at the Core of Top-N Recommenders that Use Images to Address Cold Start. In *WWW*. ACM / IW3C2, 3590–3602.

[20] Julian J. McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *SIGIR 2015*.

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI (3) (Lecture Notes in Computer Science, Vol. 9351)*. Springer, 234–241.

[22] Jinhui Tang, Xiaoyu Du, Xiangnan He, Fajie Yuan, Qi Tian, and Tat-Seng Chua. 2020. Adversarial Training Towards Robust Multimedia Recommender System. *IEEE Trans. Knowl. Data Eng.* 32, 5 (2020), 855–867.

[23] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *ICML (ACM International Conference Proceeding Series, Vol. 307)*. ACM, 1096–1103.

[24] Chenwang Wu, Defu Lian, Yong Ge, Zhihao Zhu, Enhong Chen, and Senchao Yuan. 2021. Fight Fire with Fire: Towards Robust Recommender Systems via Adversarial Poisoning Training. In *SIGIR*. ACM, 1074–1083.