ATLANTIS: A Framework for Automated Targeted Language-guided Augmentation Training for Robust Image Search

Inderjeet Singh[1], Roman Vainshtein[1]
{inderjeet.singh,roman.vainshtein}@fujitsu.com
Alon Zolfi[2], Asaf Shabtai[2]
zolfi@post.bgu.ac.il,shabtaia@bgu.ac.il
Tu Bui[1], Jonathan Brokman[1], Omer Hofman[1]
{tu.bui,jonathan.brokman,omer.hofman}@fujitsu.com
Fumiyoshi Kasahara[3], Kentaro Tsuji[3], Hisashi Kojima[3]
{kasahara.f,tsuji.kentarou,hisashi.kojima}@fujitsu.com

[1] Fujitsu Research of Europe
United Kingdom
[2] Ben-Gurion University of the Negev
Israel
[3] Fujitsu Limited
Japan

Abstract

Recent image search or content-based image retrieval (CBIR) systems rely on deep metric learning (DML) for extracting representative image features; however, their generalisation is limited by the dependency on large volumes of high-quality, diverse and unbiased training data. We introduce ATLANTIS, a framework with a novel methodology that automatically identifies training data deficiencies and then performs targeted and controlled synthetic data augmentation. Our framework comprises a Data Insight Generator for extracting contextual insights and the deficiencies from the existing training data, an Augmentation Protocol Selector to define dynamic, context-aware augmentation strategies, and an Outlier Removal and Diversity Control module to control the synthetic data's semantic coherence and diversity. ATLANTIS leverages image-to-text transformations, large language models, and text-to-image synthesis to iteratively generate and refine synthetic data while ensuring alignment with the original data and augmenting training data diversity in a controlled manner. Our comprehensive empirical evaluations reveal that ATLANTIS surpasses state-of-art in challenging domain-scarce and class-imbalanced data scenarios while also enhancing adversarial robustness, thus underscoring the generalisation gains. ATLANTIS also sets new benchmarks in standard balanced DML tasks, thereby establishing it as a robust and scalable framework for CBIR.

1 Introduction

Content-Based Image Retrieval (CBIR), also called image search systems, increasingly rely on Deep Metric Learning (DML) to deliver meaningful representations for image retrieval, using distance metrics optimised by triplet, contrastive or angular losses [4, 15, 29, 35]. However, generalization remains a critical challenge, as DML models are not only prone to overfitting when trained or finetuned on limited datasets [21] but also more susceptible to adversarial attacks [38]. To address this, prior studies perform synthetic data augmentation with traditional transformations [25, 31] or GAN-based generative models [2, 3]; yet the former is unable to adjust the content distribution of the dataset while the latter is computationally expensive and has limited scalability.

We introduce ATLANTIS, a framework designed to enhance DML model generalizability via automatic and controllable synthetic data augmentation. ATLANTIS first identifies available training data insufficiencies then generated tailored synthetic data for the subsequent DML model training. Uniquely, data insufficiencies are analysed in the text space to leverage large language models (LLMs) as virtual agents, while also leveraging state-of-art image-to-text models and metadata analysis to detect class and domain imbalance. These data insights are then transformed into text prompts using LLMs, before being fed to text-to-image models for synthetic image generation. The whole pipeline is conducted automatically, with feedback loop from DML training and filtering mechanism being enforced after the synthesised images.

The contributions of ATLANTIS are four-folds:

- We introduce ATLANTIS, a unified framework that automatically identifies and augments context-dependent, potentially missing training information while also tracking weaknesses in training. This enables the efficient generation and augmentation of only relevant training data and patching weak classes during training, resulting in improved clean data performance and adversarial robustness in the trained CBIR models.

- To the best of our knowledge, ATLANTIS is the first to effectively leverages foundational language and vision models to enhance CBIR models, while also facilitating easy integration with existing methods.

- We present a suite of components to identify data requirements, define generation objectives, and filters the generated synthetic data while controlling data diversity.

- ATLANTIS significantly outperforms current state-of-art in domain-scarce, class imbalance, adversarial, and even on standard benchmarks [1].

2 Related Work

Traditional Image Transformations. Traditional image augmentation techniques have been instrumental in enhancing the performance of computer vision models by introducing variations through manually crafted transformations such as cropping, rotation, and flipping [25, 31]. However, Goodfellow *et al.* [11] and Grill *et al.* [12] highlight the limitations in performance gains when employing a non-selective augmentation strategy, indicating the need for more sophisticated methods.

Synthetic Data Augmentation. Generative approaches, particularly Generative Adversarial Networks (GANs), have been leveraged to create synthetic data for various tasks, including image-to-image translation and data augmentation for domain-specific applications [2, 3]. Pretrained language models extend these capabilities to text-based tasks, enabling sophisticated text augmentation processes [7, 39]. Automated augmentation strategies like AutoAugment and RandAugment have shown promise, they generally remain within the

[1] Source code and related data are available at `https://github.com/intherejeet/ATLANTIS`.

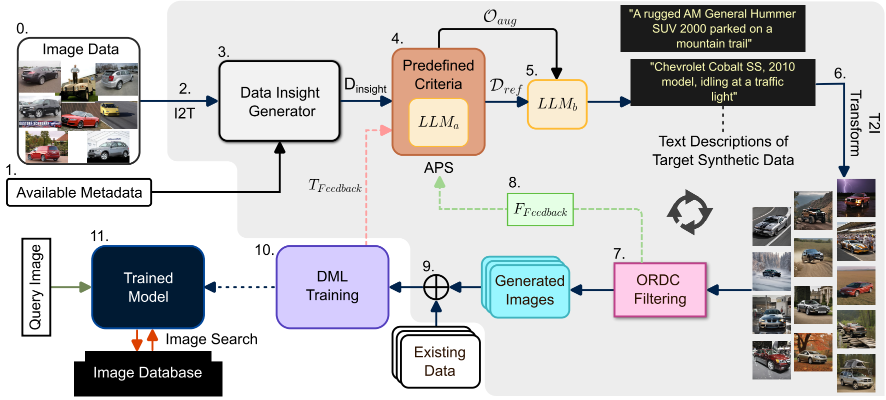

Figure 1: An overview of the ATLANTIS outlining its key integrated components: DIG, APS, ORDC filtering, and the feedback mechanisms, resulting selective and controlled synthetic data augmentation for enhanced generalizability in CBIR systems.

confines of traditional augmentation paradigms and do not exploit the rich representational power of multimodal data [5, 6].

Multimodal data augmentation efforts, such as [10, 14], leverage the intra- and cross-modal relation between training samples to generate new augmentations at raw inputs [14] or feature level [10]. These methods often employ simple augmenting strategies such as retrieval-mixing, image interpolation and text concatenation mainly for robustness training instead of addressing the data bias and class imbalance problems. Recently, Yin *et al.* [41] uses LLM and image diffusion to produce richer synthetic data but still face limitations in *controllability*, *flexibility*, and *lack an automated mechanism*.

These observed limitations have led us to the development of ATLANTIS for CBIR. The absence of a multimodal solution that combines image-to-text, LLMs with prompt engineering, and text-to-image models for controlled and targeted synthetic data augmentation in training and finetuning is a significant gap that ATLANTIS aims to fill. ATLANTIS sets a new precedent as the first to address this in the CBIR domain, while tackling the critical issues: the lack of an automatic mechanism to identify and adapt to missing training data characteristics and the limited scalability resulting from the requirement to finetune text-to-text models. This, as highlighted by other domain solutions like those in [41], not only challenges the relevance of existing techniques for direct comparison but also positions ATLANTIS as a pioneering framework for CBIR.

3 ATLANTIS Framework

ATLANTIS, standing for **A**utomated **T**argeted **L**anguage-guided **A**ugme**N**tation **T**raining for Robust **I**mage **S**earch, introduces a novel multi-modal framework designed to refine the training process of state-of-art CBIR models. The overall framework is outlined in Fig.1. ATLANTIS has 5 main building blocks: (1) Data Insight Generator *DIG* (Fig.1, step 3) for image and metadata analysis in text space to discover available data's weaknesses; (2) Augmentation Protocol Selector *APS* (Fig.1, step 4) for defining objectives for synthetic data generation; (3) *Synthetic Data Generation* for synthesising text prompts (Fig.1, step 5) and images (Fig.1, step 6); (4) Outlier Removal and Diversity Control (*ORDC*) (Fig.1, step 7) for optimising synthetic data quality; and (5) *Filtering Feedback* ($F_{Feedback}$) (Fig.1, step 8) for

Algorithm 1 Hybrid Data Insight Generator (DIG) for the Input Image Data

```
Require: Original image data $\mathcal{X}_o$, $f_{\text{ID}}$, LLM, FrequencyEval, ContextMaker,
    DomainInfer, ClassInfer, NewClassSearch, InsightExtract.
Ensure: Data insights $D_{\text{insight}}$
$\mathcal{E} \leftarrow f_{\text{ID}}(\mathcal{X}_o)$,    $\mathcal{T}_{\text{clean}} \leftarrow$ Tokenize($\mathcal{E}$) − StopWords
$\mathcal{C} \leftarrow$ FrequencyEval(AnnotatePOS($\mathcal{T}_{\text{clean}}$))
$\mathcal{M} \leftarrow$ Metadata($\mathcal{X}_o$),    $\mathcal{R} \leftarrow$ ContextMaker($\mathcal{C}, \mathcal{M}, LLM$)
$\mathcal{I}_{\text{domain}} \leftarrow$ DomainInfer($\mathcal{C}, \mathcal{R}, LLM$),    $\mathcal{I}_{\text{class}} \leftarrow$ ClassInfer($\mathcal{C}, \mathcal{M}, LLM$)
$\mathcal{N} \leftarrow$ NewClassSearch($\mathcal{I}_{\text{class}}, LLM$)
$D_{\text{insight}} \leftarrow$ InsightExtract($\mathcal{M}, \mathcal{I}_{\text{domain}}, \mathcal{I}_{\text{class}}, \mathcal{N}$)
```

Algorithm 2 Augmentation Protocol Selector (APS)

```
Require: Data insights $D_{\text{insight}}$, Filtering Feedback $F_{\text{feedback}}$, Training Feedback $T_{\text{feedback}}$,
    LLM, AugmentationObjective, ReferenceDescription, feedback initiation parameter $n$, IncorporateFeedback.
Ensure: Augmentation objectives $\mathcal{O}_{\text{aug}}$, Reference descriptions $\mathcal{D}_{\text{ref}}$
$\mathcal{O}_{\text{aug}}, \mathcal{D}_{\text{ref}} \leftarrow \{\}, \{\}$
for each class $c$ in $\{\mathcal{C} \cup \mathcal{N} \subset D_{\text{insight}}\}$ do
    $\mathcal{O}_{\text{aug}}[c] \leftarrow$ AugmentationObjective($D_{\text{insight}}[c]$, LLM)
    $\mathcal{D}_{\text{ref}}[c] \leftarrow$ ReferenceDescription($c$, LLM)
if training steps$\%n$==0 then
    $\mathcal{O}_{\text{aug}} \leftarrow$ IncorporateFeedback($\mathcal{O}_{\text{aug}}, F_{\text{feedback}}, T_{\text{feedback}}$, LLM)
```

keeping filtered synthetic data consistent with APS's defined objectives for the augmentation. We also experiment adding training feedback ($T_{Feedback}$).

3.1 Data Insight Generator (DIG)

Given $\mathcal{X}_o$ the original DML training dataset, DIG processes the visual data and metadata of $\mathcal{X}_o$ in text form (algorithm 1). First, the visual features (denoted $\mathcal{E}$) are extracted through a pretrained image-to-text model $f_{\text{ID}}(\mathcal{X}_o)$. This phase leverages state-of-art LVMs [1, 18, 19] that approximately embody Sampling Theorem precepts, ensuring discrete textual representation maintains the visual domain's continuous semantic integrity [22, 30]. Subsequent stages include tokenization (Tokenize), cleaning ($\mathcal{T}_{clean}$) with StopWords filtering, and frequency analysis (FrequencyEval) of extracted informative tokens through AnnotatePOS, pinpointing prevalent terms $\mathcal{C}$. Data context $\mathcal{R}$ is crafted using ContextMaker employed with LLM-reasoning on $\mathcal{C}$ and existing metadata $\mathcal{M}$. Domain (see section 4.1) $\mathcal{I}_{domain}$ and object class $\mathcal{I}_{class}$ distributions are deduced via DomainInfer and ClassInfer functions defined with LLM-prompted reasoning, and tailored for the task context and objectives that can handle both labeled and unlabeled input scenarios. For novel class identification $\mathcal{N}$, LLM reason on $\mathcal{I}_{class}$ and $\mathcal{R}$ with the predefined function NewClassSearch.

Finally the synthesis of target data insights is performed (InsightExtract) on $\mathcal{M}$, $\mathcal{I}_{domain}$, $\mathcal{I}_{class}$, and $\mathcal{N}$ that includes contextual overview, further augmentation information, and identified data deficiencies through domain and class distribution analysis, to provide APS with actionable insights. Specific prompt instructions given to the LLM for each stages in algorithm 1 and DIG's example outputs are detailed in Sup.Mat.

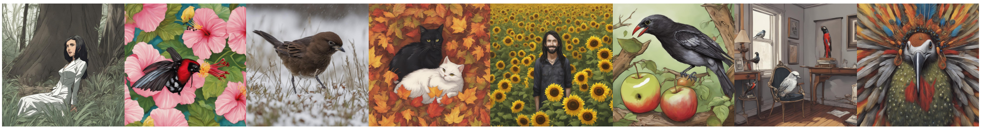

Figure 2: Examples of *noisy generations* (semantic noise) by the SDXL model [26] for the `'Chuck-Will's-Widow'` bird species. ORDC filtering ensures the removal of these kinds of outliers while controlling the diversity in the synthetic data for augmentation.

3.2 Augmentation Protocol Selector (APS)

APS defines the augmentation objectives for a downstream LLM responsible for generating text prompts for synthetic data generation. Leveraging the analytical and reasoning capabilities of foundational LLMs, the APS transforms the data insights and feedback alerts into actionable augmentation plans to guide the synthetic data generation process. As delineated in algorithm 2, the APS first processes the distilled insights D_{insight} from DIG to construct a class-wise set of augmentation objectives $\mathcal{O}_{\text{aug}}$ and reference descriptions $\mathcal{D}_{\text{ref}}$. These objectives are tailored to address the identified class imbalances $\mathcal{I}_{class}$, domain gaps $\mathcal{I}_{domain}$. Additionally, APS also identifies novel supplementable classes $\mathcal{N}$ to further enrich the generated synthetic data for increased diversity.

Finally, APS incorporates feedback alerts later in the DML model training or finetuning stages. This feedback (see Section 3.5) prompts the APS to update the augmentation objectives in order to adapt the synthetic data generation strategy to the evolving training needs. The predefined functions in algorithm 2 are primarily the defined system prompts for the used LLM; more details available in Sup.Mat.

3.3 Synthetic Data Generation

The Synthetic Data Generation is executed in two sequential phases: generation of text descriptions for the target synthetic data, and text-to-image synthesis.

Text Descriptions Generation. Given the set of enhancement objectives $\mathcal{O}_{\text{aug}}$ from the APS, an LLM as text descriptions generator $\mathcal{P}$ generates class-wise text descriptions $\mathcal{T}_{\text{desc}}$ using the set of reference class descriptions $\mathcal{D}_{\text{ref}}$ (Fig. 1, step 5). For each class $c \in \mathcal{C}$, the module constructs prompts $t_c \in \mathcal{T}_{\text{desc}}$ that are aligned with the augmentation strategy $\mathcal{O}_{\text{aug}}[c]$, converting the abstract augmentation goals into concrete linguistic constructs for the synthetic image data generation.

Image Synthesis. The image synthesis process is driven by a pretrained text-to-image model $\mathcal{G}$ [26], which maps 1-1 the generated text descriptions $\mathcal{T}_{\text{desc}}$ to a set of synthetic images $\mathcal{I}_{\text{synth}}$ (Fig.1, step 6). This mapping $\mathcal{G} : \mathcal{T}_{\text{desc}} \to \mathcal{I}_{\text{synth}}$ ensures that each synthetic image $i \in \mathcal{I}_{\text{synth}}$ visually embodies its text prompt $t \in \mathcal{T}_{\text{desc}}$, thereby systematically enriching the original dataset $\mathcal{X}_o$ with targeted synthetic instances through the APS's designed objectives in the *discrete text space*. This structured process guarantees that the augmented dataset is precisely tailored to the identified training data needs.

Algorithm 3 ORDC Filtering

```
Require: f: pretrained DML model; $\mathcal{X}_o$: original data; $\mathcal{L}_o$: original data labels; $\mathcal{X}_s$: synthetic data; $\mathcal{L}_s$: synthetic data labels; $\Delta$: diversity factor for synthetic data.
Ensure: Cleaned synthetic data $\mathcal{X}_{clean}$.
Initialize $\mathbf{C}, \mathbf{D} = \{\}, \{\}$
for each label $l$ in $\mathcal{L}_o$ do
    $n_l \leftarrow$ number of samples with label $l$ in $\mathcal{X}_o$
    $\mathbf{C}[l] \leftarrow \frac{1}{n_l} \sum_{i:\mathcal{L}_o[i]=l} \mathbf{f}(\mathcal{X}_o[i]), \quad \mathbf{D}[l] \leftarrow \frac{1}{n_l} \sum_{i:\mathcal{L}_o[i]=l} \mathrm{dist}(\mathbf{f}(\mathcal{X}_o[i]), \mathbf{C}[l])$
Initialize $\mathcal{X}_{clean}$ to an empty set
for each synthetic sample $\mathcal{X}_s[i]$ do
    $l \leftarrow \mathcal{L}_s[i]$
    if $\mathrm{dist}(\mathbf{f}(\mathcal{X}_s[i]), \mathbf{C}[l]) \leq \mathbf{D}[l] \times \Delta$ then
        Add $\mathcal{X}_s[i]$ to $\mathcal{X}_{clean}$
```

3.4 Outlier Removal and Diversity Control (ORDC) Filtering

The goal of ORDC filtering is to refine generated synthetic data by removing outliers and controlling data diversity. This filtering process is crucial for removing unintended noise introduced by the limitations of the text-to-image models and for curating the most beneficial subset for the CBIR model being trained. Inspired from the reduced-reference image quality assessment techniques [33, 40] and robust feature matching [13, 32, 36], our ORDC leverages pretrained DML models to extract embeddings that encapsulate content-based features independent of image alignment. Specifically, as delineated in algorithm 3, we employs $\mathbf{f}$ (finetuned DINO/ViT-S [9]) to project the original and synthetic images, $\mathcal{X}_o$ and $\mathcal{X}_s$ to the embedding spaces $\mathbf{f}(\mathcal{X}_o)$ and $\mathbf{f}(\mathcal{X}_s)$, respectively.

Next, we calculates the class centroids $\mathbf{C}[l]$ and average intra-class distances $\mathbf{D}[l]$ for each class label $l \in \mathcal{L}_o$. We then filter $\mathcal{X}_s$ whose embedding distance to $\mathbf{C}[l]$ is within $\mathbf{D}[l] \times \Delta$ where Δ is a diversity factor. This ensures the synthetic images have similar distribution as the original images. By adjusting Δ, ORDC can accommodate varying levels of data complexity, rendering it effective across diverse application domains. Examples of noisy images filtered out by ORDC are shown in fig. 2.

3.5 Feedbacks and Training

The Filtering Feedback module guarantees that synthetic data remains consistent with the APS's defined augmentation objectives after the ORDC filtering stage. It has predefined exceptions and criteria for alerting APS to supplement filtering stage losses (see details in Sup.Mat.). We also test a training feedback mechanism ($T_{Feedback}$) that, after a predefined number of training iterations, assesses the training classes with poor retrieval performance and then alerts APS to generate additional data for them (details in Sup.Mat.). In the final phase, the original training set is augmented with the refined synthetic data $\{\mathcal{X}_o, \mathcal{X}_{clean}, \mathcal{X}_{ff}, \mathcal{X}_{tf}\}$ and the target DML models is trained with this augmented set.

Data	PD	DINO$_H$ [9]				AD	Δ^*	**A-DINO$_H$**			
		R@1	R@2	R@4	R@8			R@1	R@2	R@4	R@8
$\mathcal{X}_{cub}$	A_f	64.2	75.6	84.4	91.0	A_s,A_a	1	**68.4**	**78.4**	**86.3**	**91.9**
	A_s	69.2	78.1	85.5	90.6	A_a,A_f	1	**72.8**	**81.6**	**88.3**	**93.4**
	A_a	63.1	74.3	83.0	89.9	A_f,A_s	1	**66.9**	**77.6**	**85.8**	**91.8**
$\mathcal{X}_{cars}$	B_s	70.3	79.2	86.8	91.7	B_c,B_p	1.5	**75.4**	**84.7**	**91.0**	**94.4**
	B_c	66.9	76.1	83.9	89.8	B_p,B_s	1.5	**75.8**	**84.2**	**90.4**	**94.5**
	B_p	56.9	68.6	78.2	85.6	B_s,B_c	1.5	**75.2**	**84.5**	**90.9**	**94.5**
		ViT$_H$ [9]						**A-ViT$_H$**			
$\mathcal{X}_{cub}$	A_f	78.3	87.0	92.4	96.0	A_s,A_a	1	**79.7**	**88.0**	**92.9**	**96.0**
	A_s	79.1	87.5	92.4	95.8	A_a,A_f	1	**81.7**	**88.4**	**93.1**	**96.1**
	A_a	77.7	86.5	92.2	95.8	A_f,A_s	1	**79.9**	**87.5**	**92.8**	**95.9**
$\mathcal{X}_{cars}$	B_s	65.1	76.2	84.6	90.7	B_c,B_p	1.5	**72.8**	**82.5**	**89.7**	**94.1**
	B_c	62.7	73.3	81.7	87.9	B_p,B_s	1.5	**74.1**	**82.8**	**90.0**	**94.2**
	B_p	51.0	63.1	73.8	83.2	B_s,B_c	1.5	**71.1**	**81.6**	**89.2**	**94.0**

Table 1: Results for automated *domain augmentation* in ATLANTIS. PD: domains present in the original training data, AD: a superset of augmented domains based on class-specific characteristics. A_f,A_s,A_a are 3 activity-based domains in $\mathcal{X}_{cub}$[34]; while B_s,B_c,B_p are 3 vehicle body-type based domains in $\mathcal{X}_{cars}$[17]. Δ^* represents optimal diversity factor Δ used in ORDC filtering.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Metric. We conduct experiments on three image retrieval benchmark datasets: CUB-200-2011 [34], Cars196 [17], and Stanford Online Products (SOP) [23]. To simulate data scarcity and distribution skewness, we engineered training subsets with selectively redacted certain patterns, while also maintaining the original test sets unaltered. Specifically, to control class distribution, we define the skewness parameter $\kappa : \mathcal{C}_{\text{total}} \rightarrow \mathbb{N}$ where $\mathcal{C}_{\text{total}}$ is the entire set of classes. For a subset of classes $\mathcal{C}_{\text{restricted}} \subset \mathcal{C}_{\text{total}}$, we applied a restriction on the number of samples to λ, generating controlled variations in class representation:

$$\kappa(i) = \begin{cases} \lambda & \text{if } i \in \mathcal{C}_{\text{restricted}} \\ N_i & \text{if } i \in \mathcal{C}_{\text{total}} \setminus \mathcal{C}_{\text{restricted}} \end{cases}$$

For domain distribution control, we first define a set of domains for each dataset based on DIG output. For example, for the CUB-200-2011 dataset, domains were categorized based on avian behaviors such as `Sitting' ($A_s$), `Swimming' (A_a), and `Flying' ($A_f$) ; while the Cars196 dataset has domain `Sedan' (B_s), `SUV-Crossover' ($B_c$), and `Performance Sport or Convertible' (B_p).

For a fair comparison across all experiments with or without novel class augmentation $\mathcal{N}$, *we maintain an unchanged test* set identical to the baselines, *ensuring that the augmented training set contains neither classes nor images from any test class*. While we restrict these generations during evaluation, in practice, the test set is always unknown. The ability of ATLANTIS to automatically infer test set information during augmentation would be a significant advantage. All images are resized to 224 × 224 input resolution. We adopt the standard recall at K (R@K) as the evaluation metric, following [16].

κ	ZS	Model	CUB-200-2011[34]				Cars196[17]			
			R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8
150	✗	DINO$_H$	72.0	80.6	87.6	92.7	46.4	57.5	68.2	77.5
		A-DINO$_H$	**76.5**	**84.5**	**90.5**	**94.4**	**77.8**	**85.6**	**90.8**	**94.2**
		ViT$_H$	80.1	87.8	92.5	96.1	43.5	54.9	65.7	75.6
		A-ViT$_H$	**82.4**	**89.1**	**93.5**	**96.3**	**74.5**	**83.7**	**90.1**	**94.1**
100	✗	DINO$_H$	75.6	83.8	89.7	93.9	66.7	77.0	84.9	90.3
		A-DINO$_H$	**78.3**	**85.7**	**91.6**	**95.1**	**79.6**	**87.5**	**92.8**	**96.0**
		ViT$_H$	82.9	89.5	93.7	95.6	59.2	70.9	80.5	87.9
		A-ViT$_H$	**83.9**	**89.9**	**93.9**	**96.5**	**76.9**	**85.6**	**91.2**	**95.0**
75	✓	DINO$_H$	64.3	75.7	84.7	90.9	61.8	73.0	81.6	88.4
		A-DINO$_H$	**76.0**	**84.3**	**90.3**	**94.1**	**80.6**	**88.1**	**92.9**	**95.7**
		ViT$_H$	78.1	86.9	92.6	95.9	58.4	70.6	80.7	88.2
		A-ViT$_H$	**83.5**	**89.9**	**93.6**	**96.3**	**76.5**	**85.5**	**91.3**	**95.0**
50	✓	DINO$_H$	71.0	81.5	88.5	93.6	75.0	83.5	90.0	94.4
		A-DINO$_H$	**76.5**	**84.8**	**90.5**	**94.7**	**82.6**	**89.2**	**93.8**	**96.5**
		ViT$_H$	79.1	87.5	92.6	95.7	70.2	79.6	87.0	92.4
		A-ViT$_H$	**83.1**	**90.0**	**93.8**	**96.1**	**78.6**	**86.7**	**92.0**	**95.6**

Table 2: Results for automated *class imbalance mitigation* in zero-shot (ZS) and full-shot learning tasks are presented. Here, κ denotes the number of training classes with $\lambda = 2$. The full-shot and zero-shot scenarios comprised 200 and 100 training classes for the CUB-200-2011 data, respectively, and 196 and 98 classes for the Cars196 data. The Δ of 1 for CUB-200-2011 data, and 1.5 for Cars196 was found optimal.

Baselines and Implementation. We compare ATLANTIS with current state-of-art in CBIR – DINO$_H$ and ViT$_H$ [9] – to evaluate performance gains in data-scarce scenarios. Additionally, for standard benchmarks, we also compare with Margin [37], NSoftmax [42], MIC [27], and IRT$_R$ [8]. The models in all experiments have ImageNet pretraining initialization and operates with embedding dimension of 128. We do not freeze the DML backbones and use the same losses and training procedures as in the original work [9]. For image-to-text metadata conversion (Fig.1, step-2), we employ BLIP-2 [18], and utilise SDXL [26] to synthesise realistic images (Fig.1, step-6) from textual descriptions. We use GPT-4 [24] (without vision component) as the LLM reasoning engine (Fig.1, steps 3, 4, 5). In the ORDC filtering phase (Fig.1, step-7), we experiment with $\Delta \in \{0.9, 1, 1.2, 1.5, 3, 5, \infty\}$ (detailed in Sup.Mat.), and also the experiments with active training feedback $T_{Feedback}$ is presented in Sup.Mat. only as we found insignificant gains *due to finetuning*.

4.2 Data-scarce Settings

Domain Augmentation. Table 1 shows that ATLANTIS exhibits significant image retrieval performance enhancements in multiple domain-scarce scenarios across all datasets and DML models. Specifically, R@1 scores increase up to 18.3% and 20.1% for the domain-specific Cars196 B_p category on DINO$_H$ and ViT$_H$ models, respectively. For the CUB-200-2011 A_s domain, the increments are 3.6% with ViT$_H$ and 2.6% with DINO$_H$.

Class Imbalance Mitigation. We evaluate class imbalance mitigation in both full-shot (sample-wise train-test split within each class) and zero-shot (class-wise train-test split) set-

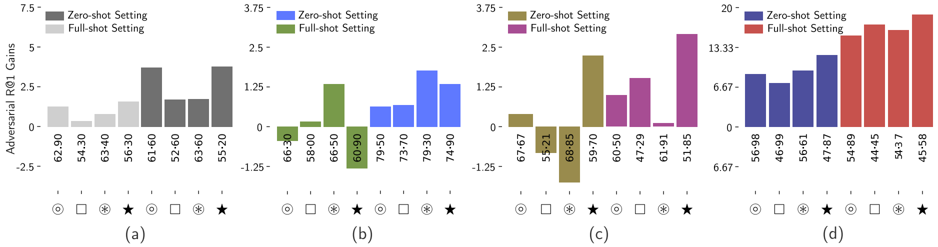

Figure 3: Adversarial Robustness Gains: The plots (a) and (b) for CUB-200-2011, and (c) and (d) for Cars196 data, present the improvements in adversarial R@1 achieved by our framework against white-box embedding-space PGD attacks of varying intensities. We measure robustness across different numbers of attack gradient steps s and the (ℓ_∞) adversarial noise size bound ε with the combinations: ⊚, □, ⊛, and ★ representing attack settings when $\{s=1,\varepsilon=0.05\}$, $\{s=1,\varepsilon=0.1\}$, $\{s=10,\varepsilon=0.05\}$, and $\{s=10,\varepsilon=0.1\}$ (ε expressing fraction of the image pixel values range). The figures (a) and (c) use original test data, while (b) and (d) use synthetic data. Model used for evaluation: DINO vs. **A-DINO**.

tings. Table 2 depicts that ATLANTIS consistently outperforms other baselines, especially in zero-shot scenarios with $\kappa = 75$, where training patterns are exceedingly scarce: R@1 scores increased by 11.7% over DINO$_H$, and by 5.4% over ViT$_H$ for the CUB-200-2011 data. For Cars196 data, the enhancements were 18.8% over DINO$_H$, and 18.1% over ViT$_H$.

Adversarial Robustness Gains To confirm ATLANTIS's effectiveness in enhancing adversarial robustness, we perform white-box feature-space [28] (embedding layer only) evasion (untargeted) attacks using the projected gradient descent (PGD) method [20] with varying attack strengths on both real and synthetic data. The untargeted PGD attack adds an optimised imperceptible noise plane (ε) to an image such that its embedding diverges significantly from its original embedding. For different attack strengths, we measure robustness across different numbers of attack gradient steps ($s \in \{1,10\}$) and magnitudes of the ℓ_∞ adversarial noise bound ($\varepsilon \in \{\frac{12.75}{255}, \frac{25.5}{255}, \frac{127.5}{255}\}$). Figure 3 presents the R@1 gains on white-box PGD attacks of varying strengths, devised with different ℓ_∞ noise bounds ε and adversarial optimisation gradient steps s. These attacks were crafted for causing evasion from the target models (DINO$_H$ and **A-DINO**$_H$) in the embedding space. The models trained with ATLANTIS are particularly robust against attacks with imperceptible noise levels ($\varepsilon \leq \frac{25.5}{255}$), resulting in up to an 18.85% increase in R@1 on the adversarial data. This confirms that ATLANTIS does not induce overfitting to the test set but rather leads to less sensitive and more generalised retrieval models.

4.3 Results on Standard Benchmarks

Although being designed for data-scarce scenarios, ATLANTIS still outperforms existing methods on standard benchmarks for all datasets as illustrated in table 3. We note that the augmentation power of ATLANTIS is constrained significantly for SOP [23] due to limited computing resources to address the vast number of classes in this dataset, yet ATLANTIS is consistently better than other baselines in most metrics. More details about SOP experiments and results are in Sup.Mat.

Method	CUB-200-2011[34] (K)				Cars196[17] (K)				SOP[23] (K)			
	1	2	4	8	1	2	4	8	1	10	100	1000
Margin [37]	63.9	75.3	84.4	90.6	79.6	86.5	91.9	95.1	72.7	86.2	93.8	98.0
NSoftmax [42]	56.5	69.6	79.9	87.6	81.6	88.7	93.4	96.3	75.2	88.7	95.2	-
MIC [27]	66.1	76.8	85.6	-	82.6	89.1	93.2	-	77.2	89.4	94.6	-
IRT$_R$ [8]	72.6	81.9	88.7	92.8	-	-	-	-	83.4	93.0	97.0	99.0
S-DeiT [9]	73.3	82.4	88.7	93.0	77.3	85.4	91.1	94.4	82.5	93.1	97.3	99.2
S-DINO [9]	76.0	84.7	90.3	94.1	81.9	88.7	92.8	95.8	82.0	92.3	96.9	99.1
H-DeiT [9]	74.7	84.5	90.1	94.1	82.1	89.1	93.4	96.3	83.0	93.4	97.5	99.2
DINO$_H$ [9]	78.3	86.0	91.2	94.7	86.0	91.9	95.2	97.2	84.6	94.1	97.7	99.3
ViT$_H$ [9]	84.0	90.2	94.2	96.4	82.7	89.7	93.9	96.2	85.5	94.9	98.1	99.4
A-DINO$_H$	**79.1**	**87.1**	**92.2**	**95.5**	**86.8**	**92.4**	**95.5**	**97.3**	**84.8**	**94.2**	**97.8**	99.4
A-ViT$_H$	**84.1**	**90.3**	**96.5**	**97.9**	**83.4**	**90.0**	**94.0**	**96.5**	85.4	**94.9**	**98.2**	**99.5**

Table 3: Standard Performance Benchmarks: ATLANTIS (**A-**) Surpasses state-of-art Models in CUB-200-2011, Cars196, and SOP datasets. Performance of all baselines is reported in [9]. Embedding size for all models was set to 128.

4.4 Ablations, Discussion, and Limitations

To assess the effectiveness and justify the presence of ATLANTIS's pipeline components, we compare the results with those of naive untargeted augmentation. For the same level of generations as in ATLANTIS, the performance of naive augmentation remained lower, but higher than in the non-augmentation case meaning the *series* containing DIG, APS, and ORDC is indeed effective. The details are included in the Sup.Mat due to space constraints.

Our evaluations confirm ATLANTIS's role in boosting CBIR model generalizability, improving outcomes with both clean and adversarial samples. ATLANTIS is particularly effective in data-limited contexts, offsetting shortages with targeted augmentation. It also performs well in balanced CBIR benchmarks, though the use of multiple LLMs and LVMs causes output variability. ORDC reduces, but does not eliminate, occasional instability from LLM biases and occasional hallucinations. As ATLANTIS has a modular design, these issues could be addressed in future with higher quality LLM and GenAI models. Regarding ethical issues, ATLANTIS currently relies on the built-in filters of LLM and GenAI to avoid generating inappropriate content, however a blacklist of objects and domains could also be integrated to our ORDC for more controllable synthesis. Another limitation is the computational overhead - as an example it takes 2.4 hours to analyze the CUB-200-2011 dataset and synthesize extra 228 images for domain scarcity A_f settings in table 1 using a standard Azure server with Intel Xenon GPU and 1x A100 GPU. Future work could leverage data parallelism and distributed training to reduce such overhead.

5 Conclusion

In conclusion, ATLANTIS advances CBIR systems by introducing a cohesive multimodal framework that mitigates available data insufficiency through targeted synthetic data augmentation and the effective leverage of foundational language and vision models. The suite of novel components in ATLANTIS, namely DIG, APS, ORDC, and the feedback mechanisms, has demonstrated superior performance in CBIR tasks, especially in data-constrained environments, while also reducing the adversarial sensitivity of these models due to improved generalisation, marking a significant milestone in the CBIR field.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

[3] Christopher Bowles, Liang Chen, Ricardo Guerrero, Paul Bentley, Roger Gunn, Alexander Hammers, David Alexander Dickie, Maria Valdés Hernández, Joanna Wardlaw, and Daniel Rueckert. Gan augmentation: Augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863*, 2018.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[5] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019.

[6] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[7] Aleksandra Edwards, Asahi Ushio, Jose Camacho-Collados, Hélène de Ribaupierre, and Alun Preece. Guiding generative language models for data augmentation in few-shot text classification. *arXiv preprint arXiv:2111.09064*, 2021.

[8] Alaaeldin El-Nouby, Natalia Neverova, Ivan Laptev, and Hervé Jégou. Training vision transformers for image retrieval. *arXiv preprint arXiv:2102.05644*, 2021.

[9] Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khrulkov, Nicu Sebe, and Ivan Oseledets. Hyperbolic vision transformers: Combining improvements in metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7409–7419, 2022.

[10] Alex Falcon, Giuseppe Serra, and Oswald Lanz. A feature-space multimodal data augmentation technique for text-video retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4385–4394, 2022.

[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. adaptive computation and machine learning. *Massachusetts, USA*, 2017.

[12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[13] Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. A new reduced-reference image quality assessment using structural degradation model. In *2013 IEEE international symposium on circuits and systems (ISCAS)*, pages 1095–1098. IEEE, 2013.

[14] Xiaoshuai Hao, Yi Zhu, Srikar Appalaraju, Aston Zhang, Wanqian Zhang, Bo Li, and Mu Li. Mixgen: A new multi-modal data augmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 379–389, 2023.

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[16] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33 (1):117–128, 2010.

[17] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[19] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[20] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[21] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

[22] Harry Nyquist. Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers*, 47(2):617–644, 1928.

[23] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016.

[24] R OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2:13, 2023.

[25] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

[26] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[27] Karsten Roth, Biagio Brattoli, and Bjorn Ommer. Mic: Mining interclass characteristics for improved metric learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8000–8009, 2019.

[28] Andras Rozsa, Manuel Günther, and Terranee E Boult. Lots about attacking deep features. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 168–176. IEEE, 2017.

[29] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[30] Claude E Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949.

[31] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[32] Rajiv Soundararajan and Alan C Bovik. Rred indices: Reduced reference entropic differencing for image quality assessment. *IEEE Transactions on Image Processing*, 21(2):517–526, 2011.

[33] William Thong, Jose Costa Pereira, Sarah Parisot, Ales Leonardis, and Steven McDonagh. Content-diverse comparisons improve iqa. *arXiv preprint arXiv:2211.05215*, 2022.

[34] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[35] Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. Deep metric learning with angular loss. In *Proceedings of the IEEE international conference on computer vision*, pages 2593–2601, 2017.

[36] Zhou Wang and Qiang Li. Information content weighting for perceptual image quality assessment. *IEEE Transactions on image processing*, 20(5):1185–1198, 2010.

[37] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE international conference on computer vision*, pages 2840–2848, 2017.

[38] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 819–828, 2020.

[39] Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. Generative data augmentation for commonsense reasoning. *arXiv preprint arXiv:2004.11546*, 2020.

[40] Guanghao Yin, Wei Wang, Zehuan Yuan, Chuchu Han, Wei Ji, Shouqian Sun, and Changhu Wang. Content-variant reference image quality assessment via knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3134–3142, 2022.

[41] Yuwei Yin, Jean Kaddour, Xiang Zhang, Yixin Nie, Zhenguang Liu, Lingpeng Kong, and Qi Liu. Ttida: Controllable generative data augmentation via text-to-text and text-to-image models. *arXiv preprint arXiv:2304.08821*, 2023.

[42] Andrew Zhai and Hao-Yu Wu. Classification is a strong baseline for deep metric learning. *arXiv preprint arXiv:1811.12649*, 2018.