

Article

ReGeNet: Relevance-Guided Generative Network to Evaluate the Adversarial Robustness of Cross-Modal Retrieval Systems

Chao Hu ¹, Yulin Yang ¹, Yan Chen ², Li Chen ¹, Chengguang Liu ³, Yuxin Li ⁴, Ronghua Shi ¹ and Jincai Huang ^{3,*}

- ¹ School of Electronic Information, Central South University, Changsha 410017, China; huchao@csu.edu.cn (C.H.); 194701017@csu.edu.cn (Y.Y.); 194702040@csu.edu.cn (L.C.); shirh@csu.edu.cn (R.S.)
² Logistics Department, Central South University, Changsha 410017, China; 37050006@csu.edu.cn
³ Big Data Institute, Central South University, Changsha 410017, China; liuchengguang@csu.edu.cn
⁴ College of Computing and Data Science, Nanyang Technological University, Singapore 639798, Singapore; lyx010329@gmail.com
* Correspondence: huangjincaisu@csu.edu.cn

Abstract

Streaming media data have become pervasive in modern commercial systems. To address large-scale data processing in intelligent transportation systems (ITSs), recent research has focused on deep neural network-based (DNN-based) approaches to improve the performance of cross-modal hashing retrieval (CMHR) systems. However, due to their high dimensionality and network depth, DNN-based CMHR systems inherently suffer from vulnerabilities to malicious adversarial examples (AEs). This paper investigates the robustness of CMHR-based ITS systems against AEs. Prior work typically formulates AE generation as an optimization-driven, iterative process, whose high computational cost and slow generation speed limit research efficiency. To overcome these limitations, we propose a parallel cross-modal relevance-guided generative network (ReGeNet) that captures the semantic characteristics of the target deep hashing model. During training, we design a relevance-guided adversarial generative framework to efficiently learn AE generation. During inference, the well-trained parallel adversarial generator produces adversarial cross-modal data with effectiveness comparable to that of iterative methods. Experimental results demonstrate that ReGeNet can generate AEs significantly faster while achieving competitive attack performance relative to iterative-based approaches.

Keywords: adversarial attack; cross-modal Hamming retrieval; adversarial generative network; deep hashing

MSC: 68T07; 68T45; 68T05



Academic Editor: Ioannis Tsoulos

Received: 1 December 2025

Revised: 23 December 2025

Accepted: 27 December 2025

Published: 30 December 2025

Copyright: © 2025 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Over the past decade, deep neural networks have been extensively studied and widely deployed in intelligent transportation systems (ITSs). In monitoring applications in particular, large volumes of streaming data are collected for traffic flow analysis [1], behavior recognition and tracking [2,3], and video summarization [4]. To enable efficient retrieval of massive multi-modal data (e.g., images and text), most commercial systems integrate cross-modal hashing retrieval (CMHR) frameworks to effectively fuse heterogeneous modalities. Specifically, a typical CMHR system allows users to submit queries in one modality, such as images or text, and returns semantically relevant results in another modality. Recent studies [5–7] have widely adopted deep neural network (DNN)-based approaches to advance

CMHR systems by developing efficient feature extraction and federated hashing learning techniques to capture latent cross-modal representations, thereby significantly improving retrieval performance.

However, Szegedy et al. [8] demonstrated that, due to the highly nonlinear nature of DNNs, a malicious user can induce substantial changes in model outputs by applying only slight perturbations to the inputs, potentially leading to system malfunction. Subsequent studies [9–11] further confirmed that deep neural network-based (DNN-based) models are inherently vulnerable to imperceptible adversarial perturbations. These intrinsic properties render even state-of-the-art deep learning models susceptible to carefully crafted adversarial examples (AEs), causing incorrect retrieval results and posing serious security risks. In general, AEs are synthesized by introducing adversarial perturbations into raw data that remain indistinguishable to human perception (e.g., slight pixel modifications in images or subtle textual manipulations such as character reordering or punctuation insertion). Adversarial attacks are typically categorized into white-box attacks [12–14] and black-box attacks [15,16]. In white-box settings, the attacker has full access to the target model, including its parameters and architecture, whereas in black-box settings, such information is unavailable. Existing studies [17–20] have extensively explored white-box attacks in single-modal hashing retrieval, while cross-modal attacks [21,22] remain relatively underexplored.

Prior work has made substantial progress in adversarial attacks on single-modal systems (e.g., image recognition [23] and text translation [24]), whereas research on cross-modal scenarios remains limited. Existing adversarial attack methods are generally divided into iterative-based approaches [8] and generative-based approaches [20,25]. Iterative-based methods formulate the attack as an optimization problem and progressively approximate the optimal adversarial point via gradient-based techniques. However, these approaches typically require repeated gradient computations or backpropagation through the target model, resulting in high computational overhead and slow generation speed. In contrast, generative-based methods [20,26] have recently attracted increasing attention due to their fast generation capability and strong attack performance achieved through well-designed generative adversarial networks. In this work, we focus on rapid AE generation under the white-box attack setting, while leaving black-box scenarios for future investigation.

In this paper, we propose a novel generative-based framework to efficiently evaluate the adversarial robustness of ITS. We introduce the **Relevance-Guided Generative Network (ReGeNet)**, which enhances relational similarity between supervised semantic information and a parallel generative network. Specifically, ReGeNet is designed to learn a general adversarial function for generating cross-modal AEs. Unlike iterative-based methods that require a large number of iterations and incur substantial computational cost due to instance-specific optimization, ReGeNet aims to approximate the decision boundary of the target model, thereby inducing retrieval results with semantically irrelevant content. The proposed framework consists of two phases: training and generation. During training, we design a pipeline with two workflows. In the first workflow, paired cross-modal training data are fed into parallel image and text generators, whose outputs are evaluated by the target CMHR system. In the second workflow, a label network leverages an adversarial similarity matrix and fuses latent features from both generators, which are then input into a graph convolutional network (GCN) for optimization. To guide training, we design three objective functions: a relevance-guided adversarial loss, a reconstruction loss, and an adversarial distinguish loss. The relevance-guided adversarial loss maximizes the Hamming distance between the generated adversarial data and the original inputs, while the reconstruction loss encourages visual indistinguishability between adversarial and original images. The adversarial distinguish loss further enforces imperceptible pertur-

bations while incorporating adversarial category information. All model parameters are jointly optimized using the ADAM optimizer [27]. During the generation phase, given an image–text pair and the trained model, the parallel generators efficiently produce effective adversarial examples.

We evaluate our approach on two widely used benchmark datasets: FLICKR-25K [28] and NUS-WIDE [29]. The experimental results demonstrate that ReGeNet can rapidly generate practical adversarial examples while achieving attack performance comparable to existing methods.

In summary, we summarize our contribution below:

- We reformulate adversarial attacks on cross-modal hashing retrieval from an iterative, instance-specific optimization problem into a relevance-guided generative learning paradigm.
- We evaluate the proposed framework on multiple cross-modal hashing retrieval systems using two widely adopted benchmark datasets: NUS-WIDE [29] and FLICKR-25K [28]. The experimental results demonstrate that our approach achieves competitive or superior attack performance while being significantly more efficient.

The rest of this paper is organized as follows: Section 2 presents the related work of CMHR and its research on adversarial attacks in the iterative and generative-based method. Section 3 provides background on cross-modal retrieval systems. Section 4 describes the problem description of the attack process on the CMHR and shows the solution of ReGeNet. Section 5 reports the experimental dataset, implementation parameters, experimental results, and discussion. Section 6 conducts the conclusions and future work.

2. Related Work

2.1. Deep Cross-Modal Hamming Retrieval

Cross-modal retrieval aims to support semantic matching across heterogeneous modalities, allowing queries in one modality, such as images or texts, to retrieve relevant content from another. However, as multi-modal datasets continue to grow in scale and complexity, approaches based on manually designed features exhibit limited efficiency and scalability in practical retrieval scenarios. Consequently, existing research has shifted toward deep neural network-based cross-modal retrieval frameworks. Recent studies [30,31] investigate deep architectures that learn mappings from multi-modal feature spaces to a shared Hamming space, demonstrating the superiority of deep models in retrieval tasks. While substantial progress has been made in single-modal retrieval, further investigation is required for cross-modal scenarios. A key challenge in CMHR systems lies in projecting heterogeneous modal features into a unified representation space. To address this challenge, ref. [5] proposed a deep hybrid architecture for visual–semantic fusion that learns a joint embedding space for images and text. Jiang et al. [6] introduced compact hash code generation to mitigate precision loss caused by relaxation, thereby improving retrieval accuracy. Cao et al. [7] designed a novel cross-modal deep hashing approach with a pairwise focal loss based on an exponential distribution. Su et al. [32] constructed a joint semantic matrix to integrate information across modalities and capture latent semantic similarity, while Chun et al. [33] further advanced cross-modal retrieval by modeling one-to-many latent relationships through probabilistic embeddings.

2.2. Adversarial Attack on Deep Hash Retrieval

The adoption of DNNs has substantially improved the retrieval performance of deep hashing models. Recent work [17,19,20,34] has extensively investigated adversarial example (AE) generation in single-modal hashing retrieval, primarily by optimizing Hamming distance-related objectives. These methods typically minimize the Hamming distance

between the hash codes of adversarial examples and predefined target codes to achieve effective attacks. In CMHR systems, however, AE generation must not only manipulate Hamming distances but also disrupt the latent cross-modal feature representations. Under white-box settings, Li et al. [21] first generated adversarial examples for CMHR by minimizing intra-modal similarity while maximizing inter-modal similarity, and Li et al. [22] further decoupled cross-modal data into modality-dependent and modality-irrelevant components to craft adversarial samples. In black-box settings, Li et al. [35] proposed a triplet-based iterative optimization strategy to generate transferable adversarial examples. Although black-box attacks are more aligned with real-world scenarios, there remains significant room for improving attack effectiveness under white-box settings. As discussed earlier, iterative-based approaches are computationally expensive and time-consuming. To the best of our knowledge, generative-based adversarial attack methods [20] have thus far been explored only in single-modal hashing retrieval, leaving cross-modal scenarios largely unexplored.

3. Background

3.1. Cross-Modal Hash Retrieval System

Compared with traditional feature extraction approaches—such as spectrum-based methods [36], color histogram-based methods [37], and bag-of-words (BoW) representations [38]—recent deep neural network-based (DNN-based) CMHR systems [5–7,39] benefit substantially from their powerful representation learning capabilities. Given cross-modal inputs such as images or text, a DNN-based CMHR system encodes them into feature vectors using carefully designed neural networks and retrieves relevant results by measuring similarity in the learned feature space. This paradigm eliminates the need for manual feature engineering and instead leverages large-scale data-driven training to learn optimal network parameters, thereby enabling high-performance retrieval in an automated manner. Moreover, deep hashing-based retrieval has become a core technique in CMHR systems for further improving efficiency. In particular, deep hash retrieval frameworks built upon the approximate nearest neighbor (ANN) search [40] are widely adopted due to their low storage requirements and efficient retrieval performance. ANN-based methods learn mappings from high-dimensional feature spaces to compact hash spaces by optimizing semantic similarity, allowing well-trained DNNs to capture semantic correlations and establish meaningful relationships between image and text data. A typical DNN-based cross-modal retrieval framework is illustrated in Figure 1.

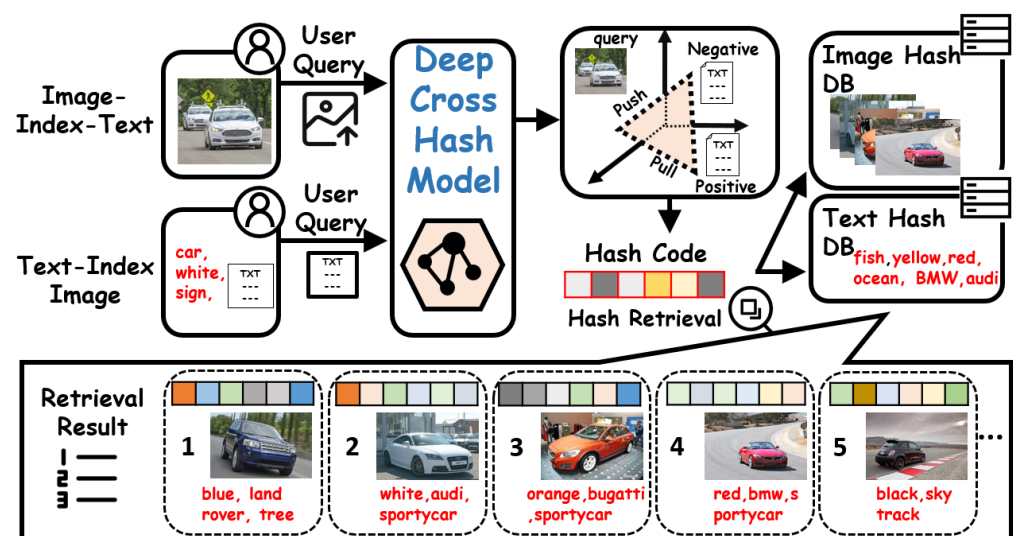


Figure 1. Illustration of the cross-modal retrieval network in the intelligent transportation system.

Leveraging a shared feature space across modalities, CMHR systems generate compact hash codes to support efficient retrieval. Existing DNN-based retrieval approaches can be broadly categorized into unsupervised learning frameworks [30,41] and supervised learning frameworks [31,42,43]. Supervised methods incorporate semantic labels to alleviate the semantic gap inherent in high-dimensional features, thereby improving retrieval accuracy. In contrast, unsupervised methods are applicable to a wider range of scenarios due to the absence of label constraints but require carefully designed high-level semantic representations to mitigate the semantic gap. The training process of a CMHR model typically consists of three stages:

- Embedding the vector of collected and semantic similar cross-modal data pairs.
- Constructing the optimization loss function to enhance the semantic and Hamming feature similarity of data pairs in hash space.
- Hashing all the cross-modal data and storing them in the database for retrieval.

Consider a cross-modal dataset $\mathcal{O} = \{o_i\}_{i=1}^M$, where each sample $o_i = \{o_i^v, o_i^t\}$ consists of paired image and text instances sharing semantic relevance. A deep cross-modal hashing retrieval (CMHR) model is designed to learn modality-specific mapping functions, $\mathcal{F}^*(\cdot)$ with $*$ $\in \{v, t\}$, such that heterogeneous inputs are embedded into a common hash space while preserving cross-modal semantic consistency. The latent output of each mapping function is a continuous representation $\mathcal{H}^*$ produced by the target model. Binary hash codes $\mathcal{B}^*$ are subsequently obtained by applying the sign operator to $\mathcal{H}^*$, yielding compact representations for efficient retrieval. Accordingly, the overall embedding and hashing process can be formulated as follows:

$$\mathcal{H}^* = \mathcal{F}^*(o^*) \quad \text{s.t.} \quad \mathcal{H}^* \in [-1, 1]^K \quad (1)$$

where K is the hash codes length, and $*$ $\in \{v, t\}$ denote the image and text data modality, and $\mathcal{H}$ is the binary-like hash code. We can obtain the binary hash code $\mathcal{B}^*$ via the sign function, i.e., $\mathcal{B}^* = \text{sign}(\mathcal{H}^*)$.

3.2. Problem Formulation

In this work, we consider a white-box adversarial setting in which the attacker has full access to the target cross-modal hashing model. An attack is deemed successful if it significantly degrades retrieval performance, as measured by a reduction in mean average precision (MAP), while satisfying a perceptual constraint (PER) to ensure imperceptibility. Unlike iterative optimization-based attacks that generate instance-specific perturbations, ReGeNet learns a relevance-guided generative mapping to directly produce adversarial samples, enabling efficient and structured disruption of retrieval rankings. Suppose a trained cross-modal hashing retrieval model $\mathcal{F}(\cdot)$ embeds heterogeneous data from different modalities into a shared hash space while preserving semantic consistency. Given a query sample o_q , the model is expected to produce a hash code that is close to semantically related samples (e.g., o_p) and distant from semantically irrelevant ones (e.g., o_n). This behavior reflects the model's objective of maintaining cross-modal semantic alignment in the retrieval space.

Under this formulation, the learning process of the CMHR model can be interpreted as optimizing its parameters δ_F to preserve relative semantic relationships among cross-modal data points. Accordingly, the training objective can be expressed as

$$\underset{\delta_F}{\operatorname{argmin}} \operatorname{Ham}(\mathcal{F}^v(o_q^v), \mathcal{F}^t(o_p^t)) - \operatorname{Ham}(\mathcal{F}^v(o_q^v), \mathcal{F}^t(o_n^t)) + \varphi \quad (2)$$

Here, φ denotes a margin parameter, which is commonly set to $\frac{K}{2}$. The function $\text{Ham}(\cdot, \cdot)$ is employed to quantify dissimilarity in the hash space and can be expressed as $\text{Ham} = \frac{1}{2}(K - \langle V, T \rangle)$, where $\langle \cdot, \cdot \rangle$ represents the inner product between hash representations. Based on the image–text triplet $\{o_q, o_p, o_n\}$, adversarial perturbations are introduced to alter relative distances among hash codes under imperceptibility constraints, thereby revealing how retrieval behavior degrades when semantic alignment is disrupted.

4. Framework

4.1. Overview

This paper investigates a supervised cross-modal generative framework termed ReGeNet. The proposed relevance-guided generative network comprises three sub-networks: parallel adversarial generative networks $\mathcal{G}$ and $\mathcal{D}$ for adversarial example (AE) generation, a label network $\mathcal{L}$ for learning latent adversarial label representations, and a relevance-guided graph convolution network $\mathcal{P}$ for modeling cross-modal correlations during training. Given a training cross-modal data pair, ReGeNet learns to invert latent representations around the decision boundary of the target hashing model to generate corresponding adversarial examples. Specifically, the input data pair is first fed into the parallel generator $\mathcal{G}$ to produce adversarial samples, whose hash codes are then obtained through the target cross-modal hashing model. The parallel discriminator $\mathcal{D}$ is employed to encourage imperceptible perturbations and effective decision boundary inversion. Meanwhile, cross-modal labels are processed by the label network $\mathcal{L}$ to embed misleading semantic information into the feature space. The semantic representations from $\mathcal{L}$ and the latent features learned by $\mathcal{G}$ are jointly integrated and iteratively refined within the graph convolution network $\mathcal{P}$ to facilitate relevance-aware feature mining. During training, we design three loss functions—a relevance-guided adversarial loss, a reconstruction loss, and a quantization loss—to jointly optimize the three sub-networks until convergence. During inference, the well-trained adversarial generator efficiently produces adversarial examples. The overall architecture of ReGeNet is illustrated in Figure 2, and detailed formulations and notations are provided in Table 1.

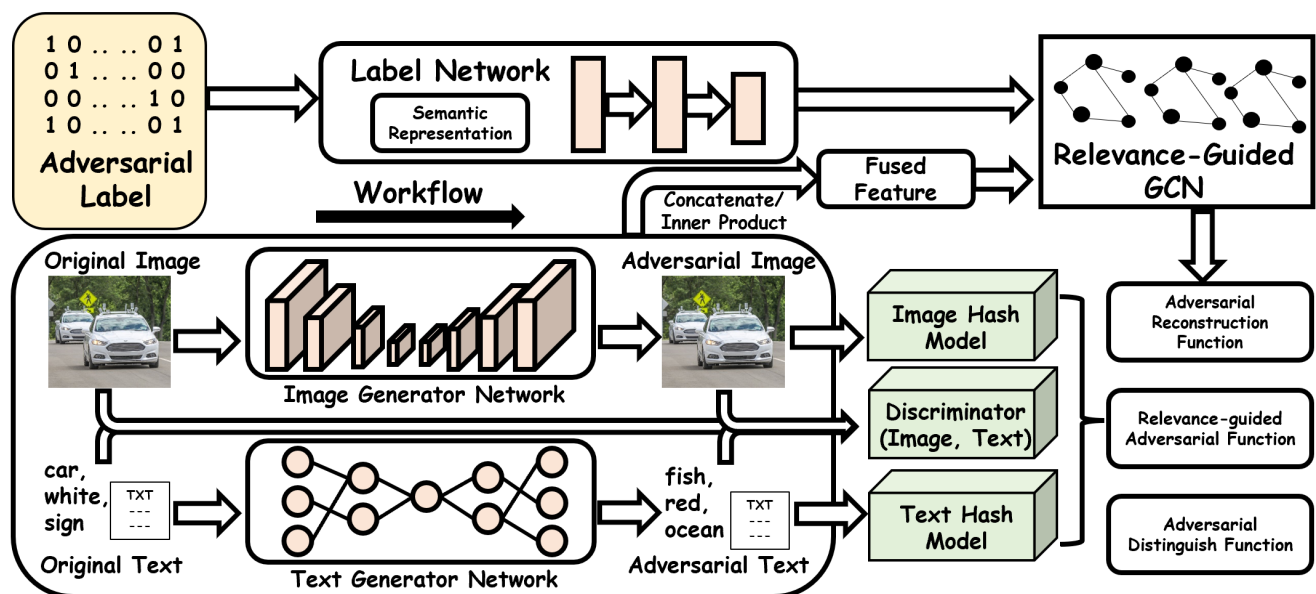


Figure 2. The framework illustrates our parallel generative network (ReGeNet).

Table 1. Symbols and definitions.

Symbol	Definition
$\mathcal{F}^*$	The cross-modal retrieval function
$\mathbf{o}^*$	The cross-modal data points
$\mathcal{H}^*$	The embedding vector of data points
$\mathcal{B}^*$	The hash codes of cross-modal data points
$\mathcal{G}^*$	The parallel generative network
$\mathcal{D}^*$	The parallel discriminator network
$\mathcal{P}$	The relevance-guided graph convolution network
$\mathcal{L}$	The label network
K	The length of hash codes
$\mathcal{J}_{\text{adv}}$	The adversarial loss function
$\mathcal{J}_{\text{rec}}$	The reconstruction loss function
$\mathcal{J}_{\text{dis}}$	The distinguish loss function

4.2. Cross-Modal Adversarial Generative Network $\mathcal{G}, \mathcal{D}$

As mentioned above, cross-modal data retrieval is subject to minimizing the cross-modal data pairs with similar semantic features. We explore the parallel cross-modal generator network $\mathcal{G}$ with the supervised information to generate the AEs, which mainly contain the image and text encoder–decoder. The latent feature of the parallel generator network $\mathcal{G}$ is fused as $\mathcal{G}_{fe}$ into the relevance-guided module $\mathcal{P}$. In practice, we follow the work [44] that the output of the encoder is used to integrate the latent feature of the cross-modal generator via the concatenate or inner product. Moreover, we follow the work [20] to exploit the skip connection flow by concatenating the original cross-modal layer and the last output layer in $\mathcal{G}^*$. Hence, the parallel adversarial generator $\mathcal{G}$ is encouraged to maximize the Hamming distance between the original and AEs in the hash code space, which needs to achieve two goals: (1) The hash code of AEs and original cross-modal data have a large Hamming distance. (2) The generated AEs are invisible from the original cross-modal data. Moreover, similar to the general adversarial generative network, we set the discriminator $\mathcal{D}$, which is designed to distinguish the visible differences between the adversarial data $\mathbf{o}'$ and the original data $\mathbf{o}$.

4.3. Adversarial Supervised Adversarial Label Network $\mathcal{L}$

Unlike the adversarial attack in the classification, the deep hash retrieval models aim to preserve the semantic feature in the hash codes rather than only directly using the categories label. Hence, especially in cross-modal data hash retrieval, we can construct the label network with supervised adversarial information to guide the generation of AEs. For instance, given the training batch of cross-modal data $\{\mathbf{o}\}^M$ with its label, we can construct the similarity matrix $\mathcal{S}$, where $\mathcal{S}_{i,j}$ denotes the semantic relevance in the cross-modal data batch, and $\mathcal{S}_{i,j} = 1$ denotes that the $\mathbf{o}_i$ and $\mathbf{o}_j$ are semantically similar; otherwise, $\mathcal{S}_{i,j} = 0$. For the AEs attack, we construct the adversarial similarity matrix $\hat{\mathcal{S}}$ based on the original similarity matrix $\mathcal{S}$. In particular, the adversarial similarity matrix can be easily extended into untargeted and targeted attacks. Since this paper discusses the untargeted paper, we randomly perturb the original similarity matrix $\mathcal{S}$ in each input cross-modal data batch. Moreover, to achieve more effective mining of supervised label information, we exploit the label network $\mathcal{L}$ to embed the adversarial similarity matrix $\hat{\mathcal{S}}$ into the latent semantic feature $\mathcal{L}_{fe}$. In particular, we follow the work of [20] to build the workflow from the latent semantic feature of $\mathcal{L}$ to the adversarial generator $\mathcal{G}$ via the upsample operation to encourage network convergence.

4.4. Relevance-Guided-Based Graph Convolution Network $\mathcal{P}$

To explore the potential feature information of supervised label network $\mathcal{L}$ and parallel adversarial generator $\mathcal{G}^*$ as much as possible, we exploit the graph convolution network (GCN) to implement semantic feature union for the above-two latent features $\mathcal{L}_{fe}$ and $\mathcal{G}_{fe}$. Hence, given the cross-modal data pair o_i and o_j , we follow the work of [39] to compute the similarity adjacency matrix $\mathcal{A}$ of the latent feature of label network $\mathcal{L}_{fe}$ with the cosine similarity as follows:

$$\mathcal{A}_{i,j} = \frac{L_{fe}^i \cdot L_{fe}^j}{\|L_{fe}^i\| \times \|L_{fe}^j\|} \quad (3)$$

where $\mathcal{A} \in \mathcal{R}^{n \times n}$. On the other hand, we exploit the multi-layer GCN structure to mine the correlations between $\mathcal{L}_{fe}$ and $\mathcal{G}_{fe}$. We implement the latent feature of adversarial generator $\mathcal{G}_{fe}$ into the first GCN layer $\mathcal{H}^0$, and the propagation rule can be formulated as

$$\mathcal{H}^{(l+1)} = \nu(\hat{\mathcal{D}}^{-\frac{1}{2}} \mathcal{A} \hat{\mathcal{D}}^{-\frac{1}{2}} \mathcal{H}^{(l)} \mathcal{W}^{(l)}) \quad (4)$$

where $\mathcal{H}^{(l)}$ denotes the l -th GCN layer, and the first layer input of the GCN is $\mathcal{G}_{fe}$. $\nu(\cdot)$ denotes the Relu activation function. $\hat{\mathcal{A}} = \mathcal{S} + \mathcal{I}_n$, and $\mathcal{I}_n$ refers to the unit matrix. $\hat{\mathcal{D}}$ is the degree matrix of $\hat{\mathcal{A}}$. $\mathcal{W}^{(l)}$ denotes the parameter of the l -th GCN layer.

Mechanism of the relevance-guided GCN: It is important to emphasize that the proposed relevance-guided GCN does not aim to learn retrieval embeddings or improve retrieval accuracy. Instead, it serves as an adversarial relevance propagation module that injects structured semantic perturbations into the adversarial generation process. By propagating adversarial relevance information across semantically related samples, the GCN encourages structurally consistent perturbations within local semantic neighborhoods. This mechanism tends to induce a smoother yet more disruptive deformation of the learned Hamming space, which can amplify retrieval ranking degradation while mitigating isolated or unstable adversarial effects.

4.5. Learning to Adversarial Hash

Regarding the mentioned sub-network $\mathcal{G}$, $\mathcal{L}$, $\mathcal{P}$, and $\mathcal{D}$ with its network parameters θ_g , θ_t , θ_p , and θ_d , respectively, the optimization goal of our proposed ReGeNet is to guarantee the attacking quality of the AEs generated by $\mathcal{G}$ guided by $\mathcal{L}$ and $\mathcal{P}$, so we set the training function for the whole ReGeNet as follows:

$$\underset{\delta_g, \delta_l, \delta_p, \delta_d}{\operatorname{argmin}} \sum \mathcal{J}_{rel} + \gamma \mathcal{J}_{rec} + \varphi \mathcal{J}_{dis} \quad (5)$$

γ and φ are the weighting constants used to balance the term in Equation (5). $\mathcal{J}_{rel}$ is the relevance-guided adversarial loss function, $\mathcal{J}_{rec}$ is the reconstruction loss function, and $\mathcal{J}_{dis}$ is the adversarial distinguish loss function. Next, we introduce the above-three loss functions below respectively.

Relevance-guided adversarial loss function $\mathcal{J}_{adv}$ aims to maximize the Hamming distance of the hash code from the original space. To solve the above, we aim to minimize the feature distance between the supervised adversarial label information and the adversarial cross-modal data. At the same time, we maximize the Hamming distance between the cross-modal data to enhance the untargeted attack performance. We formulate the above goal to obtain $\mathcal{J}_{adv}$ as follows:

$$\mathcal{J}_{rel} = \frac{1}{M^2} \left(\sum_{i,j=1}^M \log(1 + e^{\Omega_{i,j}}) + \mathcal{S}_{i,j} \Omega_{i,j} \right) + \sum_{i,j=1}^M (\log(1 + e^{-Y_{i,j}}) + \mathcal{S}_{i,j} Y_{i,j}) \quad (6)$$

where $\mathcal{S}$ denotes the adversarial similarity matrix, and $\mathcal{S}$ denotes the data label similarity matrix. $\Omega_{i,j} = \frac{1}{2}(\mathcal{B}_i^*)(\mathcal{H}_j)^T$, where $\mathcal{H}$ is the output of the correlation from the last layer of the GCN. $Y_{i,j} = \frac{1}{2}(\mathcal{B}_i^v)(\mathcal{B}_j^t)^T$, where $\mathcal{B}^*, * \in \{v, t\}$ defines the adversarial cross-modal data. The relevance-guided adversarial loss $\mathcal{J}_{rel}$ is designed to explicitly enlarge the Hamming distance gap between adversarial samples and their original counterparts rather than merely induce misclassification. By maximizing the disagreement with the original similarity matrix while aligning with the adversarial similarity matrix, $\mathcal{J}_{rel}$ perturbs the relative ordering of retrieved items in the Hamming space. As a result, semantically relevant items are pushed farther away in the retrieval ranking, which directly degrades cross-modal retrieval performance instead of classification accuracy.

Adversarial reconstruction loss $\mathcal{J}_{rec}$: To regulate the magnitude of adversarial perturbations while preserving retrieval-related representations, we introduce a reconstruction-based objective. This loss encourages adversarial samples to remain close to their original counterparts in the input space while simultaneously aligning their latent representations with the corresponding hash codes produced by the target model. As a result, the generated adversarial examples can effectively influence retrieval behavior without introducing excessive perceptual distortion.

The reconstruction loss is formulated as

$$\mathcal{J}_{rec} = \|o - \mathcal{G}(o)\|_2^2 + \alpha_1 \|\mathcal{T}(\mathcal{G}(o)) - \mathcal{B}\|_2^2, \quad (7)$$

where $\mathcal{G}(o)$ represents the adversarial sample generated from the original cross-modal input o . The ℓ_2 norm is adopted to measure reconstruction deviation and ensure stable gradient-based optimization of the generator parameters. $\mathcal{B} = \text{sign}(\mathcal{H})$ denotes the binary hash code derived from the latent representation. Here, $\mathcal{G}$ refers to modality-specific generators, and the modality index is omitted for clarity. The reconstruction loss $\mathcal{J}_{rec}$ constrains the magnitude of adversarial perturbations by penalizing deviations from the original input, ensuring imperceptibility while maintaining hash stability. The quantization-related term further regularizes the generated samples to remain close to valid hash representations, preventing excessive distortion of the Hamming space and stabilizing the adversarial training process.

The adversarial distinguish function $\mathcal{J}_{dis}$ is employed to improve the adversarial cross-model data from the original cross-modal data and encourage them into the targeted label feature space. We follow the general distinguish function in the general GAN [45] to exploit the loss function $\mathcal{J}_{dis}$ as follows:

$$\mathcal{J}_{dis} = \frac{1}{2} (\|\mathcal{D}(o) - [y, 0]\|_2^2 + \|\mathcal{D}(o') - [y_{adv}, 1]\|_2^2) \quad (8)$$

where y and y_{adv} are the original and adversarial data labels with the one-hot format, and $[\cdot, \cdot]$ denotes the concatenate operation. $\|\cdot\|_2$ denotes the ℓ_2 norm. We summarize the training algorithm in Algorithm 1. The adversarial distinguish loss $\mathcal{J}_{dis}$ complements the above objectives by encouraging the generator to produce realistic adversarial samples that are difficult to distinguish from original inputs, thereby supporting stable and effective adversarial generation.

Algorithm 1: The training process of relevance-guided generative network.

Input: Cross-modal data set $\{(o_i^*) | * \in \{v, t\}\}^M$, a pre-trained target cross-modal hash model $\mathcal{T}^*$, data label set $\{(l_i)\}^M$;
Output: Adversarial cross-modal examples v' ;
Initialize parameter ;
while *not converge* **do**
 Given the cross-modal data batch $\{(o_i)\}^n$ and the adversarial label $\{(l_i)\}^n$;
 Calculate the objective function via Equation (5);
 Update δ_g via the gradient descent;
 Update δ_l via the gradient descent;
 Update δ_p via the gradient descent;
 Update both discriminator parameter of two modality δ_d via the gradient ascent;
end

5. Experiment

In this section, we describe the experimental settings and evaluation protocol adopted to validate the proposed ReGeNet framework. Retrieval effectiveness under adversarial perturbations is assessed using Mean Average Precision (MAP), while the imperceptibility of generated adversarial examples is quantified through Perceptibility (PER). In addition, we report the computational cost associated with adversarial example generation to analyze the efficiency of ReGeNet.

5.1. Dataset

In our experiments, we evaluate our method on two popular datasets, NUS-WIDE [29] and FLICKR-25K [28], as summarized in Table 2. One dataset is NUS-WIDE, which consists of 190,421 image–text pairs in 21 categories divided into retrieval, training, and query parts with the numbers 183,321, 5000, and 2100. The other one is FLICKR-25K, which contains 20,015 image–text pairs in 24 classes. Similarly, it is divided into sections 14,015, 5000, and 1000. We utilize the query cross-modal dataset for the adversarial attacks in the two datasets, respectively.

Table 2. Data statistics of two datasets.

Datasets	Train	Query	Retrieval
FLICKER-25K	5000	1000	14,015
NUSWIDE	5000	2100	183,321

5.2. Evaluation Metrics

To evaluate the performance of our framework, we follow the work of [35] to set the Mean Average Precision (MAP). The Average Precision (AP) can be calculated as follows:

$$AP = \frac{1}{N} \times \sum_{i=1}^N \frac{i}{pos(i)}$$

where N is the size of the retrieval list. $pos(i)$ indicates the position of the i -th correct retrieval result in the returned list. Hence, the Mean Average Precision (MAP) is the mean value of the average accuracy (AP) of multiple retrieval results, and it is the widely used criterion in cross-modal Hamming retrieval. The lower the MAP value, the stronger the

attack performance. Moreover, we follow the work of [8] to measure the cross-modal data difference between adversarial and query data by calculating PER as

$$\text{PER} = \sqrt{\frac{1}{M} \|o - o'\|_2^2}$$

where o defines the original cross-modal data, and o' defines the adversarial cross-modal data. M is the total number of dimensions of cross-modal data, i.e., the total pixel number of the image. Similarly, the lower the value of PER, the better the imperceptibility of the AEs.

5.3. Implementation

All experiments were conducted using 32-bit hash codes on a server equipped with two GeForce RTX 2080 Ti GPUs. We considered two cross-modal retrieval tasks: image-to-text and text-to-image. For the image-to-text task, we adopted the architecture proposed in [46] to construct the image adversarial generator. For the text-to-image task, we employed a multi-layer perceptron to build the text adversarial generator. The text discriminator was implemented as a fully connected network, while the image discriminator consisted of two five-stride convolutional layers. We used the Adam optimizer [27] with an initial learning rate of 10^{-2} to solve Equation (5). ReGeNet was trained for 50 epochs, and the hyperparameter α_1 in Equation (7) was set to 10^{-1} . The weighting parameters γ and ϕ were chosen following common practice in the adversarial hashing and generative attack literature, where similar balancing strategies are used to trade-off attack effectiveness and perturbation constraints. We empirically observed stable training behavior over a reasonable range of these hyperparameters, with no mode collapse or oscillatory behavior during optimization. In all experiments, the generator converged consistently when trained with fixed auxiliary sub-networks, and no divergence was observed in either the loss values or retrieval performance. During training, we optimized the parameters δ_g of the adversarial generator $\mathcal{G}$ while keeping all other network parameters fixed. To ensure reproducibility, all experiments were conducted with fixed random seeds, and identical hyperparameter settings and training configurations were used across all evaluations.

5.4. Compared Methods

We employed two widely used CMHR systems, CMHH and DCMH [6], as target deep hashing models, with VGG19 [47] adopted as the image backbone. The cross-modal hashing models were trained using publicly available implementations of prior work until convergence. In addition, we evaluated two iterative-based cross-modal adversarial attack methods: CMLA [21] and DACM [22]. Each attack method was executed for 500 iterations, and the results are reported at iteration counts of 0, 200, and 500 to reflect the progressive attack behavior.

5.5. Results

MAP under different attack settings: Table 3 reports the Mean Average Precision (MAP) for both retrieval tasks, where $I \rightarrow T$ denotes retrieving text given images, and $T \rightarrow I$ denotes retrieving images given text. The Original setting corresponds to retrieval performance without attacks. We report the MAP values for the iterative-based methods CMLA and DACM, as well as for our proposed approach, on both datasets. Lower MAP values indicate stronger attack effectiveness. From Table 3, we draw the following observations. First, for both CMLA and DACM, the attack performance improved gradually as the number of iterations increased. For example, in the $I \rightarrow T$ task on FLICKR-25K, the MAP of DACM decreased from 71.73 to 50.57 after 200 iterations but dropped by only 0.81 after an additional 300 iterations. This highlights the limitation of instance-specific iterative attacks, which require substantial computational resources while yielding diminishing

performance gains. Second, across both retrieval tasks and datasets, our method achieved attack performance comparable to iterative-based approaches. For instance, on the NUS-WIDE dataset under the $I \rightarrow T$ task with CMHH, our method attained a MAP of 27.46%, whereas the strongest iterative-based method achieved 25.31%, representing a difference of 2.15%. Notably, our approach performed better on NUS-WIDE than on FLICKER-25K, which we attribute to the larger training set in NUS-WIDE that improved the generalization capability of the proposed network. We further report an ablation study analyzing the contribution of different sub-networks to attack performance in Table 4.

Table 3. MAP(%) of attack methods with 32-bit hash codes on two datasets. Best results in bold; second best underlined.

Task	Method	Iter.	FLICKER-25k		Method	Iter.	NUS-WIDE	
			CMHH	DCMH			CMHH	DCMH
$I \rightarrow T$	Original		71.73%	64.66%	Original		57.94%	49.32%
	CMLA	200	56.41%	51.78%	CMLA	200	33.47%	32.15%
		500	55.52%	51.24%		500	32.02%	31.61%
	DACM	200	50.57%	45.69%	DACM	200	26.61%	24.89%
		500	49.76%	44.55%		500	25.31%	23.90%
	Ours	1	<u>50.59%</u>	<u>46.18%</u>	Ours	1	<u>27.46%</u>	<u>24.31%</u>
$T \rightarrow I$	Original		74.08%	61.94%	Original		52.45%	45.76%
	CMLA	200	52.73%	54.02%	CMLA	200	35.68%	30.18%
		500	<u>51.45%</u>	<u>53.69%</u>		500	34.73%	29.55%
	DACM	200	48.28%	51.57%	DACM	200	27.45%	26.31%
		500	47.67%	50.42%		500	<u>26.39%</u>	25.73%
	Ours	1	51.51%	54.32%	Ours	1	24.36%	<u>28.47%</u>

Table 4. MAP(%) comparison with different sub-networks on FLICKER-25K under CMHH method.

Discussion	Settings	$I \rightarrow T$		$T \rightarrow I$	
		MAP	Per	MAP	Per
Original	$\mathcal{G}, \mathcal{D}, \mathcal{L}, \mathcal{P}$	50.59%	0.025	51.51%	0.011
Generative	$\mathcal{G}, \mathcal{D}, \mathcal{L}$	54.75%	0.031	53.92%	0.016
Generator	$\mathcal{G}, \mathcal{D}$	61.61%	0.023	55.60%	0.012

Running time and imperceptibility under different settings: Table 5 summarizes the running time and imperceptibility of the generated adversarial examples. An important objective is to improve AE generation speed while maintaining strong attack performance. Our method requires only a single forward pass through the well-trained adversarial generator to produce effective adversarial examples, which is substantially more efficient than iterative approaches that repeatedly perform backpropagation through the target model.

Table 5. Running time (min) and Perceptibility (PER) on the attack methods with the 32-bit hash code on two image datasets. We present the best results in bold, and the second best are underlined.

Tasks	Method	Iteration	FLICKER-25k		Method	Iteration	NUS-WIDE	
			Time	Per			Time	Per
$I \rightarrow T$	CMLA	100	<u>0.31</u>	0.056	CMLA	100	<u>0.42</u>	0.045
		500	1.27	0.021		500	1.87	<u>0.031</u>
	DACM	100	0.35	0.045	DACM	100	0.45	0.039
		500	1.31	<u>0.039</u>		500	1.96	0.025
	Ours	1	0.002	0.069	Ours	1	0.003	0.064

Table 5. Cont.

Tasks	Method	Iteration	FLICKER-25k		Method	Iteration	NUS-WIDE	
			Time	Per			Time	Per
T→I	CMLA	100	0.18	0.041	CMLA	100	0.14	0.029
		500	0.81	0.033		500	0.72	0.011
	DACM	100	0.16	0.058	DACM	100	0.11	0.035
		500	0.95	0.025		500	0.77	0.014
	Ours	1	0.003	0.036	Ours	1	0.003	0.054

5.6. Running Time and Imperceptibility

To better demonstrate the superiority of ReGeNet in terms of generation speed and imperceptibility of the AEs, we show in Table 5 the 32-bit hash code length on two datasets, as well as the PER and running speed of the DACM and ReGeNet. We can observe that ReGeNet had the fastest generation speed. We observed that the PER value of DACM decreased gradually with the increase in the number of iterations. This shows that the iterative attack method can generate enough AEs to be concealed. It is worth noting that ReGeNet could also produce high visual quality and covert AEs. We show the cross-modal search result of AEs in Figure 3 for two adversarial attack methods (i.e., iterative- and generative-based attack methods) and two categories of cross-modal retrieval tasks (i.e., image retrieve text and text retrieve images).

T → I	white, sporty car, audi (Example Input 1)	Original Search Result 	Adversarial Search Result 
	red, green, highway (Example Input 2)	Original Search Result 	Adversarial Search Result 
I → T	 (Example Input 1)	Original Search Result 1. sign, white, tree, ocean, brid 2. sign, white, tree, white, fin 3. sign, white, tree, sky, bird	Adversarial Search Result 1. fish, fin, red, blue, brid 2. fish, yellow, blue, brid, sky 3. fish, yellow, ocean, white, dog
	 (Example Input 2)	Original Search Result 1. green, tree, red, paper, brid 2. grass, green, tree, green, cat 3. red, tree, grass, door, dark	Adversarial Search Result 1. book, face, mouse, paper, hawk 2. dog, fish, urban, key, pilot 3. horse, car, zoo, night, fish

Figure 3. The illustration of cross-modal AEs on two retrieval tasks (i.e., Image Retrieval Text, Text Retrieval Image).

6. Conclusions

Adversarial example generation for multi-modal data can facilitate research on high-dimensional representation and visualization tasks. Prior work primarily relies on iterative adversarial attacks to generate adversarial examples against target models; however, such approaches are computationally inefficient and often impractical in real-world scenarios where target labels are unavailable. This paper investigates an adversarial training framework based on relevance-guided generative networks. Specifically, we designed parallel cross-modal image and text generators and introduced a label network to exploit adversarial similarity matrices. We further integrated the latent features learned by the parallel generators with the embedding features produced by the label network and optimized them jointly using a graph convolutional network (GCN). During training, the entire framework

was optimized using a correlation-guided adversarial loss, a reconstruction loss, and an adversarial discrimination loss. In future work, we plan to further enhance the adversarial strength of the generative network and explore the cross-domain transferability of adversarial generative models. Meanwhile, the analytical framework and empirical findings of this study provide transferable methodological insights for comprehensive assessment tasks involving large-scale multi-modal data. By emphasizing robust representation learning and reliable relevance modeling, this work helps inform the design of evaluation mechanisms that are more stable, objective, and interpretable. These explorations further establish a methodological foundation for continued advances in multi-modal data-driven comprehensive evaluation and related reform-oriented research contexts.

Author Contributions: Conceptualization, C.L. and R.S.; Methodology, C.L.; Validation, C.H., Y.Y. and J.H.; Formal analysis, C.H. and L.C.; Resources, Y.Y., L.C. and J.H.; Writing—review and editing, Y.C., C.L., R.S. and J.H.; Visualization, Y.C. and Y.L.; Supervision, C.L. and R.S.; Project administration, C.L.; Funding acquisition, Y.L., R.S. and J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Hunan Provincial “14th Five-Year Plan” Educational Science Research Base Project (No. XJK23AJD021), the Philosophy and Social Sciences Foundation of Hunan Province (No. 22YBA012), the National Natural Science Foundation of China (No. 62477046, 42471507), the Hunan Province Science and Technology Innovation Project (No. S2021GCZDYF1405), the National Key Research and Development Program of China (No. 2021YFC3340800), The Science and Technology Innovation Program of Hunan Province (No. 2025RC3012), Natural Science Foundation of Hunan Province (No. 2025JJ60229, 2023JJ40771), and the High Performance Computing Center of Central South University.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: The authors would like to thank the anonymous reviewers for their insightful comments.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Aradhya, H.R. Object detection and tracking using deep learning and artificial intelligence for video surveillance applications. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 517–530. [CrossRef]
2. Cao, Q.; Shen, L.; Xie, W.; Parkhi, O.M.; Zisserman, A. Vggface2: A dataset for recognising faces across pose and age. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; IEEE: New York, NY, USA, 2018; pp. 67–74.
3. Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4690–4699.
4. Narasimhan, M.; Rohrbach, A.; Darrell, T. CLIP-It! Language-Guided Video Summarization. In *Proceedings of the Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, Virtual, 6–14 December 2021*; Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2021; pp. 13988–14000.
5. Cao, Y.; Long, M.; Wang, J.; Yang, Q.; Yu, P.S. Deep visual-semantic hashing for cross-modal retrieval. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1445–1454.
6. Jiang, Q.Y.; Li, W.J. Deep cross-modal hashing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3232–3240.
7. Cao, Y.; Liu, B.; Long, M.; Wang, J. Cross-modal hamming hashing. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 202–218.

8. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv* **2013**, arXiv:1312.6199.
9. Shi, Y.; Wang, S.; Han, Y. Curls & whey: Boosting black-box adversarial attacks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6519–6527.
10. Guo, C.; Gardner, J.; You, Y.; Wilson, A.G.; Weinberger, K. Simple black-box adversarial attacks. In Proceedings of the International Conference on Machine Learning—PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 2484–2493.
11. Komkov, S.; Petiushko, A. Advhat: Real-world adversarial attack on arcface face id system. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; IEEE: New York, NY, USA, 2021; pp. 819–826.
12. Moosavi-Dezfooli, S.M.; Fawzi, A.; Frossard, P. Deepfool: A simple and accurate method to fool deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2574–2582.
13. Dusmanu, M.; Schonberger, J.L.; Sinha, S.N.; Pollefeys, M. Privacy-preserving image features via adversarial affine subspace embeddings. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14267–14277.
14. Che, Z.; Borji, A.; Zhai, G.; Ling, S.; Guo, G.; Callet, P.L. Adversarial attacks against deep saliency models. *arXiv* **2019**, arXiv:1904.01231. [[CrossRef](#)]
15. Ilyas, A.; Engstrom, L.; Athalye, A.; Lin, J. Black-box adversarial attacks with limited queries and information. In Proceedings of the International Conference on Machine Learning—PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 2137–2146.
16. Bhagoji, A.N.; He, W.; Li, B.; Song, D. Practical black-box attacks on deep neural networks using efficient query mechanisms. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 154–169.
17. Yang, E.; Liu, T.; Deng, C.; Tao, D. Adversarial examples for hamming space search. *IEEE Trans. Cybern.* **2018**, *50*, 1473–1484. [[CrossRef](#)] [[PubMed](#)]
18. Tolias, G.; Radenovic, F.; Chum, O. Targeted mismatch adversarial attack: Query with a flower to retrieve the tower. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)—ICCV’19, Seoul, Republic of Korea, 27 October–2 November 2019.
19. Bai, J.; Chen, B.; Li, Y.; Wu, D.; Guo, W.; Xia, S.; Yang, E. Targeted Attack for Deep Hashing based Retrieval. In *Computer Vision—ECCV 2020, Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020*; Springer: Berlin/Heidelberg, Germany, 2020.
20. Wang, X.; Zhang, Z.; Wu, B.; Shen, F.; Lu, G. Prototype-supervised adversarial network for targeted attack of deep hashing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16357–16366.
21. Li, C.; Gao, S.; Deng, C.; Xie, D.; Liu, W. Cross-modal learning with adversarial samples. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
22. Li, C.; Tang, H.; Deng, C.; Zhan, L.; Liu, W. Vulnerability vs. reliability: Disentangled adversarial examples for cross-modal learning. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, 6–10 July 2020; pp. 421–429.
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
24. Tang, Y.; Pino, J.; Li, X.; Wang, C.; Genzel, D. Improving speech translation by understanding and learning from the auxiliary text translation task. *arXiv* **2021**, arXiv:2107.05782. [[CrossRef](#)]
25. Li, S.; Neupane, A.; Paul, S.; Song, C.; Krishnamurthy, S.; Roy-Chowdhury, A.K.; Swami, A. Stealthy Adversarial Perturbations Against Real-Time Video Classification Systems. In Proceedings of the NDSS’19—Network and Distributed Systems Security (NDSS) Symposium, San Diego, CA, USA, 24–27 February 2019.
26. Hu, W.; Tan, Y. Generating adversarial malware examples for black-box attacks based on GAN. *arXiv* **2017**, arXiv:1702.05983. [[CrossRef](#)]
27. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
28. Huiskes, M.J.; Lew, M.S. The mir flickr retrieval evaluation. In Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, Vancouver, BC, Canada, 30–31 October 2008; pp. 39–43.
29. Chua, T.S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; Zheng, Y. Nus-wide: A real-world web image database from national university of singapore. In Proceedings of the ACM International Conference on Image and Video Retrieval, Fira, Greece, 8–10 July 2009; pp. 1–9.
30. Liu, X.; Mu, Y.; Zhang, D.; Lang, B.; Li, X. Large-scale unsupervised hashing with shared structure learning. *IEEE Trans. Cybern.* **2014**, *45*, 1811–1822. [[CrossRef](#)] [[PubMed](#)]

31. Yan, C.; Xie, H.; Yang, D.; Yin, J.; Zhang, Y.; Dai, Q. Supervised hash coding with deep neural network for environment perception of intelligent vehicles. *IEEE Trans. Intell. Transp. Syst.* **2017**, *19*, 284–295. [[CrossRef](#)]
32. Su, S.; Zhong, Z.; Zhang, C. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3027–3035.
33. Chun, S.; Oh, S.J.; De Rezende, R.S.; Kalantidis, Y.; Larlus, D. Probabilistic embeddings for cross-modal retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8415–8424.
34. Bai, J.; Chen, B.; Wu, D.; Zhang, C.; Xia, S. Universal Adversarial Head: Practical Protection against Video Data Leakage. In Proceedings of the ICML’21, Online, 18–24 July 2021.
35. Li, C.; Gao, S.; Deng, C.; Liu, W.; Huang, H. Adversarial Attack on Deep Cross-Modal Hamming Retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2218–2227.
36. Weiss, Y.; Torralba, A.; Fergus, R. Spectral hashing. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–10 December 2008; Volume 21.
37. Pushpalatha, K.R.; Chaitra, M.; Karegowda, A.G. Color Histogram based Image Retrieval—A Survey. *Int. J. Adv. Res. Comput. Sci.* **2013**, *4*, 119.
38. Zhang, Y.; Jin, R.; Zhou, Z.H. Understanding bag-of-words model: A statistical framework. *Int. J. Mach. Learn. Cybern.* **2010**, *1*, 43–52. [[CrossRef](#)]
39. Xu, R.; Li, C.; Yan, J.; Deng, C.; Liu, X. Graph Convolutional Network Hashing for Cross-Modal Retrieval. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19), Macao, China, 10–16 August 2019; Volume 2019, pp. 982–988.
40. Wang, J.; Zhang, T.; Sebe, N.; Shen, H.T. A survey on learning to hash. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 769–790. [[CrossRef](#)] [[PubMed](#)]
41. Qin, Q.; Huang, L.; Wei, Z.; Xie, K.; Zhang, W. Unsupervised deep multi-similarity hashing with semantic structure for image retrieval. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 2852–2865. [[CrossRef](#)]
42. Wang, X.; Shi, Y.; Kitani, K.M. Deep supervised hashing with triplet labels. In *Computer Vision—ACCV 2016, Proceedings of the 13th Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 70–84.
43. Liu, H.; Wang, R.; Shan, S.; Chen, X. Deep supervised hashing for fast image retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2064–2072.
44. Han, J.; Dong, X.; Zhang, R.; Chen, D.; Zhang, W.; Yu, N.; Luo, P.; Wang, X. Once a man: Towards multi-target attack via learning multi-target adversarial network once. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5158–5167.
45. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Proceedings of the 18th International Conference, Munich, Germany, 5–9 October 2015*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
46. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
47. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.