Ads and Anomalies: Structuring the Known and Probing the Unknown

Thesis submitted in partial fulfilment
of the requirements for the degree of

Master of Science in ***Computer Science and Engineering*** *by Research*

by

K N Amruth sagar
2021701013
amruth.sagar@research.iiit.ac.in

International Institute of Information Technology, Hyderabad
(Deemed to be University)
Hyderabad - 500 032, INDIA
December 2024

International Institute of Information Technology

Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Ads and Anomalies: Structuring the Known and Probing the Unknown" by K N Amruth sagar, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date	Advisor: Prof. Ravi Kiran Sarvadevabhatla

To my brother, **Ksheera sagar**, with gratitude and love for your unwavering support and inspiration.

Acknowledgments

I would like to express my deepest gratitude to my advisor, Prof. Ravi Kiran Sarvadevabhatla, whose guidance has been instrumental throughout my master's journey. He taught me invaluable skills, such as identifying open questions in research, choosing research directions, and presenting ideas and results with clarity and simplicity. From crafting impactful presentations to understanding the importance of elements like typography, colors, diagrams, and layout, he emphasized how these details can shape the readability and reception of work. He also instilled in me the importance of thoroughness, to double-check all code and experimental results. His mentorship extended beyond research, as he guided me through efficient time management, teaching me the importance of planning tasks, like preparing posters and presentations post-submission. I am also grateful for the enriching collaboration with Arjun Reddy Akula and Varun Jampani, a collaboration made possible by my advisor's support and encouragement.

I am also immensely thankful to my friends, who are like brothers—Aditya, Bhoomeendra, Dhaval, Dhruv, Prateek, and Nayan. Their unwavering support in every aspect—from brainstorming ideas and coding assistance to discussions, trips, and celebrations—provided a strong foundation that helped me grow as a person.

Finally, I dedicate this thesis to my brother, Ksheera Sagar. Words cannot fully convey the extent of his support. Throughout my academic journey, he was always there, offering advice, helping me overcome research obstacles, and providing feedback on my ideas. His constant encouragement made even the most challenging tasks seem achievable.

Abstract

The convergence of computer vision and advertising analysis has seen progress, but existing advertisement datasets remain limited. Many are small subsets of larger datasets, and while larger datasets may offer multiple annotations, they often lack consistent organization across all images, making it challenging to structure ads hierarchically. This lack of clear categorization and overlap in labeling hinders in-depth analysis. To address this, we introduce MAdVerse[1], a comprehensive, multilingual dataset of over 50,000 advertisements sourced from websites, social media, and e-newspapers. MAdVerse organizes ads into a hierarchy with 11 primary categories, 51 sub-categories, and 524 specific brands, facilitating fine-grained analysis across a diverse range of brands. We establish baseline performance metrics for key ad-related tasks, including hierarchical classification, source classification, and hierarchy induction in other ad datasets and, in a multilingual context, thereby providing a structured foundation for advertisement analysis.

In our second work, we investigate foundational aspects of out-of-distribution (OOD) detection. Existing OOD benchmarks typically focus on broad, class-level shifts but lack controlled environments for assessing how individual attribute changes such as color or shape affect OOD detection. To bridge this gap, we created two synthetic datasets, SHAPES and CHARS[2], each designed to allow controlled experimentation with isolated shifts in attributes. Through variations in color, size, rotation, and other factors, these datasets facilitate a targeted examination of OOD detection performance under specific conditions, providing insights into how OOD detection is affected under different attribute shifts. Later, we apply OOD detection methods to advertisements, where models face real-world distribution shifts characteristic of diverse advertising styles.

Our contributions, MAdVerse for structured ad analysis and SHAPES and CHARS for controlled OOD studies emphasize the importance of robust, adaptable models for both foundational research and practical applications in advertisement analysis.

[1] MAdVerse: A Hierarchical Dataset of Multi-Lingual Ads from Diverse Sources and Categories. Project Page

[2] In the known, Out of ordinary: Probing OOD detection with synthetic datasets. Github Repo

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The ways we connect, communicate, and learn have changed dramatically in a world characterized by rapid digital transformations and constant connectivity. The emergence of social media, the growth of e-commerce, and the accessibility of information have all contributed to the development of an ecosystem in which people are continuously interacting with a variety of content types. Because of this, businesses in every industry are now investigating ways to stay visible, relevant, and competitive in a market that is becoming more and more crowded. Organizations face the difficult task of differentiating their brand in this dynamic and fragmented marketplace, where more brands than ever are fighting for consumers' attention.

Reflecting the pressures of this environment, companies have dramatically increased their investment in advertising and marketing to boost brand recognition and foster consumer engagement. This heightened expenditure underscores the critical role advertising plays in competitive markets, where visibility has become as crucial as product quality or service excellence. The average consumer is estimated to encounter approximately 20,000 brand messages each day [1], highlighting the immense challenge of standing out amid this daily deluge of promotional content.

In recent decades, advertising strategies have changed dramatically, moving from print, radio, and television to a wide range of digital platforms. Brands can engage with audiences in new and highly interactive ways through data-driven programmatic advertising, social media, and search engine marketing. For instance, social media makes it possible to share content that encourages direct interaction, and search engines enable intent-based targeting that draws users in at strategic times. With the help of each digital tool's distinct features, advertisers can instantly modify and improve their messaging, increasing the efficacy of their ads based on engagement metrics and real-time feedback.

This thesis can be viewed in two main parts. In the first part, we introduce a new multilingual, hierarchical ad dataset, *MAdVerse*, designed to go beyond traditional plain class labels, which are often too simplistic to capture the full scope of brand identity and product type in an advertising context. By utilizing hierarchical labels, we enable a classification structure that reflects the natural hierarchy in which brands and products can be organized—a structure that is both more intuitive and more informative. This approach not only enriches the dataset with additional contextual information but also allows for more analytical insights.

The above work naturally leads to a broader, intriguing question: Given the wide variation in visual styles, colors, and compositions across images, what specific visual attributes most influence a model's ability to detect unfamiliar inputs? This leads us to Out-of-Distribution (OOD) detection—a field focused on identifying inputs that are novel or significantly different from the data a model was trained on. More specifically, we examine which visual attributes of an image most influence the OOD detection. Additionally, we investigate whether the choice of model backbone plays a significant role in these OOD outcomes, and explores whether these methods reveal any consistent patterns in how they respond to these features. Understanding these factors could deepen our insights into the robustness of detection methods across a range of visual contexts. By uncovering these influencing factors, we can better anticipate how detection methods handle this variability in attributes.

Finally, we evaluate the effectiveness of hierarchical classifiers trained on our structured ad dataset, MAdVerse, when they encounter new images. This will allows us to observe how well these classifiers operate not only within established brand and product categories but also in managing the challenge of novel advertisements, offering insights into the classifiers' ability to generalize and adapt in the ever-evolving landscape of advertising.

Chapter 2

MAdVerse: A Hierarchical Multi-Lingual Ad Dataset

2.1 Existing ad datasets

Hussain et al. [2] were the first work to propose automatic advertisement understanding. The dataset provides annotations related to ad topic, sentiment, ad strategy and symbols. However, the annotations do not include any notion of hierarchy. The work of Fosco et al. [3] predict visual importance in graphic designs and includes a subset of ad images from Pitt ads [2] since ads can be perceived as graphic designs. Liang et al. [4] propose an ad dataset of 1000 images dubbed ADD1000 and explore saliency prediction on advertising images using eye movement data. More recently, Kumar et al. [5] use a subset of 3000 ad images from Pitt ads [2] and annotate them with persuasive strategies utilised in advertisements. A smaller subset of 250 images in their dataset also have segmentation masks for various ad regions. Cao et al. [6] generate image-conditioned layouts for advertisements. This work also provides an ad poster layout dataset with saliency maps and layout annotations for all ad layouts.

With the exception of the Pitt Ads dataset [2], other works typically operate within the confines of utilizing subsets of ad images sourced from Pitt Ads, which they then augment with task-specific annotations. Alternatively, some works employ ad images, not as their primary focus on advertisements but as supplementary tools to address their specific research objectives. What stands out as a significant gap in this landscape is the absence of a comprehensive ad dataset that encompasses a wide array of advertisements from diverse sources, languages, styles, and also incorporates semantic grouping of brands or products.

2.2 Motivation & Contribution

To empower computer vision systems to excel in comprehending this diverse ad landscape, exposure to various sources and languages is essential. This exposure enables systems to recognize different ad types across contexts, adapt to evolving advertising trends, and effectively process ads on diverse platforms. Surprisingly, existing works on advertising datasets [2, 5, 6, 4, 3] do not emphasize collecting ads from multiple sources and languages. Most of these datasets either consist of scraped images or are subsets of larger datasets with annotations specific to particular tasks.

For example, the ad dataset by Hussain et al. [2] is a large-scale resource with multiple types of annotations, yet it has limitations in consistency. Not all images in the dataset contain all annotation types, so attempting to organize them into a hierarchical structure reveals cycles and closed loops instead of forming a directed acyclic graph (DAG) or a tree. This overlap in annotations creates ambiguities about image categorization; grouping images based on certain annotations results in divisions that closely resemble an n-class classification, offering limited flexibility for hierarchical organization.

Rest of the datasets are curated as small subsets of larger collections, with annotations tailored to specific research objectives, such as eye fixation patterns in ads [4] or identifying persuasive strategies used in ads [5]. In other cases, advertisements appear only incidentally within datasets where they are not the primary focus. For example, the dataset by Fosco et al. [3] is aimed at predicting visual importance in graphic designs; advertisements are included only because they contain similar graphic elements. Consequently, these datasets lack a comprehensive focus on advertisements, limiting their applicability for broader ad-specific analysis."

Consequently, there is a noticeable gap in datasets that truly capture the multilingual, multi-source diversity of advertisements, underscoring the need for a more robust dataset that facilitates advanced understanding across varied advertising contexts.

To address these limitations, we present MAdVerse, a novel multi-source, multilingual, and hierarchical advertising dataset. MAdVerse is designed to capture the complex relationships between brands and products, providing a deeper understanding of their characteristics and how they are represented visually and textually in ads. This dataset employs a hierarchical taxonomy, organizing information from broad product categories to specific brand names, which enables detailed analysis of product-brand associations. Comprising over 50,000 ads, MAdVerse includes data from a variety of platforms—social media, websites, newspapers—reflecting current advertising trends and practices.

In addition to being multi-source, MAdVerse is multilingual, spanning 11 languages, which allows for a broader examination of the linguistic and cultural dimensions of advertising. This multilingual scope enables researchers to investigate how language choice impacts ad interpretation and consumer response. By including annotations related to both products and brands, MAdVerse supports in-depth analyses of advertising strategies, potentially revealing insights into how language and cultural context influence ad effectiveness.

2.3 MAdVerse

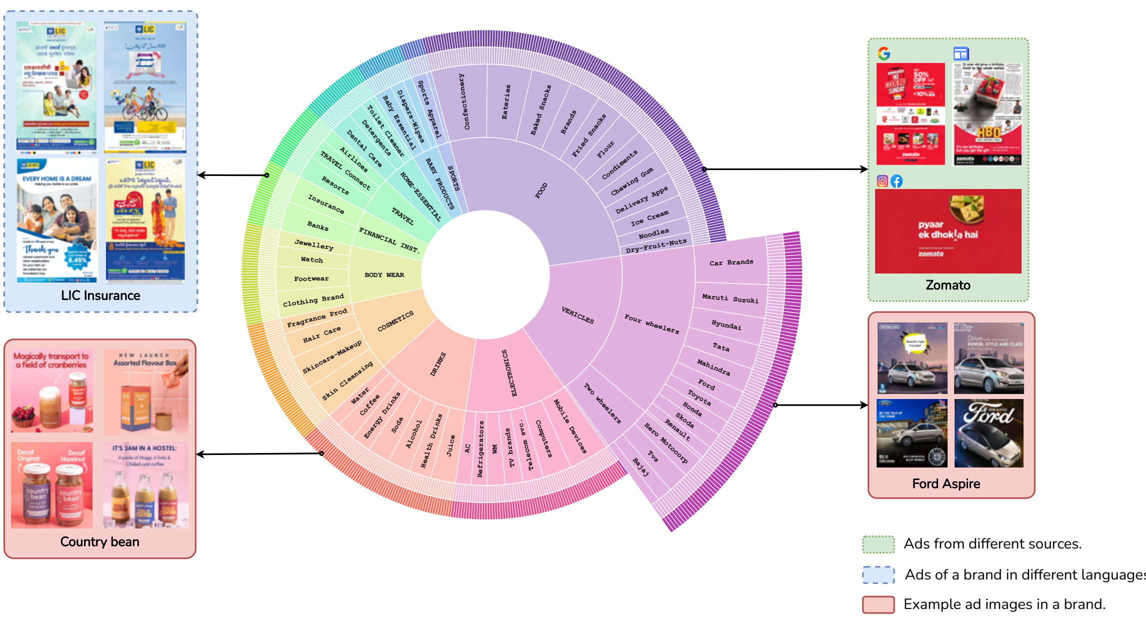

Figure 2.1: Hierarchical Taxonomy of the Dataset.

2.3.1 Collecting ad images from web

We created a diverse and high-quality dataset of image advertisements by scraping from multiple sources – Google images, social media platforms (Facebook and Instagram), dedicated websites (e.g. Advert Gallery [7]) and digital newspapers in various languages. We compiled a comprehensive list of popular brands to ensure representation across various advertisement topics such as food, clothing, sports, and hygiene. Our objective was to collect images for each brand from these sources.

For Google images, we used a fixed prompt prepended to the brand name to query and retrieve the top 200 images associated with the brand. To maintain dataset quality and uniqueness, we employed a CNN-based method from the imagededup library [8] to identify and remove exact duplicates and similar images above a similarity threshold of 0.95. This resulted in a diverse collection of images without any exact duplicates.

We then implemented a two-stage process to remove non-advertisement images to refine the dataset further. First, we formulated a rule-set to define the criteria for an image to be classified as an advertisement. Specifically,

- images that featured only brand names, arbitrary text, or individuals without any evident advertising intent were categorized as non-advertisements.
- Frames extracted from video advertisements were similarly labeled as non-ads unless they explicitly showcased a connection to the brand through identifiable logos or contextual brand elements.
- images displaying solely the product on a plain background, devoid of text, taglines, metaphors, or even a logo, were deemed insufficient for classification as an advertisement

This systematic classification allowed us to maintain a consistent standard and mitigate ambiguity in identifying true advertisements within the scraped images. The examples of what is considered as an Ad and not an Ad is given in Figure 2.2

Next, we manually annotated a subset of images according to the rule-set and used them to train an Ad classifier incorporating a Vision Transformer (ViT) [9] and a classifier head, which demonstrated accuracy of 93.77% and F1 score of 0.898 on the test set with train, test, val split distribution of 70%, 20%, 10%. The trained classifier was then applied to the unannotated set of images, effectively extracting ad images from the dataset.

We developed an annotation tool to address artifacts such as ads of similar products but not similar brands within the same brand folder and to remove hard negative non-ads. We categorized images as ad or non-ad using this tool and placed them into the appropriate brand folders. During this process, we encountered ads featuring brands not initially included in our brand list and collaborative advertisements involving multiple brands. We categorized such images as miscellaneous.

To address under-represented brands containing only a few images, we scraped advertisements from popular social media platforms like Instagram, Facebook using the Apify [10] scraping tool. After manually selecting relevant ads, we placed them in their corresponding brand folders.

Figure 2.2: Ads and Non-Ads examples

2.3.2 Collecting ads from newspapers

We gathered ad images from Advert-gallery [7], an image repository of ads from digital newspapers and websites of e-newspapers in various languages. We randomly sampled pages from different newspapers and annotated the ads using the MS-COCO [11] format. In this section, we outline our methodology for extracting newspaper advertisements from digital newspapers. Our approach involved multiple steps, beginning with the acquisition of 30,987 ads from digital newspaper sources.

We employed Faster R-CNN[12] model with a ResNet-50 Backbone, with a Feature Pyramid Network (FPN) architecture. This model was pre-trained on the Newspaper Navigator Dataset [13], setting the foundation for our subsequent fine-tuning efforts.

Fine-tuning of the network was carried out by freezing the weights of the backbone and adjusting the weights of the fully connected layers. This fine-tuning process relied on a carefully annotated dataset of 2,863 English newspaper page images, utilizing an 80-20 percent train-test split strategy.

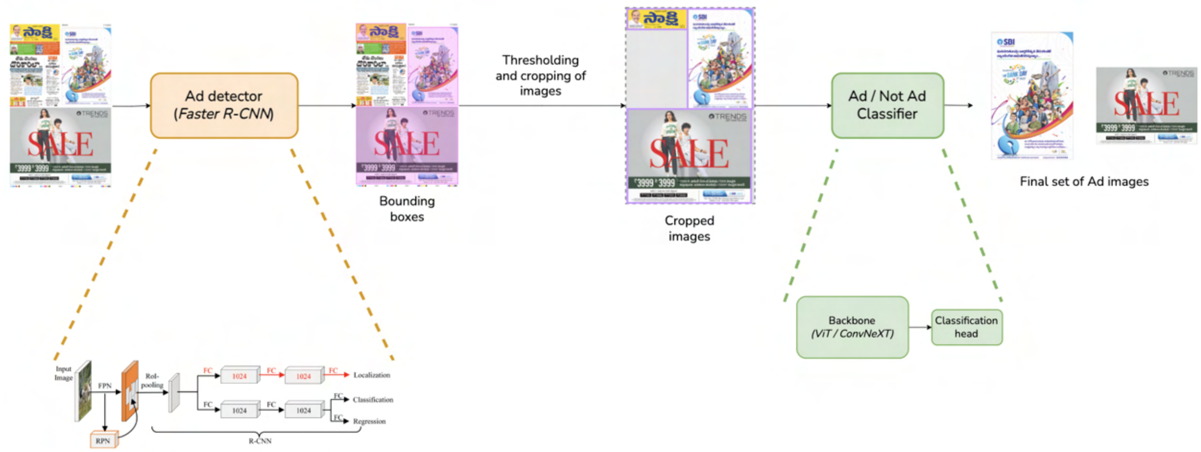

Figure 2.3: A Faster R-CNN-based detection pipeline designed to identify potential advertisement images, extract bounding boxes, crop images, and filter ads from non ads using an Ad/Non ad classifier

Subsequently, we implemented an ad classifier based on a ViT/ConvNext Backbone to further enhance the precision of our ad extraction pipeline. This classifier played a crucial role in eliminating false positives from the data obtained through the detection stage. To ensure dataset accuracy, we performed manual validation to eliminate false-positive images.

Finally, to address any remaining duplicates within the dataset, we conducted an additional deduplication step similar to the initial one. The final dataset is comprised of **53,406** images. Of these, **22,428** images came from WebAds and **1,947** images were from Advert Gallery. The dataset images contain hierarchical annotation of categories and sub-categories, reflecting meaningful relationships between the images and their respective brands.

The hierarchy is unbalanced, meaning leaf nodes are at different heights in the tree and the hierarchy has 5 levels, providing a detailed and well-organized dataset. The statistics of the hierarchy are presented in Table 2.1. A pictorial illustration of the dataset can be viewed in Fig. 2.1.

2.3.3 Dataset annotations

2.3.3.1 Hierarchical Annotations

Our dataset has a systematic organization of products and their associated brands into a structured taxonomy that captures not only their individual attributes but also the relationships that exist between them. By structuring the data in this manner, we can provide a more comprehensive understanding of

Dataset	# images	Dataset source	Hierarchy present	Taxonomical levels	Languages
Pitt Ads[2]	64,832	Google images	✗	-	-
Persuasion strategies[5]	3000	Subset of Pitt Ads	✗	-	-
Fixation Prediction[4]	1000	Subset of Pitt Ads	✓	2	-
Visual Importance [3]	1000	Subset of Pitt Ads	✗	-	-
MAdVerse	**53,406**	**Google, Instagram & Facebook, Newspapers**	✓	**5**	**11**

Table 2.1: Comparitive overview of related datasets

advertisements.

The taxonomy comprises 5 levels that progressively detail the categorization of ad images. Each hierarchical level maintains the same level of abstraction. The structure of our hierarchy is shown in Fig. 2.1. The highest level has product categories such as "Food", "Electronics", and "Vehicles" which get subdivided into fine-grained brand examples such as "Ice cream", "Washing Machine", and "Ford" as we go down the hierarchy. The taxonomy is designed such that the last-level nodes are brand names, while the non-leaf nodes are clustered into a hierarchy on the basis of their product features.

Images from the internet are primarily retrieved using keyword query-based techniques. The keywords are organized hierarchically, ensuring the downloaded data is automatically structured within this hierarchy.

2.3.3.2 Ad Source Annotation

Our dataset has a wide array of ad images collected from various online sources, each contributing to the diversity and comprehensiveness of our dataset. The sources consist of popular social media platforms such as Facebook, Instagram, and Google databases, as well as other websites, thus reflecting a realistic representation of digital advertising campaigns. Additionally, ad images sourced from e-newspapers further expand the dataset's scope by incorporating images from a publicly available website

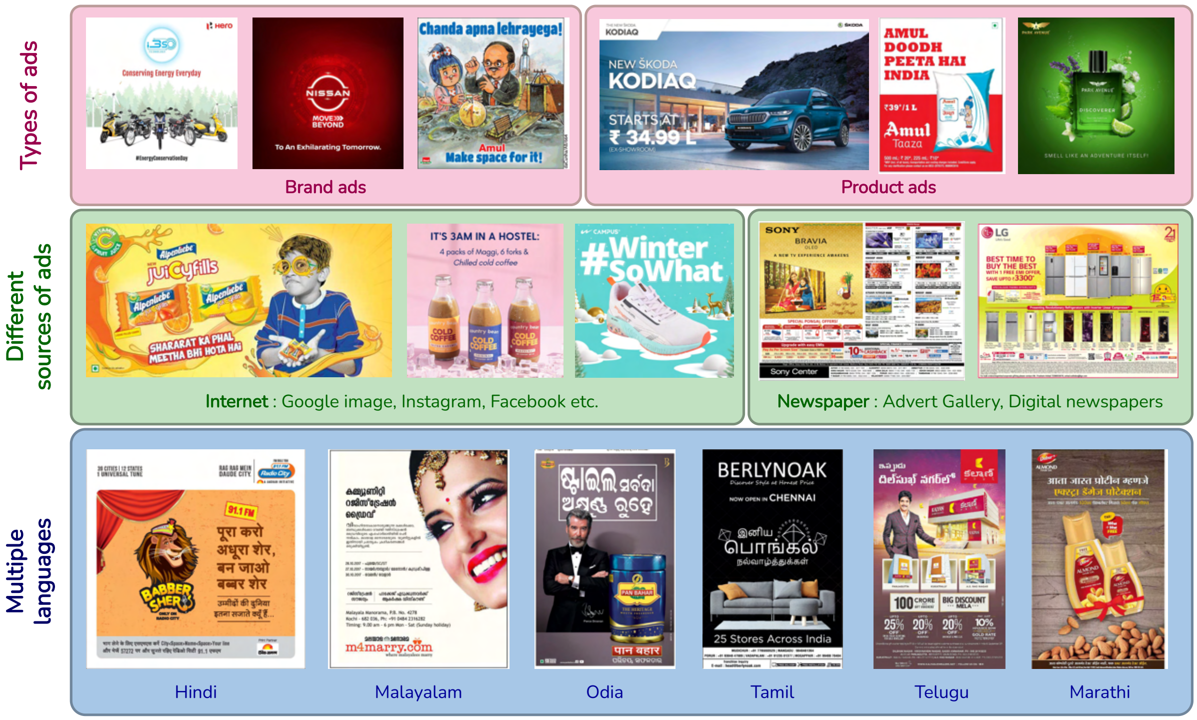

Figure 2.4: An illustration of the diverse annotations in the MAdVerse dataset.

the advert gallery, and those appearing in publicly available newspapers. Combining ad images from multiple sources results in a robust and versatile dataset that captures a variety of modern advertisements.

Images sourced from platforms like Facebook, Instagram, and Google databases capture the dynamic nature of digital advertising on social media. These sources offer a range of ad formats, including banners, carousel ads, and sponsored posts. Incorporating such images allows our dataset to mirror the real-world scenario of ad campaigns tailored for social media platforms, showcasing different styles, messaging strategies, and product showcases. This inclusion ensures that our dataset is not limited to a specific subset of advertisers or industries, but rather encompasses a broader spectrum of products and promotional approaches (see Fig. 2.4).

The incorporation of ad images from e-newspapers introduces a dimension of traditional advertising, in contrast to the digital realm. These images are obtained from a publicly available website advert gallery, which is a collection of ads from Indian newspapers. Ads from e-newspapers stand out due to presence of more text compared to digital ads, featuring more extensive written content and occasionally including contact information. Including these sources underscores the dataset's ability to capture both historical and contemporary advertising practices, catering to research on evolving trends and long-term impact analysis.

By bringing together ad images from such diverse sources, our dataset is no longer bound to one type of advertising and it helps researchers to investigate differences in various formats of advertising, analyze cross-platform strategies, and explore the evolution of ad strategy overtime. This diverse range of ad sources underpins the dataset's potential to address a multitude of research questions, making it a valuable resource for advancing the field of computer vision in the context of advertising analysis and understanding consumer behavior.

2.3.3.3 Product vs Brand Annotations

Our dataset contains two types of images: one for products and another for brands. Each type has its own role in showing the different aspects of advertising content. This differentiation is integral to capturing the diverse strategies and messaging techniques employed by advertisers.

Product images constitute a category that revolves around highlighting specific offerings or services. These images centre on showcasing the distinct features, benefits, and selling points of individual products. In contrast, brand images occupy a different sphere within the dataset. These images are crafted to enhance a company's overall brand image and identity, transcending the promotion of individual products.

By distinguishing between product and brand annotations, our dataset provides a representation of advertising content strategies (see Fig. 2.4). This separation allows us to analyze how brands convey their messages differently when focusing on specific products versus building a cohesive brand identity. Moreover, this distinction allows us to explore the interplay between product-centric and brand-centric advertising and the varying impacts these approaches may have on consumer perception and engagement.

2.3.3.4 Multilingual Annotations

Including ads in different languages is crucial because it helps us understand how advertising messages are adjusted to suit various languages and cultural settings. Multilingual annotations enable researchers to study how advertisers use different languages to connect with people. They can look at how words and meanings change, and how ads consider different cultures and feelings to reach people who

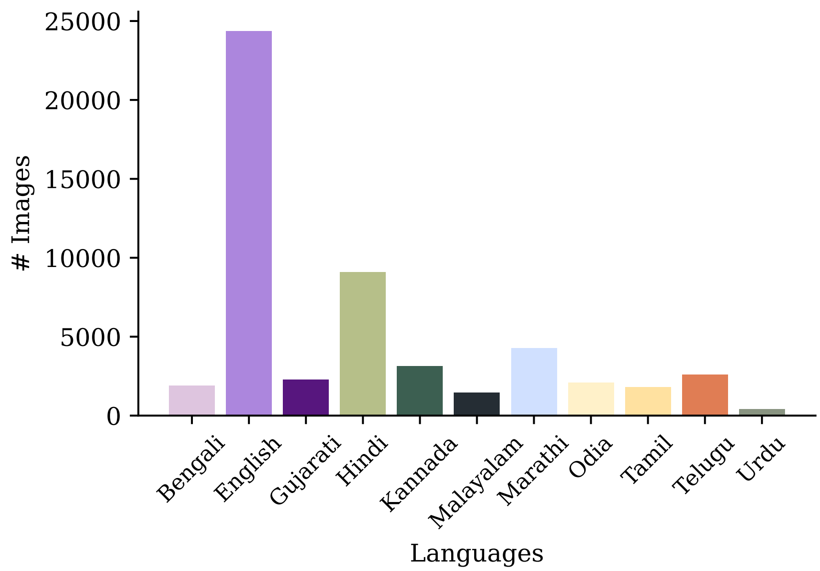

Figure 2.5: Image distribution per language in the dataset.

speak different languages. This multilingual perspective is crucial for uncovering universal principles as well as region-specific dynamics that underlie effective advertising campaigns and consumer behavior.

Our dataset is a testament to the linguistic diversity inherent in modern advertising campaigns, featuring ads in 11 distinct languages: Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Tamil, Telugu, and Urdu. The fact that our dataset includes multiple languages adds a special aspect that matches the multinational nature of advertising and accommodates a wider range of research questions.

We provide multilingual annotations for all 53,406 images in a total of 11 languages, the majority of them being in English, Hindi, and Marathi. The distribution of languages in the dataset can be seen in Fig. 2.5.

Chapter 3

Setup and Evaluation of Hierarchical ad classification and other tasks

We have discussed the dataset, the data collection process, its multiple sources and annotations and the different types of ad images in the previous chapter. Now, we present baselines for hierarchical classification, inducing hierarchy in other ad datasets and other tasks.

3.1 Hierarchical image classification

Hierarchical image classification predicts classes for a given ad image, at every level in the hierarchy. The first step in hierarchical image classification is to define the structure of the hierarchy. The nodes and the relationship between them can be represented in multiple ways, like Directed Acyclic Graphs and trees where the presence of an edge between node indicates a type of relationship between them. Deng et al. [14] introduce the concept of Hierarchy and Exclusion (HEX) graphs, a graph formalism capturing different types of semantic relations between labels, and they provide rigorous analysis of the properties of the proposed graphs.

The studies take different directions determined by the point at which the hierarchy is integrated. [15, 16, 17, 18, 19] imbue the architecture with a hierarchical structure, like having a branch net per level of hierarchy, to accommodate hierarchical classification, while semantic embedding from [20], HXE(Hierarchical cross entropy) and Soft labels from [21] incorporate the underlying hierarchy into loss functions via tree-based metrics. It appears that the literature on hierarchical classification may not be extensively developed.

HD-CNN [15] is the first work to use deep learning for hierarchical image classification by using a two-level category hierarchy to enhance image classification. Works like B-CNN [16], H-CNN [17] and [22] employ a shared approach of introducing hierarchy to datasets without a hierarchy or with inadequate levels by either manually creating necessary class groupings or utilizing WordNet [23] for semantic class grouping, facilitating hierarchical classification. B-CNN [16] does this with MNIST [24], CIFAR100 [25] datasets and H-CNN [17] adds hierarchy artificially to Fashion MNIST [26]. [22] adds hierarchy to fine-grained datasets like CUB-200-2011 [27] and Stanford cars [28] to show the benefits of hierarchical approach for fine-grained visual classification.

Instead of using a single label structure like most of the works, MMF [18] integrates different label structures to obtain a diverse set of prior information about the categories and fuse them to achieve better hierarchical classification. Chen et al. [19] do hierarchical multi-granularity classification, in which images can be classified to a particular level of granularity in the hierarchy. Level category prediction proposed by Wang et al. [29] utilizes cross-attention between hierarchical class word embeddings and image embedding, followed by pooling of output embeddings for score-based level and category prediction.

As previously mentioned, some works introduce hierarchy by formulating the loss, which leverages hierarchical information, and few others make the target label hierarchy aware, thus transforming the classification task into a label prediction. [20] maps n classes onto an n-dimensional unit hypersphere, where the distance between any two labels correlates to the distance between them in a tree. Bertinetto et al. [21] introduce two kinds of loss; one is Soft-labels, a label embedding method that uses tree-based metrics like LCA height to map one-hot labels to hierarchy aware labels; the second is hierarchical cross-entropy loss (HXE), which calculates conditional probabilities based on the relation between the two classes, thus bringing in the hierarchy. Discrete optimal transport based loss from [30] incorporates information about hierarchy by using tree-based distances in ground distance matrix of discrete optimal transport framework.

In the context of our ad dataset, we explore a combination of representative backbones, hierarchical/non-hierarchical architectures, and hierarchical /non-hierarchical losses. We are the first to benchmark hierarchical approaches on an ad dataset.

3.2 Model Architectures and Backbones

Hierarchical classification predicts classes for images at various levels of granularity in an underlying structured hierarchy. The structure of a hierarchy depends on the relationship between the classes in the dataset. In our case, nodes present in each level are mutually exclusive classes, and our hierarchy is a tree, and every node has only one parent. To conduct a comprehensive evaluation, we consider various combinations of backbones, classifier configurations, loss functions and metrics - see Fig. 3.1.

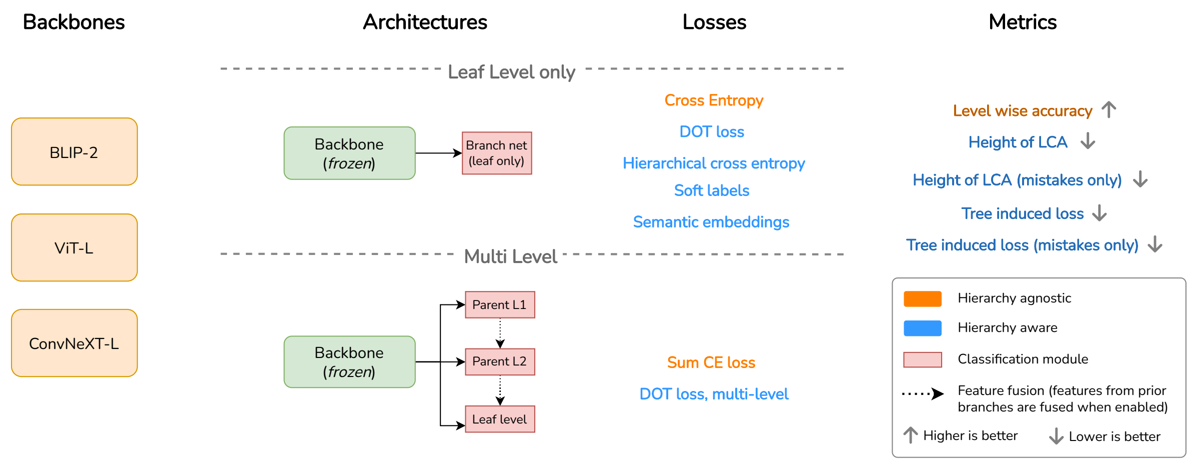

Figure 3.1: All combinations used in the experiment.

Backbones: We consider ViT-L [31], ConvNeXT large[32] and BLIP-2[33] as representative backbones. While the ViT model has shown great results in tasks like object detection and classification tasks, a CNN-based backbone like ConvNeXT [32] is robust to noise in input and is deemed efficient compared to transformer-based vision backbones. BLIP-2 is a backbone which is aware of both the visual and the textual modality. The latter is crucial since ads consist of pictures and text.

Classification Architectures: We consider two classifier architectures to perform hierarchical classification. One approach is to directly predict the fine-grained brands (i.e. leaves in the hierarchy), and then obtain parent predictions by traversing from predicted leaf to the topmost level. Another approach is have a classification branch for each level of the hierarchy to obtain class predictions.

3.3 Losses and Metrics

3.3.1 Losses

This is a brief overview of the loss function we used.

- Sum CE: Sum of cross entropy loss from all levels, which is like treating all the levels individually and equally without taking the hierarchy into account.
- Simple CE: Just a simple cross entropy loss

- HXE [21]: This is hierarchical cross entropy, which takes level wise softmax probabilities and uses conditional probability to encode the information about the hierarchy.
- Soft labels [21]: Instead of 1-hot vectors, this makes the target level hierarchy aware using level wise softmax probabilities.
- Semantic Embedding [20]: This maps all the one-hot targets onto a unit-hyper-sphere of the same dimension as the number of classes, enforcing tree based distances between the vectors to make the target embedding hierarchy aware.
- DOT loss [30]: It uses tree based metrics to construct a matrix called ground distance matrix, which is used in discrete optimal transport framework.

3.3.2 Metrics

Two classes of metrics were used. One is hierarchy agnostic, and the other is hierarchy aware. **Hierarchy agnostic**: These are standard metrics like level-wise accuracy. They don't take into account the hierarchical structure of the data and treat each level as a separate set of n classes.

Hierarchy aware: We have used 4 types of hierarchy aware metrics. For all hierarchy-aware metrics, lower values indicate better performance:

- Height of LCA: LCA represents the common ancestor at the lowest level connecting predicted and ground truth nodes. Higher values of this metric indicate errors made in predicting the upper levels of the hierarchy.
- Height of LCA (mistakes only): Similar to the above metric but focused specifically on misclassifications. This metric helps us assess the severity of mistakes made.
- TIE (Tree Induced Loss):TIE measures the shortest distance between the predicted and ground truth nodes. When both nodes are on the same level, it simplifies to twice the LCA height. In imbalanced hierarchies, it represents a distance from the ground truth to predicted node, indicating how far the predictions have deviated. Similar to Height of LCA, higher TIE value indicates the predictions deviated from ground truth in the earlier levels of hierarchy.

- TIE (mistakes only):This metric is the same as TIE but only considers misclassifications. It provides insight into the severity of mistakes made by the hierarchical classifier.

3.4 Dataset split

We create 70-10-20 split for training, validation and testing respectively. This includes english ads from all the sources, including web advertisements and e-newspaper ads, encompassing all brands present in the dataset. To ensure that each set is representative, the split was conducted in a stratified manner, maintaining the presence of all classes in the hierarchy across the training, validation, and test sets. This splitting was performed at the image level: within each class, 70% of images were randomly selected for training, with the remaining 30% divided between validation (10%) and testing (20%). This method ensures that each subset retains a diverse representation of classes, supporting balanced evaluation.

3.5 Results

3.5.1 Hierarchical classification

The top two performers from each approach are given in the table 3.1. As expected, fine-grained classification is generally harder given the number of classes and per-class image distribution. Since BLIP-2 is trained on various vision-language tasks, it exhibits best performance across the board, demonstrating the advantage of a language-and-text based backbone choice. Compared to hierarchy-agnostic cross entropy losses, hierarchy aware label embedding methods such as Semantic embedding [20] and Soft Labels [21] enable superior performance in predicting the classes accurately across the tree levels.

3.5.2 Inducing hierarchy in other ad datasets

To assess the generalization of our hierarchical classifiers and to demonstrate the efficacy of providing hierarchical annotations to ads beyond our dataset, we conducted experiments by inducing hierarchy onto a sampled set of images from Pitt ads [2]. As [2] does not provide a test set explicitly, we sampled slightly more than 200 images from the ads, keeping in mind that images will be removed from the test set that are already present in our dataset. Finally, we had 234 images.

Architecture		Backbone	Loss	Accuracy (%)			Height of LCA			Height of LCA (mistakes only)			TIE			TIE (mistakes only)		
				L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3
Leaf level		BLIP-2	Semantic embedding	98.17	94.56	52.54	0.02	0.07	0.55	1.00	1.34	1.15	0.04	0.15	1.09	2.00	2.67	2.31
		BLIP-2	Soft-labels	97.46	93.74	85.25	0.03	0.09	0.24	1.00	1.41	1.60	0.05	0.18	0.47	2.00	2.81	3.19
Multi level	with FF	BLIP-2	Sum CE	95.16	89.25	81.73	0.05	0.15	0.29	1.00	1.41	1.60	0.10	0.30	0.58	2.00	2.82	3.20
	without FF	BLIP-2	Sum CE	95.51	89.34	80.87	0.04	0.15	0.31	1.00	1.41	1.61	0.09	0.30	0.62	2.00	2.83	3.23

Table 3.1: A comparison between leaf-only and multilevel approaches, with a focus on the top two performers

The images were manually annotated up to the sub-category level (L1 and L2) due to the difficulty in annotating Level 3, as the brand of the images did not align with those in MAdverse. Annotating images only till level 2 ensured consistency with the hierarchical structure of our dataset. We used the hierarchical classifiers which were trained on our English ads, as mentioned in the previous section. We chose the BLIP-2 as the backbone with multilevel architecture with feature fusion, trained with level-wise sum of cross entropy loss.

Table 3.2 shows the level-wise accuracy of the model used for hierarchical placement. The results show that the classifier trained on our dataset shows reasonable cross-dataset generalization, which is particularly relevant for researchers seeking to leverage hierarchical annotations in diverse domains.

Metrics	Hierarchy Agnostic								Hierarchy Aware							
	Accuracy (%)		Precision		Recall		F1 Score		Height of LCA		Height of LCA (mistakes only)		TIE		TIE (mistakes only)	
	L1	L2	L1	L2	L1	L2	L1	L2	L1	L2	L1	L2	L1	L2	L1	L2
Pitt's Ad Dataset	86.67	72.15	0.74	0.49	0.73	0.47	0.73	0.45	0.12	0.41	1	1.46	0.24	0.81	2.93	5.63
Multilingual Ads	96.91	87.29	0.71	0.61	0.70	0.62	0.71	0.60	0.03	1.62	1	1.27	0.06	0.32	2.00	2.54

Table 3.2: Results of inducing hierarchy on a subset of Pitt's Ad Dataset and our multilingual dataset.

3.5.3 Inducing hierarchy on multilingual ad images

Similar to the previous task, we wanted to assess the ability of our hierarchical classifiers trained on English ads to classify multilingual ads. We selected a sample of slightly more than 200 images from multilingual ads. We repeated the same step of removing images will be removed from the sampled set that is present in the English dataset. Finally, we had 291 images. Similar to the previous task, we chose BLIP-2 as the backbone with multilevel architecture with feature fusion, trained with Sum_CE loss, since it has the best level-wise accuracy out of all the configurations tried. Table 3.2 shows the level-wise accuracy of the model on multilingual images. The classifier performs very well in predicting the first level of the hierarchy, but the prediction accuracy drops slightly. The main reason behind this is that the images used for inference contain text in multiple languages, and most of these images are from newspapers. Newspapers tend to have more text content in various languages, which makes it a bit harder for the classifier to classify these images accurately.

3.5.4 Ad source classification

This task deals with the classification of ad sources like social media, newspapers, and Google databases. Leveraging our dataset's comprehensive ad image annotations from online platforms like Facebook, Instagram, Google images, and newspapers, we chose two separate types of grouping of data for our classification tasks. For the first task, we grouped the images of the ads into newspaper ads and web ads, giving a total of 30,978 newspaper ads and 22,428 web ads. For the second task, we grouped the images into three classes Google ads, social media ads (Instagram/Facebook), and newspaper ads. For this task, we have 13,194 images from the Google database, 9,234 images from social media websites, mainly Facebook and Instagram, and 30,978 newspaper images combining both the advert gallery [7] and the e-papers. For both classification tasks, we split the data into training, validation, and testing sets, allocating 70%, 20%, and 10% respectively.

In our dataset, the images are accompanied by annotations indicating their sources. We can broadly categorize the data into two main groups based on their origin. The first group comprises advertisements that were extracted from traditional newspapers. Additionally, this group includes ads obtained from various online platforms, which we refer to as "web ads."

The second grouping encompasses advertisements obtained from multiple sources. It includes newspaper ads as well as advertisements found on social media platforms such as Facebook and Instagram.

Lastly, this group comprises ads that were scraped from the Google database, commonly known as ”Google ads.” These two distinct groupings help us analyze the diverse sources of advertising content within our dataset.

Our chosen architecture consisted of a straightforward non-frozen backbone with a classification head. We had a separate learning rate for each backbone and classification head in order to benefit the training process. For the backbones, ViT-large [31], and ConvNeXT [32] were chosen. Table 3.3 shows the results of models on ad source classification.

The results of the experiment clearly show that our classifier works well in both situations. We noticed that telling apart ads from Instagram and Google databases is harder than distinguishing between web and newspaper ads.

Classification	Metrics	Architecture	
		VIT	Convnext
Two-class classification (Web Ads vs Newspaper Ads)	Accuracy (%)	97.90	96.96
	Precision	0.98	0.96
	Recall	0.97	0.94
	F1 score	0.97	0.95
Three-class classification (Google Ads vs Social Media Ads vs Newspaper Ads)	Accuracy (%)	90.07	90.70
	Precision	0.85	0.87
	Recall	0.84	0.87
	F1 score	0.85	0.87

Table 3.3: Results of ad source classification

Chapter 4

Towards Effective OOD Detection: Exploring Key Influences

So far, we have explored the creation of MAdVerse, along with the various hierarchical loss functions, evaluation metrics, and architectural strategies used for classifying advertisements and related tasks. Advertisements for the same product can vary substantially in their visual elements, such as color palette, branding, and packaging. These differences introduce a challenge: even when trained on a large dataset, a model may struggle to correctly interpret or classify unseen images that deviate from what it has seen before. This variability raises an intriguing question: which specific visual characteristics have the greatest impact on a model's ability to recognize or misclassify unfamiliar inputs? This question is crucial because, in real-world applications, models frequently encounter data that may differ from their training examples.

To address this, we turn to the field of Out-of-Distribution (OOD) detection, which provides a framework for identifying whether a given input is 'seen' or 'unseen' by assigning an OOD score for the input. This score, assessed against a threshold allows us to gauge how novel a particular input is relative to the training data. Our second study focuses on identifying which visual features of an image, such as color or shape, most strongly influence these OOD scores, ultimately helping us understand what makes an input appear unfamiliar to a model.

4.1 Out-of-Distribution (OOD)

Out-of-distribution (OOD) detection is a critical task in machine learning and deep learning, where the goal is to identify inputs that differ significantly from the data used during training. These out-of-distribution samples represent scenarios that a model has not been exposed to, potentially leading to unreliable or incorrect predictions. OOD detection addresses challenges in real-world applications by ensuring models can handle such "novel" inputs more effectively, thereby enhancing the robustness and safety of artificial intelligence (AI) systems, especially in sensitive areas like autonomous driving, healthcare, and financial systems.

The concept of OOD detection emerged in the early 2000s, with one of the foundational works being by Bishop (1994), who proposed a method for monitoring network outputs to identify inputs far from the training distribution [34]. Building on this foundation, Hendrycks and Gimpel [35] introduced a pivotal method to detect OOD samples using the softmax scores from neural networks, marking a modern approach to OOD detection in deep learning models. Their method demonstrated that samples from the training distribution often yield high-confidence softmax outputs, while OOD samples tend to exhibit lower confidence scores, thus providing a straightforward mechanism for OOD detection.

In recent years, significant advancements have been made in this domain, including the development of energy-based models and contrastive learning approaches for better OOD detection performance. This field continues to evolve rapidly as researchers explore innovative strategies to improve the reliability and scalability of OOD detection in various machine learning applications.

4.2 Definitions of OOD

Out-of-Distribution (OOD) detection might initially seem straightforward, as the primary objective is to identify inputs that markedly differ from those seen during training. However, the complexity arises when attempting to define what truly constitutes an OOD instance. Does OOD include entirely new classes, slight variations of existing classes (such as a new breed of dog within a familiar category), shifts in attributes like color or texture, or even small distributional shifts (whether covariate or semantic)? These distinctions lead to important considerations.

For instance, categorizing hand-written alphabets as OOD when a model has been exclusively trained on hand-written numbers intuitively aligns with OOD principles better than a broad definition of OOD as the universal set Ω minus the set of N classes. In other scenarios, a more granular approach might be necessary, where only unseen variations within known classes are considered OOD. The scope of what qualifies as OOD can differ significantly in terms of specificity and difficulty, with no universal rule governing its definition. Classes may be visibly distinct yet deceptively similar, or visually similar despite semantic distinctions, complicating the creation of reliable OOD data splits based solely on class names or hierarchical structures like WordNet.

Building on these considerations, the complexity of Out-of-Distribution (OOD) detection lies in defining what qualifies as truly ”out-of-distribution” in varied real-world contexts. OOD detection doesn't adhere to a singular framework but instead involves several types of distributional shifts, each capturing distinct aspects of unfamiliar data. Researchers have categorized these as "transformed distributions," where known classes are altered (e.g., scaled or rotated), "related distributions," loosely resembling known data but with distinct features, "complement distributions" that introduce entirely unrelated inputs, and "synthetic distributions," which generate artificial unfamiliar patterns [36].

These distinctions illustrate that OOD detection varies not only in type but also in the degree of difficulty, requiring adaptable criteria. These nuanced categories emphasize the need for flexible OOD frameworks that recognize beyond binary narrow range of OOD types. Recognizing OOD data involves assessing both the level of novelty and similarity, making it a dynamic, application-specific process. By accounting for different kinds of unfamiliar data, robust OOD detection systems can more accurately interpret diverse, unseen inputs without sacrificing interpretability or reliability.

4.3 OOD detection methods

In the context of out-of-distribution (OOD) detection, methods can be grouped broadly into categories based on different assumptions and approaches, each addressing specific distributional shifts, such as covariate shifts or semantic shifts. Here's an overview of the types and key themes within the main categories of OOD detection approaches.

4.3.1 Classification-Based Methods

Classification-based methods for OOD detection are commonly used when the OOD data has similar features to the in-distribution (ID) data but lacks corresponding label categories. These methods typically involve modifying the softmax output of a classifier to produce a lower confidence score for OOD samples. Techniques like temperature scaling and input perturbations aim to improve the separability of ID and OOD scores. For instance, Hendrycks and Gimpel [35] proposed using the maximum softmax probability as a simple yet effective indicator of ID and OOD separation. Another enhancement in this category is ODIN [37], which combines temperature scaling and input perturbations to improve separability between ID and OOD samples by modifying the classifier's softmax response. Energy-based models further advance this approach by directly computing the energy of input samples; EBO [38] showed that using an energy score rather than softmax confidence can more accurately differentiate between ID and OOD data, particularly in deep neural networks.

4.3.2 Density-Based Methods

Density-based methods operate on the assumption that ID and OOD samples follow distinct probability distributions. Variational Autoencoders (VAEs) [39] and Generative Adversarial Networks (GANs) [40] have been utilized to learn these distributions, and they assign likelihood values to new samples, with low values indicating OOD. Recent advancements have led to the incorporation of likelihood-ratio tests, improving the effectiveness of these models by comparing conditional likelihoods that better capture OOD characteristics.

4.3.3 Distance-Based Approaches

Distance-based methods measure the distance between test samples and ID data points in feature space. Methods like Mahalanobis distance, which relies on feature embeddings, are popular for their simplicity and effectiveness. These techniques often use pre-trained classifiers to extract embeddings, then apply distance metrics to determine if a test sample falls within the expected distribution. Lee et al. (2018) introduced a technique that combines Mahalanobis distance with softmax scores, enhancing OOD detection performance by capturing nuanced relationships within ID data.

4.3.4 Outlier Exposure

Outlier exposure is a more recent development aimed at enhancing traditional OOD methods by using a small set of known OOD data during training. By training models to recognize these outlier data points, the models can better generalize to unseen OOD samples. Outlier exposure shifts the paradigm of OOD detection by not solely relying on a model's ability to generalize ID data but by directly training on dissimilar data to encourage model robustness. This category has been influential in improving the reliability of OOD detection, especially in safety-critical applications where the cost of undetected OOD samples is high.

Each category tends to address either covariate or semantic shifts with a unique approach to preserving or enhancing the ID classification accuracy while minimizing OOD misclassification. Classification-based and distance-based methods are typically lightweight and can be integrated into existing classifiers. Density-based methods, though often computationally expensive, excel in applications where high confidence in probability estimates is needed. Outlier exposure, meanwhile, is gaining traction due to its proactive approach in handling OOD samples by expanding the training data space.

For more detailed information on OOD methods in specific categories, refer to Yang et al.'s survey [41] on generalized OOD detection, which systematically explores each category and highlights prominent methods.

4.4 Existing OOD Benchmarks

Datasets such as OpenImage-O [42] were developed to overcome problems such as OOD datasets relying on predefined class labels, which can overlap with in-distribution (ID) classes and offer limited coverage. OpenImage-O provides more diverse and realistic OOD examples. Similarly, ImageNet-OOD [43] focuses on reducing covariate shifts and resolving semantic ambiguity by selecting OOD classes that do not overlap with ImageNet-1K, allowing for a more targeted evaluation of semantic shifts. ImageNet-O [44], on the other hand, addresses models' failures to detect OOD data by using adversarial filtering to stress-test models' high-confidence misclassifications. Other benchmarks, such as NINCO [45], tackle the contamination of OOD samples with ID examples, providing a cleaner and more diverse dataset for OOD evaluation. C-OOD [46] introduced a versatile framework for evaluating OOD detection across varying levels of difficulty, addressing the biases of earlier benchmarks.

While these efforts have significantly advanced the field, most of the focus was on improving the quality of a specific type of attribute shift or developing benchmarks with varying OOD difficulty levels. However, there is still a lack of understanding of how OOD methods perform when individual image characteristics, such as color or class, are changed, which we have investigated in this manuscript.

Trends in the evaluation of OOD Methods: The evaluation of out-of-distribution (OOD) detection methods has evolved significantly over time. Initially, OOD detection methods were evaluated on simpler, small-scale datasets with low-resolution images. Common choices for in-distribution (ID) datasets during this early phase included CIFAR-10, CIFAR-100 [47], and SVHN [48]. As for the out-of-distribution (OOD) datasets, selections were often visually distinct and low-resolution datasets such as LSUN [49] (Crop and Resize), Places365 [50] and Textures [51]. While these dataset pairs offered some insight into OOD detection performance, their limitations became increasingly apparent. The ID and OOD datasets were typically quite different in terms of both visual appearance and resolution, often leading to an overestimation of OOD detection performance, and failed to reflect real-world distribution shifts encountered in more complex domains. Recognizing these limitations, more recent methods such as ViM [42] and NNGuide [52] shifted toward using ImageNet-1k [53] as the ID dataset, introducing more realistic scenarios for OOD detection. This shift also brought about the adoption of larger, more challenging OOD datasets such as subsets of ImageNet-21k [54] and iNaturalist [55].

4.5 Motivation and Contribution

While these efforts have significantly propelled advancements in out-of-distribution (OOD) detection, much of the research has primarily concentrated on enhancing specific types of attribute shifts or establishing benchmarks that vary in difficulty levels for OOD tasks. However, a comprehensive understanding remains lacking regarding the performance of OOD methods when individual image characteristics, such as color or class, undergo controlled changes, a gap that this study aims to address.

To bridge this gap, we propose a synthetic approach utilizing two synthetically constructed datasets, SHAPES and CHARS. The SHAPES dataset includes fundamental 2D and 3D geometric figures, such as squares, cubes, and spheres, while CHARS contains alphanumeric characters. Each dataset offers test sets with systematically varied image attributes, enabling a focused investigation. Specifically, our analysis centers on three controlled scenarios: (1) Known classes with novel attributes, where object colors are altered while preserving the original class from the training distribution, modeling a covariate shift; (2) Unseen classes with familiar attributes, where object class is new but color remains constant, representing a semantic shift with visual continuity from the training data; and (3) Completely novel classes and attributes, where both object class and color are unfamiliar to the model, posing the highest level of distributional divergence. Additionally, we introduce image corruption into the test sets to evaluate the robustness of OOD methods against noisy inputs.

By investigating the response of OOD detection methods to these controlled distribution shifts and analyzing their score patterns in corrupted samples, we aim to offer a more nuanced understanding of the conditions under which OOD methods may falter and to evaluate their robustness against minor perturbations, such as noise or visual distortions. This study provides insights into the operational limits of OOD methods and their sensitivity to subtle input variations, thus contributing to the development of more resilient models.

Chapter 5

Probing OOD methods with synthetic datasets: SHAPES and CHARS

We introduce two new synthetic datasets, SHAPES and CHARS. SHAPES includes basic 2D and 3D geometric primitives, such as squares, cubes, and spheres, whereas the CHARS dataset comprises alphanumeric characters. To create samples in both datasets, a consistent process is followed. Each dataset has its own configurations specifying attributes such as color, rotation, size range, and other dataset-specific properties. These configurations also define the number of samples for train, validation, ID (in-distribution) test, and 3 OOD (out-of-distribution) splits (see Figure 5.1). The attribute values for ID and OOD are disjoint sets, ensuring clear separation between in-distribution and out-of-distribution samples. The background color for all images remains the same across all samples.

The three OOD splits are defined as follows:

- *OOD in color*: Images have their colors sampled from OOD colors, while their classes remain a subset of ID classes.
- *OOD in class*: Images use ID colors but belong to completely unseen classes not present in the training set.
- *OOD across both*: Configurations are generated by sampling entirely from OOD attributes, meaning both class and color are OOD.

The configuration generator pre-generates all configurations required for train, validation, and test splits with a fixed random seed, ensuring reproducibility and consistency.These configurations are created in a single step and only once, guaranteeing that the test sets remain fixed. Images are rendered dynamically in the dataloader, which fetches the necessary configurations for each batch and renders the images using moderngl [56]. For test splits, a specified percentage of images, as defined in the configuration, are pre-assigned with corruption details during the configuration generation phase. The corruption details include a corruption method and a severity level (either 1 or 2), selected from one of the ten common image corruption strategies mentioned in [57].

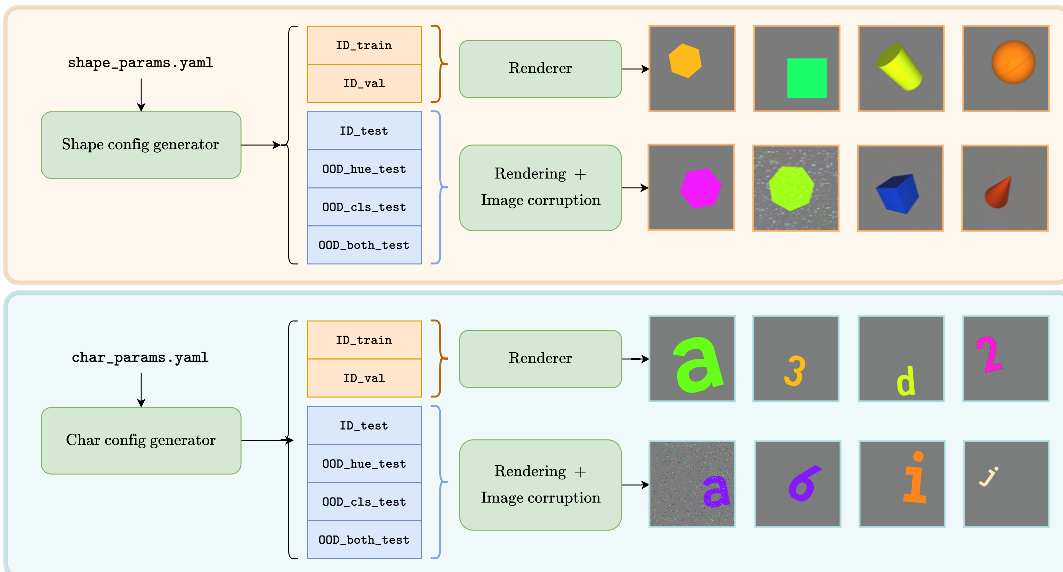

Figure 5.1: Dataset generation pipeline for SHAPES and CHARS datasets.

5.1 Image Attribute Specifications

Color:

Colors are in HSV format. The background is set to [0, 0, 155] for both SHAPES and CHARS datasets, with shared in-distribution and out-of-distribution hues:

```
color:
  id_hues: [30, 45, 60, 75, 90, 105, 120, 135, 150]
  ood_hues: [210, 225, 240, 255, 270, 285, 300, 315, 330]
  bg_color: [0,0,155]
```

Classes:

Both datasets have in-distribution (ID) and out-of-distribution (OOD) classes. For SHAPES, 8 ID and 9 OOD shapes are defined; for CHARS, 10 ID and 10 OOD characters. SHAPES ID-OOD classes are conceptually related, while CHARS ID and OOD classes use different characters.

```
shape:
id_shapes: [circle, square, rp6, rp8, eq_triangle, sphere_3d, cube_3d, cylinder_3d]
  ood_shapes: [ellipse, rectangle, rp7, rp9, is_triangle, random, ellipsoid_3d,
  cuboid_3d, cone_3d]

chars:
  id_chars: [a, b, c, d, e, 0, 1, 2, 3, 4]
  ood_chars: [f, g, h, i, j, 5, 6, 7, 8, 9]
```

Note: Here `rp` means *Regular polygon*

Size, Rotation, and Parameters:

SHAPES and CHARS define size and rotation ranges. Size bounds are set as a percentage range $[35, 55]$ of the image dimension; font size for CHARS ranges from 60 to 150. Rotation is specified for 2D and 3D objects:

```
size_bounds: [35,55]
font_size_min_max: [60,150]

shapes:
  rot_min_max_2d: [-180,180]
  rot_min_max_3d: [-60,60]
  step_angle: 10

chars:
  rot_min_max: [-60,60]
  step_angle: 5
```

Additional parameters for synthetic images include point counts for random shapes and lighting for 3D shapes:

```
shape:
  rnd_shape:
    num_points: [5-12]
    min_smoothness: 30
```

```
      max_smoothness: 80
      min_radius_mult: 0.7
      max_radius_mult: 1.3

    solid_shape_params:
      ambient_strength_bounds: [0.25, 0.5]
      specular_strength_bounds: [0.15, 0.3]
      light_pos_offset: 80
```

Dataset Splits:

Three seeds (1, 2, 3) are used to create splits. Each split has the following image counts:

```
imgs_per_split:
  train: 100000
  val: 5000
  test:
    id: 5000
    ood_hue: 5000
    ood_cls: 5000
    ood_both: 5000
```

Corruption Methods:

We apply 10 corruption strategies (30

```
corruption_methods = [ "gaussian_noise", "shot_noise", "impulse_noise",
"speckle_noise", "gaussian_blur", "glass_blur",
"spatter", "contrast", "brightness",
"saturate"]
```

5.2 Synthetic image generation

This is a simplified overview of the entire process shown in Figure 5.1, from generating image configurations to rendering the final images.

Algorithm 1 Dataset Configuration and Rendering

Input: Dataset attributes, and a random seed for reproducibility **Output:** Configuration files for dataset splits, Rendered images

Step 1: Read image attribute configurations and set the random seed.

Step 2: For each image:

- Sample attributes such as rotation, class, color, size, and other additional parameters using appropriate random sampling methods (e.g., `random.choices`)
- Calculate margins for movement in X and Y directions based on the size and rotation of the image.
- Sample random offsets in the X and Y directions from center, to place the shape/character.

Step 3: If the image belongs to the OOD test split:

- Add additional OOD-related information, such as the type of OOD and image corruption strategies, to the configuration.

Step 4: Save all configuration files for each dataset split (Train, Validation, ID Test, OOD Test). These files will be used in the rendering process.

Rendering Process:

- Using `moderngl`'s headless contexts, configurations from dataloader's collate function are sent to respective graphical contexts.
- Images are rendered based on these configurations and are returned as batches.

5.3 Experiments and analysis

5.3.1 Experimental setup

Problem Setup:. Let $\mathcal{D}_{\text{in}} = \{(x_i, y_i); x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$ represent the In-distribution data (i.e., data the model is trained on) sampled from distribution $P_{\text{in}}(x, y)$, where $x \in \mathbb{R}^d$ is the input image and y is the corresponding label. The Out-of-distribution data $\mathcal{D}_{\text{out}}$ comes from a different distribution $P_{\text{out}}(x, y)$ which is not seen during training.

Given a classifier $f : \mathbb{R}^d \rightarrow \mathbb{R}^N$ trained on $\mathcal{D}_{\text{in}}$ that classifies input data to N In-distribution classes, the goal of Out-of-distribution detection is to design a scoring function $S(x)$ that helps in distinguishing between in-distribution data $\mathcal{D}_{\text{in}}$ and out-of-distribution data $\mathcal{D}_{\text{out}}$. The decision is made on a threshold τ, where:

$$S(x) = \begin{cases} \text{In-Distribution}, & \text{if } S(x) \geq \tau, \\ \text{Out-of-Distribution}, & \text{if} S(x) < \tau. \end{cases}$$

We evaluate the OOD methods under three types of OOD scenarios: 1) where only the color of the shape/character changes 2) where only the class changes, and 3) where both color and class change. We will use the terms 'OOD in color,' 'OOD in class,' and 'OOD in both' as shorthand for addressing these OOD types.

Datasets: We prepare the SHAPES and CHARS datasets by setting a random seed to generate 125,000 image configurations for training, validation, ID test and three OOD test splits (OOD in color, class, and both). Across all four test splits (one in-distribution and three OOD) in both datasets, A portion of the images is corrupted using one of ten corruption methods applied to each image at a specific severity level (either 1 or 2). To ensure consistency, we repeat the training and evaluations using three random seeds.

Backbones: We select 10 architectures combined across the ResNet [58], DenseNet [59], Vision Transformer (ViT) [9], and Wide-ResNet [60] families, each with a single linear layer as the classification head. The output dimension of the classification head corresponds to the number of in-distribution (ID) classes for each dataset. All models are trained independently from scratch on both datasets.

OOD methods: We evaluate 13 OOD detection methods comprising of logit-based, feature-based and energy-based methods across the three OOD scenarios. Logit-based methods include ODIN [37], MaxLogit [61], MSP [35], and ViM [42], all of which operate directly on logits or modify them to compute OOD scores. Feature-based methods include SCALE [62], SHE [63], GradNorm [64], KNN [65], and NNGuide [52], which work on feature representations, typically from the penultimate layer. Lastly, energy-based methods, include EBO [38], GEN [66], ASH [67], and ReAct [68], which calculate an energy score derived from logits or modified activation's. All OOD detection methods are implemented using the OpenOOD framework laid out by [69].

Evaluation Metrics: We evaluate the OOD-detection performance using the commonly used metric AUROC. It represents the probability that a positive example receives a higher detection score than a negative example [70]. Higher value indicates better detection performance. AUROC values are calculated for each dataset, for each type of OOD scenario, both with and without image corruption. The reported AUROCs correspond to the median AUROC across the three seeds. The observed absolute deviation from the median (MAD) for AUROC across the three seeds for all OOD methods and backbone combinations was in the order of 10^{-2}.

We use the *Overlap Coefficient* (eq-5.1) to measure the overlap between the smoothed densities of min-max normalized OOD scores of ID and OOD test samples.

$$\text{overlap}(A, B) = |A \cap B|, \quad 0 \leq \text{overlap}(A, B) \leq 1. \tag{5.1}$$

The set notations used in Equation 5.1 are for the sake of brevity.

5.3.2 Sensitivity of OOD methods to color

Tables 5.1, 5.2, 5.3 and 5.4 presents the AUROC scores for un-corrupted test images across all combinations of OOD methods and architectures on both datasets in two OOD scenarios: OOD in color and OOD in class. Except for methods such as KNN, React, and ViM, other methods perform poorly, with AUROC as low as 0.01, which is worse than random coin flip. The reason for this can be seen in Figure 5.3, where OOD samples are given higher scores than ID samples in 'OOD in color' scenario. This observation is quite opposite to the intended behavior of OOD detection methods, where ID samples should have received higher scores than OOD samples.

The AUROC values in the scenario when both color and class change, is nearly identical to that of the 'OOD in color' scenario and the results are presented in Figure 5.2. It can again be seen in Figure 5.3, where the score distributions for OOD in color and OOD in both color and class scenarios remain similar. This further reinforces the evidence that OOD detection methods are highly sensitive to changes in visual attributes like color. Further among the selected architectures, we observe that ViT performs the best across all OOD methods and amongst the 13 OOD methods, KNN and ViM are robust across all the architectures (Tables 5.1, 5.2, 5.3 and 5.4).

5.3.3 Impact of Image Corruption

To assess the impact of image corruption on OOD scores for both ID and OOD samples, we first extract the OOD scores for corrupted ID and OOD test sets from all OOD methods, across all architectures, and apply min-max normalization such that the relative order of score distributions and scales

	D.121	D.169	D.201	R.101d	R.200d	R.50	ViT_B	ViT_S	WR.101	WR.50
ASH	0.01	0.02	0.04	0.29	0.45	0.13	0.66	0.89	0.08	0.09
EBO	0.01	0.03	0.08	0.29	0.42	0.04	0.66	0.93	0.10	0.04
GEN	0.01	0.03	0.20	0.36	0.40	0.07	0.54	0.92	0.22	0.05
GRADNORM	0.00	0.01	0.01	0.14	0.09	0.05	0.63	0.54	0.02	0.03
KNN	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00
MAXLOGIT	0.01	0.03	0.08	0.29	0.42	0.04	0.65	0.92	0.10	0.04
MSP	0.01	0.10	0.74	0.48	0.50	0.18	0.24	0.90	0.37	0.11
NNGUIDE	0.01	0.04	0.17	0.35	0.48	0.06	0.70	0.95	0.16	0.08
ODIN	0.01	0.03	0.06	0.20	0.16	0.17	0.65	0.84	0.09	0.11
REACT	1.00	0.99	1.00	0.94	0.96	0.20	0.65	0.93	0.93	0.28
SCALE	0.01	0.03	0.11	0.52	0.52	0.41	0.65	0.92	0.26	0.35
SHE	0.00	0.03	0.04	0.20	0.23	0.07	0.99	0.98	0.06	0.06
VIM	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 5.1: AUROC of OOD detection methods across all architectures on uncorrupted test images. *Dataset*: SHAPES, *scenario*: color is OOD

	D.121	D.169	D.201	R.101d	R.200d	R.50	ViT_B	ViT_S	WR.101	WR.50
ASH	0.57	0.60	0.61	0.63	0.62	0.62	0.51	0.54	0.58	0.59
EBO	0.64	0.65	0.65	0.69	0.67	0.65	0.63	0.59	0.62	0.64
GEN	0.63	0.65	0.65	0.69	0.66	0.64	0.63	0.60	0.63	0.63
GRADNORM	0.55	0.56	0.56	0.58	0.56	0.57	0.58	0.56	0.53	0.57
KNN	0.74	0.74	0.75	0.75	0.75	0.71	0.65	0.66	0.71	0.72
MAXLOGIT	0.64	0.66	0.65	0.69	0.67	0.65	0.64	0.59	0.62	0.64
MSP	0.64	0.64	0.67	0.69	0.67	0.63	0.61	0.61	0.63	0.62
NNGUIDE	0.62	0.65	0.65	0.67	0.65	0.64	0.63	0.58	0.59	0.64
ODIN	0.62	0.64	0.64	0.51	0.46	0.52	0.56	0.50	0.52	0.54
REACT	0.65	0.65	0.66	0.72	0.70	0.66	0.63	0.59	0.66	0.66
SCALE	0.55	0.56	0.55	0.62	0.61	0.62	0.63	0.59	0.60	0.62
SHE	0.56	0.58	0.56	0.59	0.57	0.54	0.59	0.60	0.51	0.55
VIM	0.77	0.77	0.77	0.72	0.72	0.66	0.63	0.63	0.65	0.67

Table 5.2: AUROC of OOD detection methods across all architectures on uncorrupted test images. *Dataset*: SHAPES, *scenario*: class is OOD

	D.121	D.169	D.201	R.101d	R.200d	R.50	ViT_B	ViT_S	WR.101	WR.50
ASH	0.09	0.32	0.07	0.07	0.06	0.19	0.65	0.68	0.18	0.13
EBO	0.13	0.42	0.09	0.15	0.16	0.25	0.88	0.81	0.25	0.15
GEN	0.18	0.54	0.19	0.32	0.27	0.40	0.89	0.76	0.58	0.17
GRADNORM	0.04	0.07	0.02	0.03	0.03	0.09	0.94	0.64	0.09	0.09
KNN	1.00	1.00	1.00	1.00	1.00	1.00	0.79	0.75	1.00	1.00
MAXLOGIT	0.13	0.43	0.10	0.16	0.17	0.25	0.88	0.81	0.25	0.15
MSP	0.79	0.82	0.24	0.73	0.63	0.79	0.87	0.77	0.96	0.29
NNGUIDE	0.23	0.61	0.19	0.23	0.25	0.29	0.81	0.79	0.37	0.20
ODIN	0.11	0.35	0.08	0.13	0.16	0.22	0.94	0.88	0.25	0.13
REACT	0.96	0.98	0.97	0.97	0.95	0.99	0.87	0.73	1.00	0.99
SCALE	0.09	0.27	0.08	0.10	0.08	0.38	0.88	0.80	0.29	0.24
SHE	0.09	0.24	0.05	0.05	0.06	0.14	0.69	0.71	0.18	0.13
VIM	1.00	1.00	1.00	1.00	1.00	1.00	0.76	0.83	1.00	1.00

Table 5.3: AUROC of OOD detection methods across all architectures on uncorrupted test images. *Dataset*: CHARS, *scenario*: color is OOD

	D.121	D.169	D.201	R.101d	R.200d	R.50	ViT_B	ViT_S	WR.101	WR.50
ASH	0.83	0.87	0.88	0.75	0.75	0.77	0.49	0.47	0.72	0.80
EBO	0.88	0.91	0.93	0.92	0.93	0.91	0.91	0.93	0.86	0.94
GEN	0.90	0.93	0.94	0.94	0.95	0.92	0.91	0.94	0.88	0.95
GRADNORM	0.83	0.85	0.85	0.64	0.66	0.69	0.83	0.82	0.64	0.71
KNN	0.99	0.99	0.99	1.00	1.00	0.99	0.93	0.97	0.98	0.98
MAXLOGIT	0.88	0.91	0.93	0.92	0.93	0.91	0.91	0.93	0.86	0.94
MSP	0.91	0.93	0.95	0.96	0.97	0.95	0.90	0.93	0.92	0.96
NNGUIDE	0.88	0.88	0.92	0.90	0.93	0.90	0.91	0.93	0.86	0.94
ODIN	0.88	0.92	0.92	0.45	0.51	0.80	0.90	0.90	0.68	0.77
REACT	0.95	0.94	0.95	0.95	0.95	0.93	0.92	0.94	0.90	0.94
SCALE	0.79	0.70	0.73	0.56	0.59	0.57	0.91	0.93	0.56	0.59
SHE	0.90	0.94	0.89	0.73	0.77	0.76	0.89	0.95	0.77	0.82
VIM	1.00	1.00	1.00	0.99	1.00	0.96	0.92	0.96	0.96	0.95

Table 5.4: AUROC of OOD detection methods across all architectures on uncorrupted test images. *Dataset*: CHARS, *scenario*: class is OOD

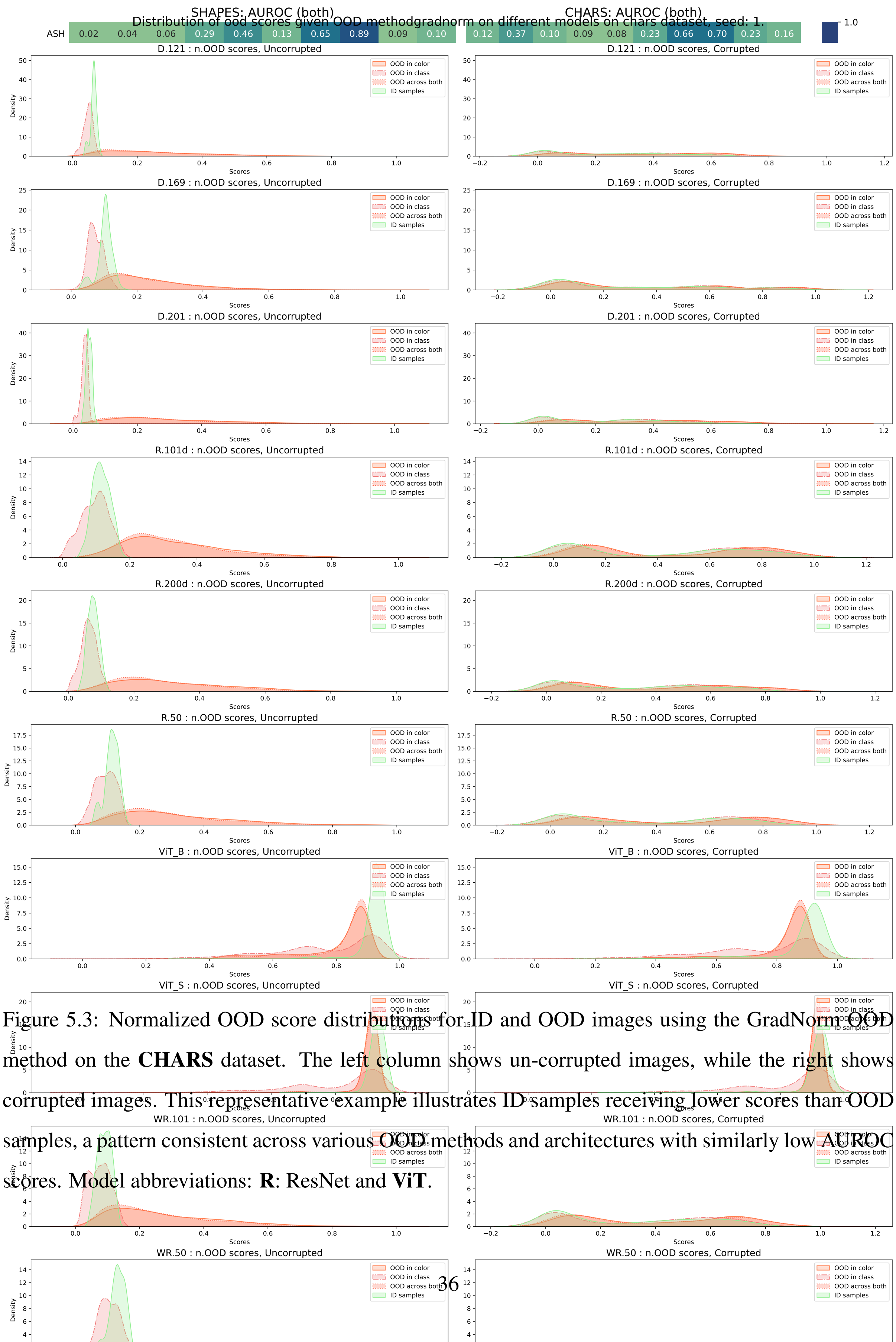

Figure 5.3: Normalized OOD score distributions for ID and OOD images using the GradNorm OOD method on the **CHARS** dataset. The left column shows un-corrupted images, while the right shows corrupted images. This representative example illustrates ID samples receiving lower scores than OOD samples, a pattern consistent across various OOD methods and architectures with similarly low AUROC scores. Model abbreviations: **R**: ResNet and **ViT**.

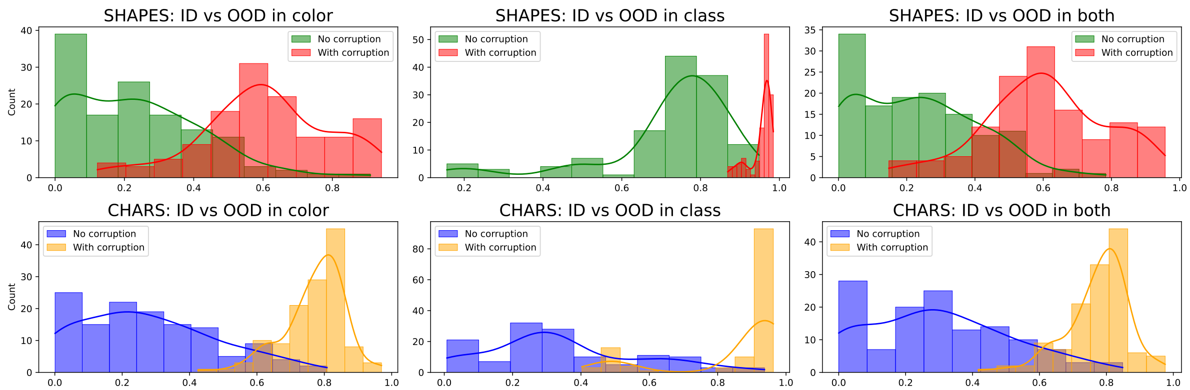

Figure 5.4: Histograms of overlap coefficients between ID and OOD score distributions, measured across all combinations of OOD methods and models, with two rows for the datasets SHAPES(top) and CHARS(bottom), with each row containing 3 subplots for different OOD test types: color, shape, and both. Each subplot compares overlap coefficients with and without image corruption.

are preserved. Then using the overlap coefficient 5.1, we measure the overlap between the smoothed densities of normalized OOD scores of ID and OOD test samples. Figure 5.4 shows the histogram of 130 such overlap coefficients obtained by all the OOD method and backbone combinations. Intuitively, when there is no corruption, the overlap coefficients across all the OOD methods and architectures should be relatively lower, indicating the ability of OOD methods to assign a higher score to the ID samples. But with image corruption, one might expect a higher overlap given the poor performance of OOD methods. Henceforth pointing towards a conjecture that, an ID corrupted image is as bad as an OOD image (with or without corruption). We precisely corroborate this intuition in Figure 5.4, where we find that overlap coefficients across the OOD method and backbone combinations increase in the presence of image corruption and significant in the case of OOD in class.

As seen in the above tables (5.1, 5.2, 5.3, 5.4), KNN and ViM remain robust OOD methods and ViT, the best amongst the chosen architectures. Though we can observe a slight increase in the AUROC for OOD in color and OOD in both color and class cases (relative to the tables and Figure 5.2 respectively), there is a decrease in AUROC for OOD in color scenario. These observations can be attributed to the same line of observation we made in Figure 5.4 that an ID corrupted image is as bad as an OOD image, which inflates or shrinks the AUROC values which suit a random coin toss.

Chapter 6

Out-of-distribution detection in Ads

Building on our hierarchical ad classification work, and leveraging two synthetic datasets to investigate how specific image factors influence OOD detection performance, we now proceed to apply these methods to the task of OOD detection in ads, specifically at the leaf level of the hierarchy. Here, both the category and subcategory are known, with only the brand remaining unseen—making this a particularly challenging application within the diverse landscape of digital advertising.

6.1 Ads-OOD dataset preparation

To prepare the datasets for out-of-distribution (OOD) detection in advertisements, we first selected a diverse set of countries based on their annual digital advertisement spending, ranked in descending order. This selection aimed to represent a wide array of advertising styles and spending power from various regions worldwide. The countries chosen include the USA, Canada, Mexico, Brazil, Argentina, Egypt, South Africa, Morocco, Japan, South Korea, Germany, the United Kingdom, France, Italy, Spain, Russia, Australia, Saudi Arabia, and the UAE. From each of these countries, we then identified brands within various sub-categories that demonstrated a strong online presence, particularly on widely-used platforms such as Facebook.

For each selected brand, we scraped 300 advertisement images from their official Facebook pages to ensure a substantial representation of each brand's visual style and content. After collecting this initial set of ad images, a manual inspection was performed to confirm that the images met the study's definition and standards for ads. This review process was essential for filtering out irrelevant or low-quality images, resulting in a refined dataset comprising 41 brands, and a total of 12281 images.

The final test set consists of two components: the Out-of-Distribution (OOD) set, comprising the previously mentioned scraped data from unseen brands, and the In-Distribution (ID) set, represented by MAdVerse's ID test set. *Brands in these two subsets are mutually exclusive*. The chosen brands offer a strong basis for evaluating OOD detection in advertisements, encompassing a diverse range of global advertising styles and strategies.

6.2 Results

	ASH	EBO	GEN	GRADNORM	KNN	MAXLOGIT	MSP	NNGUIDE	REACT	SCALE	SHE	VIM
Semantic Embedding (Leaf only)	0.40	0.76	0.34	0.74	0.71	0.81	0.81	0.37	0.66	0.66	0.73	0.45
Soft Labels (Leaf only)	0.76	0.82	0.83	0.61	0.80	0.84	0.87	0.82	0.78	0.81	0.64	0.83
Sum CE Multi (fusion)	0.69	0.85	0.87	0.71	0.80	0.86	0.85	0.85	0.68	0.78	0.68	0.84
Sum CE Multi (No fusion)	0.79	0.79	0.87	0.75	0.82	0.81	0.84	0.79	0.76	0.78	0.75	0.80

Table 6.1: AUROC plot 12 OOD methods (X) vs top 4 performing hierarchical classifiers with BLIP v2 as backbone, and their respective loss and architecture used (Y)

Table 6.1 shows that most of the OOD method and model combinations achieve good AUROC scores, with only a few OOD methods exhibiting lower AUROC values when the hierarchical classifier is leaf-only with Semantic Embedding technique. To evaluate and compare Out-of-Distribution (OOD) detection methods and various loss-architecture combinations, we employ consensus ranking. Consensus ranking aggregates rankings from multiple sources, yielding a single averaged ranking that reflects overall performance, where a smaller rank indicates better performance. From Tables 6.3 and 6.2, it is evident that simpler logit-based OOD methods, such as MaxLogit and MSP, consistently achieve the highest ranks. Additionally, a straightforward multi-level architecture, which does not incorporate feature fusion and utilizes a simple loss function (Sum CE) that sums cross-entropy losses from all levels, achieves a superior consensus ranking.

Interestingly, the rankings reveal a grouping pattern among the different types of OOD methods. Logit-based methods like MaxLogit and MSP secure the top ranks, followed by entropy and energy-based methods such as GEN and EBO. Next in the ranking are distance-based methods like KNN and NNguide, while more complex activation and gradient-based methods, such as ASH and REACT, are positioned lower. Furthermore, the multilevel architectures occupy the top two positions, while leaf-only architectures are ranked at the bottom.

Loss Type	Consensus Ranking
Sum CE (Multi (No Fusion))	1.0
Sum CE (Multi (Fusion))	2.5
Soft Labels (leaf only)	2.5
Semantic Emb. (leaf only)	4.0

Table 6.2: Consensus ranking of hierarchical losses based on AUROC values across different OOD methods.

OOD Method	Consensus Ranking
msp	2.25
maxlogit	2.50
ebo	4.25
gen	4.25
knn	6.00
vim	6.00
nnguide	7.00
scale	8.00
gradnorm	9.00
ash	9.50
react	9.50
she	9.75

Table 6.3: Consensus ranking of OOD methods based on AUROC values.

Chapter 7

Conclusion

In this thesis, we addressed the critical gap in existing datasets for advertising analysis, where diversity in advertisement styles, semantic groupings, and granular annotation levels was lacking. To fill this void, we introduced MAdVerse—a large, multilingual dataset of over 50,000 advertisements from diverse sources like the web, social media, and digital newspapers. This dataset is organized into a hierarchical structure, ranging from broad categories to fine distinctions, and is annotated with essential information, including ad source, type (product or brand), language, and hierarchical labels. MAdVerse thus provides a comprehensive foundation for studying ads across linguistic and cultural contexts, supporting multiple advanced advertising analysis tasks such as hierarchical classification, source identification, and inducing hierarchy in ads.

We also investigated the factors influencing out-of-distribution (OOD) detection, focusing on which specific visual attributes, such as color or shape, significantly impact detection performance. To address this, we introduced two synthetic datasets, SHAPES and CHARS, enabling controlled experiments by isolating visual variables to assess how different OOD detection methods handle unseen data. Our findings reveal that OOD detection models are highly sensitive to changes in attributes like color, suggesting that existing models struggle with subtle shifts in visual distribution. By further introducing image corruption, we gained insights into the robustness of these models under real-world conditions. These findings set the foundation for applying OOD detection in advertising, allowing us to assess the robustness of models in recognizing unseen or unexpected ads.

By combining the contributions of MAdVerse with a focused investigation into the factors affecting OOD detection, this thesis provides a dual advancement: (1) a versatile, linguistically and visually diverse dataset for advertisement analysis, and (2) a benchmark that highlights the limitations of current OOD detection methods under specific shifts in visual attributes.

In practice, classifiers trained on MAdVerse can serve as ad-categorization models for large-scale advertising repositories, aiding in efficient sorting and labeling of new ads. Moreover, our OOD detection work offers valuable insights into how different attributes of an image affect detection scores, establishing a benchmark for evaluating and refining new OOD methods. Together, these contributions support deeper insights into advertising and foster innovation in building more robust, adaptable advertising models for real-world applications.

Bibliography

[1] "Investment in advertising," https://www.theworldcounts.com/economies/global/spending-on-advertising/. 1

[2] Z. Hussain, M. Zhang, X. Zhang, K. Ye, C. Thomas, Z. Agha, N. Ong, and A. Kovashka, "Automatic understanding of image and video advertisements," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1100–1110. 3, 4, 9, 17

[3] C. Fosco, V. Casser, A. K. Bedi, P. O'Donovan, A. Hertzmann, and Z. Bylinskii, "Predicting visual importance across graphic design types," in *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 249–260. [Online]. Available: https://doi.org/10.1145/3379337.3415825 3, 4, 9

[4] S. Liang, R. Liu, and J. Qian, "Fixation prediction for advertising images: Dataset and benchmark," *Journal of Visual Communication and Image Representation*, vol. 81, p. 103356, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047320321002339 3, 4, 9

[5] Y. Kumar, R. Jha, A. Gupta, M. Aggarwal, A. Garg, T. Malyan, A. Bhardwaj, R. Ratn Shah, B. Krishnamurthy, and C. Chen, "Persuasion strategies in advertisements," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, pp. 57–66, Jun. 2023. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/25076 3, 4, 9

[6] Y. Cao, Y. Ma, M. Zhou, C. Liu, H. Xie, T. Ge, and Y. Jiang, "Geometry aligned variational transformer for image-conditioned layout generation," in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1561–1571. [Online]. Available: https://doi.org/10.1145/3503161.3548332 3, 4

[7] "Advert gallery," https://www.advertgallery.com/. 5, 7, 19

[8] T. Jain, C. Lennan, Z. John, and D. Tran, "Imagededup," https://github.com/idealo/imagededup, 2019. 6

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. 6, 32

[10] "Apify," https://apify.com/. 6

[11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755. 7

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf 7

[13] B. C. G. Lee, J. Mears, E. Jakeway, M. Ferriter, C. Adams, N. Yarasavage, D. Thomas, K. Zwaard, and D. S. Weld, "The newspaper navigator dataset: Extracting and analyzing visual content from 16 million historic newspaper pages in chronicling america," *CoRR*, vol. abs/2005.01583, 2020. [Online]. Available: https://arxiv.org/abs/2005.01583 7

[14] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam, "Large-scale object classification using label relation graphs," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 48–64. 13

[15] Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, and Y. Yu, "Hd-cnn: Hierarchical deep convolutional neural networks for large scale visual recognition," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2740–2748. 13

[16] X. Zhu and M. Bain, "B-CNN: branch convolutional neural network for hierarchical classification," *CoRR*, vol. abs/1709.09890, 2017. [Online]. Available: http://arxiv.org/abs/1709.09890 13, 14

[17] Y. Seo and K. shik Shin, "Hierarchical convolutional neural networks for fashion image classification," *Expert Systems with Applications*, vol. 116, pp. 328–339, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417418305992 13, 14

[18] X. Li, Y. Zhou, Y. Zhou, and W. Wang, "Mmf: Multi-task multi-structure fusion for hierarchical image classification," in *Artificial Neural Networks and Machine Learning – ICANN 2021*, I. Farkaš, P. Masulli, S. Otte, and S. Wermter, Eds. Cham: Springer International Publishing, 2021, pp. 61–73. 13, 14

[19] J. Chen, P. Wang, J. Liu, and Y. Qian, "Label relation graphs enhanced hierarchical residual network for hierarchical multi-granularity classification," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4848–4857. 13, 14

[20] B. Barz and J. Denzler, "Hierarchy-based image embeddings for semantic image retrieval," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Los Alamitos, CA, USA: IEEE Computer Society, jan 2019, pp. 638–647. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/WACV.2019.00073 13, 14, 16, 17

[21] L. Bertinetto, R. Mueller, K. Tertikas, S. Samangooei, and N. A. Lord, "Making better mistakes: Leveraging class hierarchies with deep networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2020, pp. 12 503–12 512. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.01252 13, 14, 16, 17

[22] D. Chang, K. Pang, Y. Zheng, Z. Ma, Y. Song, and J. Guo, "Your "flamingo" is my "bird": Fine-grained, or not," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2021, pp. 11 471–11 480. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.01131 13, 14

[23] G. A. Miller, "WordNet: A lexical database for English," in *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994. [Online]. Available: https://aclanthology.org/H94-1111 14

[24] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012. 14

[25] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *Tech Report*, 2009. 14

[26] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *CoRR*, vol. abs/1708.07747, 2017. [Online]. Available: http://arxiv.org/abs/1708.07747 14

[27] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011. 14

[28] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *2013 IEEE International Conference on Computer Vision Workshops*, 2013, pp. 554–561. 14

[29] P. Wang, J. Chen, and Y. Qian, "Semantic guided level-category hybrid prediction network for hierarchical image classification," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 21, no. 06, p. 2350023, 2023. [Online]. Available: https://doi.org/10.1142/S0219691323500236 14

[30] Y. Ge, S. Li, X. Li, F. Fan, W. Xie, J. You, and X. Liu, "Embedding semantic hierarchy in discrete optimal transport for risk minimization," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 2835–2839. 14, 16

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy 15, 20

[32] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los

Alamitos, CA, USA: IEEE Computer Society, jun 2022, pp. 11 966–11 976. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01167 15, 20

[33] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models," in *ICML*, 2023. 15

[34] C. M. Bishop, "Novelty detection and neural network validation," *IEE Proceedings-Vision, Image and Signal Processing*, vol. 141, no. 4, pp. 217–222, 1994. 22

[35] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *Proceedings of International Conference on Learning Representations*, 2017. 22, 23, 32

[36] S. Farquhar and Y. Gal, "What 'out-of-distribution' is and is not," in *NeurIPS ML Safety Workshop*, 2022. [Online]. Available: https://openreview.net/forum?id=XCS_zBHQA2i 23

[37] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=H1VGkIxRZ 23, 32

[38] W. Liu, X. Wang, J. D. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20, 2020. 23, 32

[39] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2022. [Online]. Available: https://arxiv.org/abs/1312.6114 24

[40] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 24

[41] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2110.11334 25

[42] H. Wang, Z. Li, L. Feng, and W. Zhang, "Vim: Out-of-distribution with virtual-logit matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 25, 26, 32

[43] W. Yang, B. Zhang, and O. Russakovsky, "Imagenet-ood: Deciphering modern out-of-distribution detection algorithms," in *International Conference on Learning Representations (ICLR)*, 2024. [Online]. Available: https://openreview.net/forum?id=VTYg5ykEGS 25

[44] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," *CVPR*, 2021. 25

[45] J. Bitterwolf, M. Mueller, and M. Hein, "In or out? fixing imagenet out-of-distribution detection evaluation," in *ICML*, 2023. [Online]. Available: https://proceedings.mlr.press/v202/bitterwolf23a.html 25

[46] I. Galil, M. Dabbah, and R. El-Yaniv, "A framework for benchmarking class-out-of-distribution detection and its application to imagenet," in *International Conference on Learning Representations (ICLR)*, 2023. [Online]. Available: https://openreview.net/forum?id=Iuubb9W6Jtk 25

[47] A. Krizhevsky, "Learning multiple layers of features from tiny images," University of Toronto, Toronto, Ontario, Canada, Tech. Rep., 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf 25

[48] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. [Online]. Available: http://ufldl.stanford.edu/housenumbers 25

[49] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015. [Online]. Available: https://arxiv.org/abs/1506.03365 25

[50] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. [Online]. Available: http://places.csail.mit.edu 25

[51] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. [Online]. Available: https://www.robots.ox.ac.uk/~vgg/data/dtd/ 25

[52] J. Park, Y. G. Jung, and A. B. J. Teoh, "Nearest neighbor guidance for out-of-distribution detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1686–1695. 26, 32

[53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009. [Online]. Available: https://image-net.org 26

[54] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Imagenet-21k pretraining for the masses," *arXiv preprint arXiv:2104.10972*, 2021. [Online]. Available: https://arxiv.org/abs/2104.10972 26

[55] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. [Online]. Available: https://www.inaturalist.org 26

[56] S. Dombi. (2020) Moderngl, high performance python bindings for opengl 3.3+. https://github.com/moderngl/moderngl. 27

[57] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=HJz6tiCqYm 28

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE, 2016. 32

[59] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017. 32

[60] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proceedings of the British Machine Vision Conference (BMVC)*, vol. 2016, 2016, p. 87. 32

[61] D. Hendrycks, S. Basart, M. Mazeika, A. Zou, M. Mostajabi, J. Steinhardt, and D. X. Song, "Scaling out-of-distribution detection for real-world settings," in *International Conference*

on Machine Learning, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:227407829 32

[62] K. Xu, R. Chen, G. Franchi, and A. Yao, "Scaling for training time and post-hoc out-of-distribution detection enhancement," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=RDSTjtnqCg 32

[63] J. Zhang, Q. Fu, X. Chen, L. Du, Z. Li, G. Wang, xiaoguang Liu, S. Han, and D. Zhang, "Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=KkazG4lgKL 32

[64] R. Huang, A. Geng, and Y. Li, "On the importance of gradients for detecting distributional shifts in the wild," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, ser. NIPS '21, 2021. 32

[65] Y. Sun, Y. Ming, X. Zhu, and Y. Li, "Out-of-distribution detection with deep nearest neighbors," *ICML*, 2022. 32

[66] X. Liu, Y. Lochman, and C. Zach, "Gen: Pushing the limits of softmax-based out-of-distribution detection," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 23 946–23 955. 32

[67] A. Djurisic, N. Bozanic, A. Ashok, and R. Liu, "Extremely simple activation shaping for out-of-distribution detection," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=ndYXTEL6cZz 32

[68] Y. Sun, C. Guo, and Y. Li, "React: Out-of-distribution detection with rectified activations," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://openreview.net/forum?id=IBVBtz_sRSm 32

[69] J. Zhang, J. Yang, P. Wang, H. Wang, Y. Lin, H. Zhang, Y. Sun, X. Du, Y. Li, Z. Liu, Y. Chen, and H. Li, "Openood v1.5: Enhanced benchmark for out-of-distribution detection," *arXiv preprint arXiv:2306.09301*, 2023. 32

[70] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006. 32