Temporal Dynamics in Visual Data: Analyzing the Impact of Time on Classification Accuracy

Tom Pégeot[*] Eva Feillet[*,+] Adrian Popescu[*] Inna Kucher[*] Bertrand Delezoide[†]

(*) Université Paris-Saclay, CEA, LIST
F-91120 Palaiseau, France
name.surname@cea.fr

(+) Université Paris-Saclay, CentraleSupélec, MICS
F-91190 Gif-sur-Yvette, France
name.surname@centralesupelec.fr

(†) Amanda
F-75008, Paris, France

Abstract

Visual datasets are generally constructed from the samples available at the time of their collection and are not further updated. However, these static datasets do not reflect the distribution changes that occur in real data. We analyze how different collection times lead to a shift in class distribution by collecting a set of Flickr images published over 14 years. The proposed "Visual Classes through Time" (VCT-107) dataset contains images tagged by their publication date and includes 107 classes covering various topics (human-made objects, animals, plants, food, etc.). Images from each class are divided into five collection periods to study the impact of time on classification accuracy. When training different classification models using linear probing, we observe an accuracy loss when training on data from one period and testing on other periods. This happens even in the case of a strongly pre-trained model like DinoV2 ViT-B/14. Intuitively, the performance loss is generally more significant when the collection periods between the training and test data are further apart. Our analysis reveals that the temporal shift varies between classes, with the largest shifts observed for human-made objects and the smallest for natural concepts such as animal species. Our results stress the importance of regularly updating models to adapt to time-induced changes in the distribution of visual classes, even when using a strongly pre-trained model. We release the VCT-107 dataset to facilitate research on temporal shifts.

1. Introduction

The performance of image classification models highly depends on the quality of their training data. Many existing datasets are composed of samples collected at a fixed point in time and do not contain temporal information, as is the case for popular resources such as CIFAR-100 [35], ImageNet [14], or YFCC100M [81]. Other datasets are available in different versions that extend or refine the dataset initially proposed, e.g., Google Landmarks [53,92], Google Open Images [36, 39], or Mapillary [50]. However, even these datasets lack temporal metadata and are not designed to study the robustness of models to distributional shifts arising from the passage of time [94].

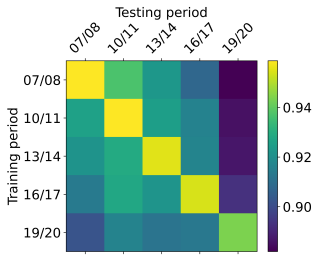

Figure 1. Evolution of the classification accuracy on VCT-107 when training a linear classifier on one data collection period and testing on another. Five periods between 2007-2008 and 2019-2020 are considered. Image features are computed using a ViT-B/14 network pre-trained with DinoV2 [55].

To address the challenge of temporal shift, we introduce the dataset VCT-107. VCT-107 comprises 107 classes and 951,176 images uploaded to the Flickr platform between 2007 and 2020. A timestamp accompanies each image. The dataset encompasses different topics, including household objects, vehicles, buildings, plants, and animals.

We evaluate the generalization ability of various models on VCT-107 in the context of imbalanced and balanced classification tasks and domain-incremental learning [87]. The results of our experiments highlight the existence of a distribution shift that alters classification accuracy. Figure 1 illustrates the impact of time on classification performance. It shows the accuracy of a linear classification layer fitted on top of a pre-trained DinoV2 ViT-B/14 model [55] using data from one period and tested on data from another. We observe that performance generally decreases when the interval between the training and test samples increases, whether

in the future or the past. Nonetheless, our experiments with several continual learning algorithms [20, 43, 44] show that this accuracy gap can be reduced. These algorithms maintain knowledge of past distributions while adapting to new distributions without storing past examples. Furthermore, we find that temporal shift impacts classes differently. The magnitude of the shift can be quantified by analyzing the distances between the distributions of data collected at different moments. This quantification enables a class-level control of the required training dataset updates. These findings suggest good practices for dealing with temporal shifts in visual class representations. We hope this contribution will encourage work to mitigate the obsolescence of visual representations. We will release the image URLs and the code to encourage research on the topic.

2. Related work

Visual classification datasets. Recent advances in visual classification learning have been fostered by the publication of visual datasets [85]. Classification models are commonly evaluated on CIFAR-100 [35], ImageNet-1000 [70] or on domain-specific, fine-grained datasets such as Food-101 [6], Stanford Cars [34] or Oxford Flower-102 [52]. However, these datasets are not designed to challenge the robustness of models against distribution shifts.

Specific visual datasets have been proposed for the task of domain adaptation, including MNIST with diverse backgrounds [18], Office-Home [89], Citiscapes [11] and DomainNet [60]. Datasets like ImageNet-R [26] or ImageNet-D [69] are designed to benchmark the robustness of ImageNet-trained models against domain shifts. The CORE50 dataset [42] comprises 50 objects filmed in 11 settings and is built specifically for continual domain adaptation. In addition to disparate backgrounds and image styles, distribution shifts may arise from different geographies [3, 8, 68], weather conditions [30, 47], or ethnicities [28, 41]. Another line of work proposes to use synthetic images. For example, Visda [61] focuses on the simulation-to-reality shift, and SHIFT [76] uses a generative model to control shifts in scene elements for autonomous driving.

Works on temporal shifts. One of the rare vision datasets with a distribution shift directly caused by time is AmsterTime [95], a collection of 2,500 images matching a street view in Amsterdam (using Mapillary navigation platform) to a historical archival image from the same scene. AgeDB [49] is a dataset for face verification in the wild and contains temporal information about a person's age in each image. The Wild-Time benchmark [94] focuses on temporal shift and covers two visual tasks: gender prediction with the Yearbook dataset [19] and prediction of land use with satellite images from FMoW [10]. We compare our proposed VCT-107 with these datasets in Table 1. Natural language processing works study temporal changes in lexical semantics and propose methods to detect such changes, e.g., [5, 37]. The authors of [23] distinguish between semantic shifts that are more cultural or more linguistic. We refer to the survey of [38] for more details on semantic shifts in word embeddings. In our experiments, we use models trained on visual data only and vision-language models.

Biases and generalization. Datasets partially represent the visual world and are inherently biased [85]. The authors of [17] identify three main types of biases, arising from (1) selecting a subset of items that differ from the general population, (2) framing the object to convey a specific message via the image composition, or (3) assigning different labels or wrong semantic categories. Biases lead to distributional shifts between the data used to train a model and the data encountered during its operational phase, challenging the model on unseen data.

Several lines of work aim to increase a model's ability to generalize to new domains or tasks. Generalization is favored by the quantity, quality, and diversity of its training data [32, 51, 55]. Pre-training with large corpora [86] and multiple data augmentation techniques [58, 75] is now common practice. Multimodal language vision models such as CLIP [63] and DALL-E [65] show strong transferability without per-sample labels. Their self-supervised training uses up to billions of image-text pairs. Diversifying representations using features from intermediate layers of a pre-trained model [16] or combining multiple encoders [78] can also improve transferability. Finally, transfer learning and domain adaptation focus on reusing knowledge gained for solving a source task in a different but related problem [7]. We refer to the surveys of [73, 84] for a detailed review of transfer learning and domain adaptation algorithms.

Continual learning. Continual Learning (CL) builds models that can adapt to their environment and incrementally develop more complex skills and knowledge [4, 83]. Domain-Incremental Learning (DIL) [87] is a CL scenario that learns a classification model sequentially, with each step in the sequence introducing data from a new domain. The set of target classes remains the same throughout the process, but class distributions change. Thus, the challenge is to recognize classes in an increasing number of domains without storing all previous data, a challenge addressed in different ways. The approach of [40] does not require task boundaries but relies on a costly clustering step. The work of [82] leverages self-supervised learning. An adapter method [62] is applied in [56] to adapt a pre-trained model on the initial subset efficiently and then incrementally train a classifier based on a linear discriminant analysis layer. Similarly, RanPAC [44] combines a Parameter-Efficient Transfer Learning (PETL) procedure with a random layer that projects samples in a higher dimensional space to improve discrimination. FeCAM [20] also uses a fixed feature extractor and focuses on incrementally updat-

Dataset	Yearbook [19]	FMoW [10]	AgeDB [49]	AmsterTime [95]	Core50 [42]	**VCT-107**
Input	Yearbook photos	Satellite images	Faces in the wild	Landmarks	Video frames	**Web images**
Prediction	Gender	Land use	Face identification, Age, Gender	Visual place recognition	Object recognition	Object recognition
Time range	1930-2013	2002-2017	∼1890-2017	∼1850-2020	/	**2007-2020**
#domains	-	-	-	2	8 train + 3 test	**5 periods**
#classes	2	63	568	1,231	50	**107**
#samples	37,189	118,886	16,488	2,462	164,866	**951,176**

Table 1. Comparison of visual datasets containing temporal information.

ing a classifier based on the Mahalanobis distance.

3. Constitution of VCT-107

We describe the VCT-107 collection, processing, and labeling process. Then, we analyze the resulting dataset.

Data collection. We downloaded images from the Flickr platform because its content covers diverse visual concepts over a long interval, and its API facilitates the collection of images using predefined temporal intervals. We collected images for five distinct periods: 2007-2008, 2010-2011, 2013-2014, 2016-2017, and 2019-2020, denoted as 07/08, ... 19/20. Grouping images in two-year intervals ensures enough training and test images for all classes. The one-year gap between intervals facilitates the analysis by better separating the data subsets. We initially collected data for the 22/23 period but dropped it because the number of images was insufficient for most classes.

To ensure diversity in the dataset, we prompt ChatGPT-4o to provide class names and definitions from the following nine topics: plants, animals, food, buildings, vehicles, household objects, electronic devices, sporting equipment, and apparel: *Please provide a list of 50 popular [TOPIC_NAME] types using a JSON format for the output*. Since the LLM answers sometimes include less than 50 items, the initial class count is 439. We verify the correctness of the proposed class names and descriptions to filter out hallucinations. We then collect up to 3000 Flickr images and associated metadata for each target year using Flickr's internal search engine ranking. This initial, uncurated dataset includes nearly 11 million images.

Image rights and safety. Following [81], we collected only freely redistributable images, but this approach did not provide enough samples per period for most classes. Therefore, we broadened the search and collected Flickr images with all licenses. This change has practical implications for the distribution of copyrighted content. We follow recent practice in sharing visual datasets [71] and provide the URLs rather than the image files themselves.

Concerned about image safety issues [80], we instructed the annotators to remove any image that could be considered "not safe for work" and to flag any image that might have been taken without the subject's consent. We provided them with clear textual safety guidelines and interacted with them when in doubt. If an image from a cluster was flagged, we removed the entire cluster.

Dataset preprocessing. We preprocess the dataset to minimize the labeling effort. We compute the embeddings of all the collected images using a ViT-B/14 pre-trained using DinoV2 [55]. We remove near-duplicates using a 0.9 cosine similarity threshold between each pair of images uploaded in the same year. We cluster images using K-means [59] with 50 clusters per year. We keep only clusters involving at least two Flickr users to ensure a minimal social consensus on the class's visual representation. We use these clusters to accelerate the annotation process.

Content annotation. We implement a dedicated labeling interface (illustrated in the appendix). Each row of images represents the visual summary of a cluster and contains at most ten images. These images are sampled uniformly based on their L2 distance to the cluster centroid and shown in increasing order of distances from left to right in the interface. This sampling relies on the hypothesis that there is a correlation between the distance to the centroid and the representativeness of an image for a given class. We provide annotators with textual instructions illustrated by examples. The instructions require them to annotate the rightmost image of each row, including a depiction of the visual class according to the LLM's definition. They state that the object may be located in any image region and that other objects can be visible. Three participants contributed to the annotation task, and one participant annotated each cluster. To reduce the annotation effort, the participants first label the image subset from 2020 because it contained the fewest images. Then, we rank the classes according to the number of relevant images labeled for 2020 and keep the 125 most populated classes. Finally, we ask participants to label the images from the remaining nine collection years for these 125 classes. This step provides a fast labeling of the images, but some noise might subsist. Next, we check the annotations of the test subset to ensure a reliable evaluation.

Candidate images for the test set are sampled uniformly from the selected clusters and labeled by the other annotators. They are included in the final test subset if the three annotators agree on their relevance. The specific annotation of test images also validates the clustering-based annotation. We find that the three participants agree on the relevance of over 98% of the images sampled from the clustering-based annotations. We keep a class only if it has at least 40 valid test images and 100 training images per year.

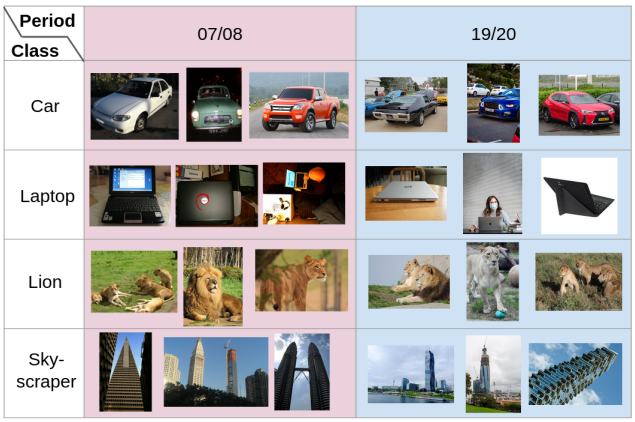

Figure 2. Samples representing four VCT-107 classes during the 2007-2008 and 2019-2020 periods. The car, laptop, and skyscraper classes illustrate the appearance changes of human-made objects whose design changes over time, shifting the representations learned for these classes. Lion has a stable appearance, and the representation shift is much smaller in this case.

VCT-107 summary and illustration. The dataset includes 107 classes from 9 topics, ranging from 31 animal classes to 2 types of electronic devices. The dataset includes between 2881 and 21237 samples per class, with at least 483 images per period. The class names and sample distribution are detailed in the appendix. The images were uploaded by over 248772 Flickr users, who each contributed an average of 4.4 images. The minimum, mean, and maximum user counts per class are 1106, 4289, and 11593, respectively. These numbers ensure that VCT-107 class representations benefit from social consensus. Nevertheless, a selection bias occurs, as with any visual dataset [17].

Figure 2 illustrates the impact of time on visual classes. Due to space restrictions, we sample three images of four classes taken during the earliest and most recent VCT-107 periods. Changes over time in the representations of human-made objects are mainly determined by the lifespan of these objects [72], itself determined by technological advancements, visual design trends, regulation, and brand strategies [90]. Vehicles illustrate the complex interaction between these factors with a continually evolving technological and visual design. For instance, the shift toward electric batteries changes the appearance of cars to match technical requirements [2] but also to highlight their difference from fossil-fuel-based vehicles and increase their appeal [48]. Similar considerations apply to mass-consumption electronic devices, such as laptops and smartphones [1]. Their usage and representations depend on technical advancement and their functions for users of different ages, incomes, and world regions. Interestingly, the visual representations of human-made objects mix the old and the new, highlighting users' fascination for the past [67]. Figure 2 shows that users upload vintage cars during both periods. Visual representation changes are also observed in architecture, with increasing stylistic diversity and the availability of new building materials and techniques [22].

The impact of time is reduced for natural classes such as lions because their appearance does not significantly change. However, trends also appear, particularly for classes closely associated with humans, such as pets. For instance, the popularity of dog breeds evolves [27], influencing the class' visual representation. Equally important, framing biases [17] might still affect their depictions regarding how they are photographed and in which contexts.

While we focus on the impact of time, multiple factors influence visual class depictions. VCT-107 classes are subject to a selection bias [17, 85]. This bias is amplified in operational conditions due to the long-tailed nature of visual datasets [93]. Another important bias comes from the demographic characteristics of the users of Flickr, with variations of social status, ethnicity, gender, and location across time [54]. In particular, some regions of the world tend to be more represented than others in visual datasets [68]. This leads to an imbalanced depiction of visual concepts, particularly for classes such as buildings. The cameras used to take the photos influence image quality and can affect the representations learned. Finally, disparities due to lighting conditions or image colorimetry also occur [79]. Together, these factors create temporal shifts in visual classes. We quantify their effect on image classification in Section 4 and provide an embedding-based analysis in Section 5.

4. Experiments

4.1. Experimental setup

We split VCT-107 into five temporal periods, as described in Section 3. We run experiments with the entire training set and in low-shot scenarios by sampling 200, 100, 50, or 20 images per period. To assess the models' generalization ability, we train them on each period and measure their test accuracy on the other periods. In some experiments, we also accumulate training samples over time to evaluate the effect of retraining from scratch. The test set of each period is fixed across experiments and contains 80 images per class.

We use SGD with a momentum set to 0.9, a weight decay set to $4 \cdot 10^{-5}$, and a cosine learning rate scheduler initialized at 0.1. We train for 100 epochs for full training and 20 epochs for linear probing (LP). This transfer learning method freezes all parameters except the classification layer [33]. Unless otherwise stated, data preprocessing is the same and consists of rescaling the images to $256 \cdot 256$ pixels, then randomly cropping to $224 \cdot 224$ pixels and normalizing using ILSVRC [70] statistics.

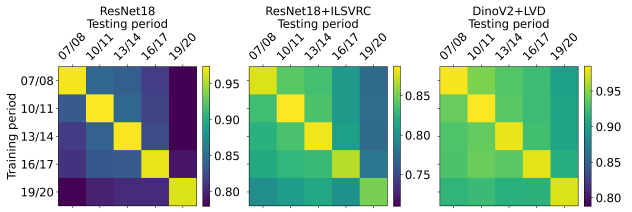

Figure 3. Accuracy across temporal periods when training with the entire VCT-107 dataset using three different backbone models. To facilitate comparison, the range of values is displayed from 80% of the maximum accuracy value of each backbone.

4.2. Impact of the training strategies

We evaluate the capacity of pre-trained and fully-trained models to mitigate the temporal shift. Full training involves the entire VCT-107 dataset because it requires more samples. We use a smaller ResNet18 instead of a ViT as it requires less sample to train. Therefore in Figure 3 we experiment with: (1) a ResNet-18 [25] trained from scratch, (2) a ResNet-18 pre-trained on ILSVRC [70] (3) a DinoV2 ViT-B/14 pre-trained on the LVD for easier comparison with subsequent experiments. The primary objective of these experiments is to assess the accuracy stability over time, not to compare the accuracies obtained with each backbone. Figure 3 shows that the backbone trained from scratch exhibits the largest performance variation. This highlights the importance of pre-training for mitigating temporal shifts. The pre-trained ResNet-18 comes second, with the ViT-B/14 network pre-trained using DinoV2 achieving the highest stability across time. The results from Figure 3 confirm that combining strong pre-training and linear probing constitutes a competitive baseline for mitigating temporal shift.

Due to significant architectural differences, the model trained using DinoV2 is not directly comparable to the ResNet-18 model. A comparison with more similar architectures is necessary to assess whether all pre-training methods yield the same robustness to temporal shifts. To isolate the effect of DinoV2's unsupervised pre-training, we compare its generalization performance against a ViT-B/16 model pre-trained in a supervised manner on ILSVRC. Additionally, we include a ViT-B/16 variant pre-trained on the full ImageNet-21k to evaluate the impact of pre-training dataset size. We evaluate those by training linear probes with 200 samples per class and period. This removes the issue of having an imbalance in the training data, which may slightly alter our results. We provide the results for all combinations of training and testing periods in Figure 4.

With 200 samples per period, DinoV2 achieves a higher average accuracy and improves generalization over time. DinoV2 experiences a maximum accuracy drop of 7.2%, whereas the ViT-B model pre-trained on ILSVRC sees a loss of up to 9.0%. Although the ViT model pre-trained on ImageNet-21k performs slightly worse than DinoV2, it also exhibits a maximum accuracy loss of 6.7%. These results suggest that the primary limitation of the ILSVRC pre-training method lies in the quantity of data rather than its supervised nature.

Finally, we also consider the increasingly popular Contrastive Language-Image Pre-training (CLIP) [64], as its multimodal approach could offer greater robustness. To maintain consistency with our previous experiments, we experimented with two models: the standard ViT-L/14 and a ViT-B/16. For linear probing, we attach the linear classifier after the projection to the shared latent space, retaining only the vision component of the model. This method follows the original approach described by Radford et al. [64]. Figure 4 indicates that temporal shifts also affect multimodal models. However, the ViT-B/16-based CLIP experiences a maximum accuracy loss of only 5.5%, which is lower than that of the other ViT-B models, suggesting that CLIP training provides increased robustness. The appendix provides zero-shot classification scores for each period to illustrate its relative classification difficulty.

4.3. Impact of the training set size

The size of the training set strongly influences the generalization ability in static datasets [51, 85]. Following the findings from Subsection 4.2, we use DinoV2 with linear probing. We experiment with $n \in \{200, 100, 50, 20\}$ training samples per class and period to assess the influence of time in low-shot settings. We repeat each experiment 4 times for each low-shot scenario using four random seeds for sampling and report average results in Figure 5.

Reducing the number of images per class harms the overall performance since individual class representations progressively weaken. Figure 5 highlights that when n decreases, the accuracy on periods other than the training period decreases slower than on the same training period. The average accuracy obtained when testing on the same period as training drops by 5.2% when n goes from 200 to 20. Meanwhile, the average accuracy for the other periods only drops by 3.3%. We also observe that when testing on periods other than the training period, the *relative* accuracy loss decreases slowly as n decreases. In this case, the average accuracy loss is 3.8% and 2.0% for 200 and 20 training images per period, respectively. This result highlights the ability of strong pre-training to handle temporal shifts. This is important in practice, as many real-life datasets include limited training data per class [74].

4.4. Domain-incremental learning

The experiments in Subsections 4.2 and 4.3 do not include any mitigation strategy other than using a strongly pre-trained backbone. Here, we test the effectiveness of continual learning (CL) [57] algorithms against temporal shifts. Domain-incremental learning (DIL) [87] is a sequen-

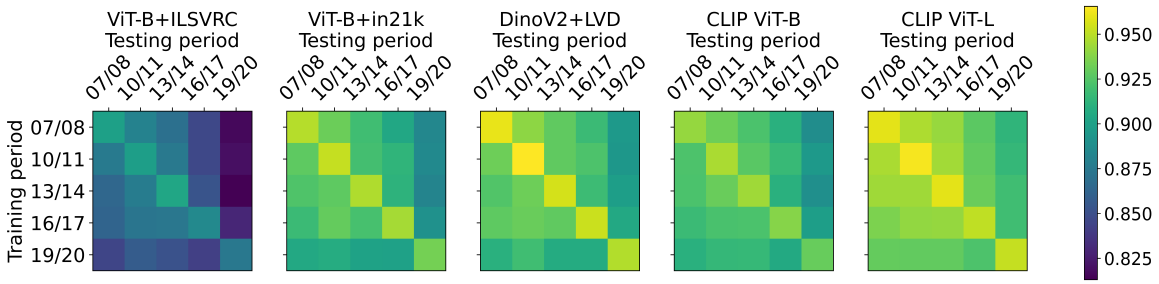

Figure 4. Accuracy across time for cross-period training and testing. All models use linear probing with 200 samples per class and period.

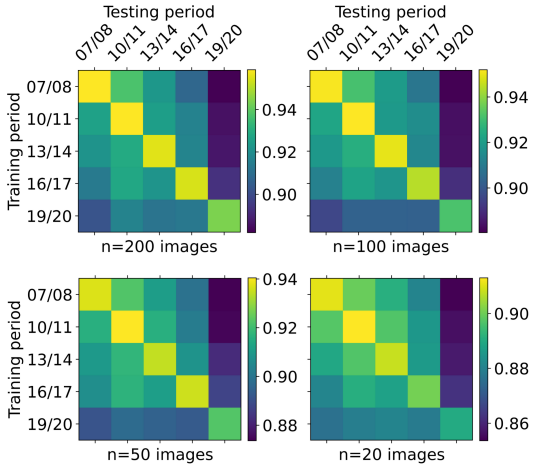

Figure 5. Accuracy over time for DinoV2 ViT-B/14 and linear probing for $n = \{200, 100, 50, 20\}$ samples per class and period.

Algorithm	DinoV2 ViT-B14 LVD-142m [55]	ViT-B16 Aug-Reg IN21k [86]	#Stored params
NCM	92.6	90.7	$82 \cdot 10^3$
FeCAM-1	92.7	92.6	$672 \cdot 10^3$
FeCAM-n	<u>94.3</u>	92.9	$63 \cdot 10^6$
RanPAC	**94.8**	**94.6**	$108 \cdot 10^6$
Replay20	93.4	92.5	$1.6 \cdot 10^9$
Accumulate	94.1	<u>93.0</u>	$16 \cdot 10^9$

Table 2. Average accuracy for six algorithms and two pre-trained backbones. The algorithms are ordered by the number of parameters added to the backbone. The storage needs are computed for images of size 3*224*224. **Best results**, <u>second best</u>.

tial learning process where each step corresponds to a new domain. Here, a domain is a data collection period, e.g., 2007-2008, 2010-2011, etc. The set of classes to recognize remains the same, but their distribution changes. Each step of the process aims to obtain a model that can recognize all classes, regardless of the data collection period. We follow the DIL protocol from [44] and include all $T = 5$ periods in the test set. The average accuracy is computed as the mean value of the test accuracy across the T training steps: $A = \frac{1}{T}\sum_{t=1}^{T} Acc(M_t, \bigcup_{i=1}^{T} D_i)$, where M_t is the model trained at step s_t on data collected at time t and D_i is the test dataset corresponding to period i.

We experiment with several competitive CL algorithms using a fixed encoder. The Nearest Class Mean classifier (NCM) [31] updates a running mean embedding vector for each class and predicts the class using the cosine similarity to class prototypes. FeCAM [20] also stores a mean vector for each class and computes a shared feature covariance matrix ("FeCAM-1") or one feature covariance matrix per class ("FeCAM-n"), used to compute the Mahalanobis distance between the embedding of a test sample and the mean class vectors. RanPAC [44] combines a PETL step with a random projection from dimension 768 to 10,000 to better separate classes. At inference, distances to class means are computed using the Gram matrix. These algorithms do not store past images, which is useful when storage or privacy issues must be considered. Still with a fixed encoder, we also consider linear probing with a cumulative replay buffer of 20 images per period ("Replay-20") and a cumulative replay buffer containing all the training images seen so far ("Accumulate").

We report the average DIL accuracy in Table 2. The results show that DIL algorithms match or outperform naive replay and accumulation strategies while requiring at least 250 times less additional memory. RanPAC and FeCAM-n, the two algorithms that perform the most refined modeling of past knowledge, obtain the best accuracy. Figure 6 indicates the DIL algorithms reduce the accuracy losses for test data from past periods but are ineffective for future data. The results confirm the effectiveness of CL algorithms in mitigating the effects of domain shifts when combined with a pre-trained model. However, higher accuracies tend to be obtained with higher memory requirements.

5. Temporal shifts analysis

We investigate the importance of temporal shift in VCT-107 by analyzing the embedding space and the performance variations per general topics over time.

5.1. Topic-based analysis of temporal shifts

We discuss the effect of temporal shifts for eight VCT-107 general topics by refining the analysis of results from Subsection 4.3 obtained with 200 images per class and period. Figure 7 shows that the intra-period accuracy varies significantly depending on the topic. *Household Objects* and *Apparel* are the most challenging topics, while *Animals*

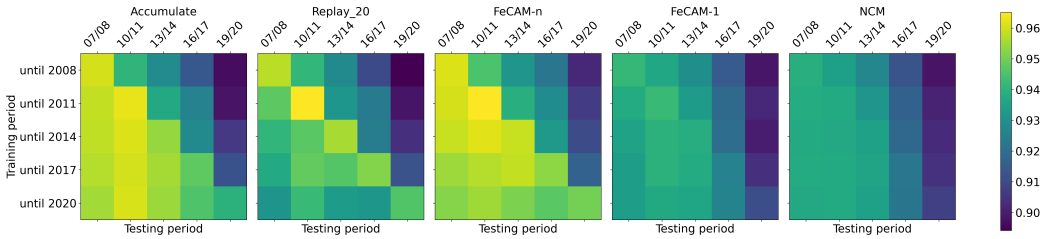

Figure 6. Accuracy comparison when accumulating samples ("Accumulate" and "Replay_20") and using linear probing vs. updating the model using incremental learning ("FeCAM-1", "FeCAM-n" "NCM"). Experiments with a pre-trained DinoV2 ViT-B/14 network.

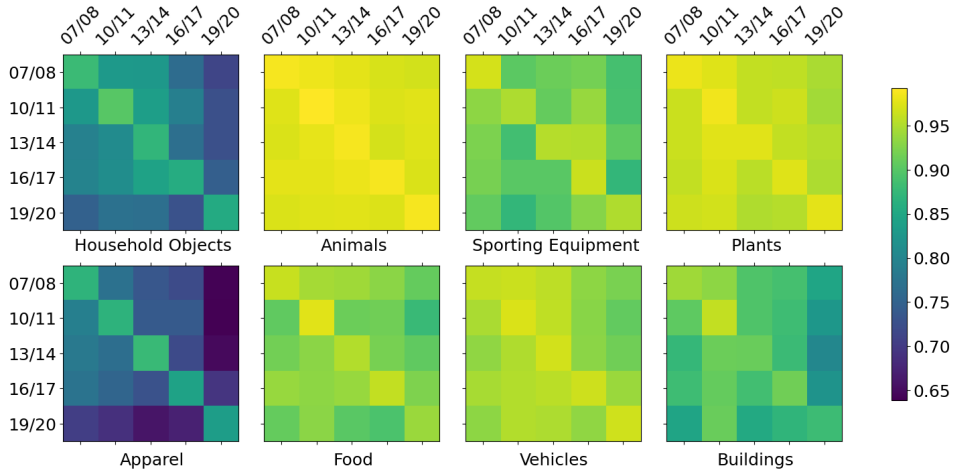

Figure 7. Accuracy across temporal periods for the general topics included in VCT-107. The results are obtained using a DinoV2 backbone with linear probing and 200 training images per class. We exclude *Electronic Devices* because this topic has only two classes.

and *Plants* are the easiest ones. The effect of temporal shifts is also more significant for human-made classes than natural ones. The fact that time has variable effects for different topics is important in practice since it indicates that temporal adaptation could be tailored at the class level.

5.2. Embedding-based analysis of temporal shifts

After observing the effect of temporal shift on VCT-107 classes, we attempt to predict shift before training. We compare the following distances: (1) the $L2$ distance between the centroids of two distributions, (2) the Fréchet Inception Distance [15] (FID) used by [46] to measure the gap between two distributions when studying generalization, (3) the energy distance [21, 77] that tests for equal distribution in high dimensions without distributional assumption [77] and (4) the Sinkhorn distance for optimal transport, a popular approximation of the Wasserstein distance [13, 45, 91].

We measure the distance of each class's mean DinoV2 embedding distributions for each pair of temporal periods in which the training and test periods are distinct. We compare these distances to the loss of accuracy for the same pairs of periods. Let A_{origin} and A_{target} be the test accuracy for the training and target periods. The relative loss in accuracy is given by: $(A_{target} - A_{origin})/A_{origin}$. For readability, in Figure 8, we average the distances and the accuracy losses by general topic for every pair of a train and a test period. We observe that for each considered metric, the average distances generally grow with the accuracy loss. FID and Sinkhorn's algorithms successfully assigned higher values to the two most affected topics. They could be used in practice to decide whether to update the visual representation of a topic (or even an individual class).

Finally, we check if the magnitude of the shift increases with the time difference between the two distributions by the FID metric. We average the distances corresponding to each topic based on the temporal interval between the target period and the others. Figure 9 confirms that the average distance grows for all topics as the interval increases.

6. Discussion and conclusion

We introduce VCT-107 to analyze the impact of time on visual classification models. We experiment in several settings and observed an accuracy drop when training and testing during different periods. The performance loss generally grows with the temporal distance between the training and testing periods. We also observe that the classification accuracy loss depends on the type of classes.

Practical guidelines. Based on these results, we propose the following recommendations for improving classification performance under temporal shifts:

- *Use self-supervised pre-training* with linear probing to re-

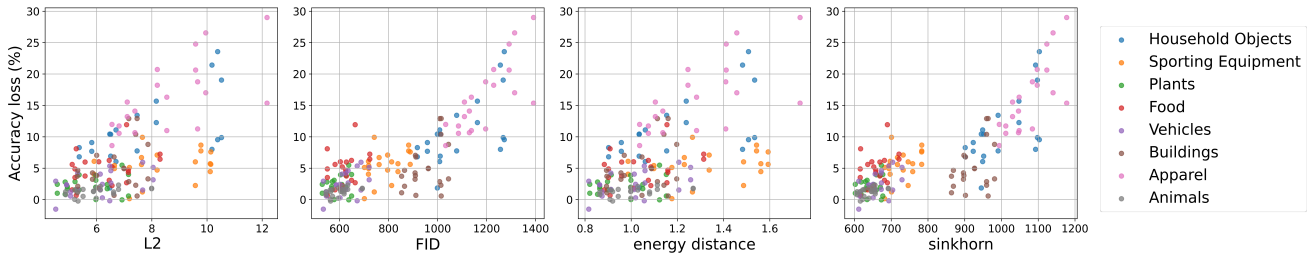

Figure 8. Relative accuracy loss over time for the classes of the general VCT-107 topics as a function of distribution shift measured with four metrics. Results aggregate distances and accuracies for individual classes for the assessed training-test period pairs.

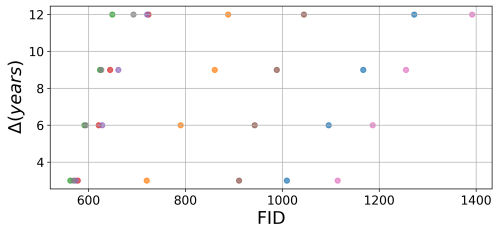

Figure 9. VCT-107 topic distributions shift measured with the FID distance as a function of the temporal interval between training and test periods. We use the same colors for topics as Figure 8.

duce the performance variability over time. The results confirm the improved generalization ability of pre-trained models [55] in a temporal context. The relative pre-training performance depends on the implemented type of learning, the dataset size, and the dataset diversity, but the relation is not always straightforward. In particular, our experiments indicate that self-supervised visual learning outperforms multimodal training in an image classification task despite visual models having a smaller parametric footprint and using a smaller training set.

- *Implement continual learning algorithms* to further mitigate performance loss on past data if retraining with all historical data is not an option. CL algorithms require the storage of samples or statistical information but make the training process much more efficient. They benefit costly learning processes, such as training foundation models with huge datasets [12].
- *Consider the type of visual classes* when learning over time. Our experiments confirm the intuition that human-made objects are more impacted by temporal shifts. However, there are important differences between the different types of human-made objects. The analysis of class embeddings indicates that using an appropriate distance can predict the need to update the training set. Adapting the update rate for different classes is particularly interesting when training foundation models, whose updates are needed to keep pace with novelty but are also costly.

Limitations and future work. We discuss limitations and suggest future work directions to mitigate them.

- The dataset is sourced from Flickr. Adding supplementary sources would increase the generality of the findings, but access to photos with temporal metadata over such a long period is not straightforward. We can only hope that social platforms will facilitate researchers' access to data, but we observe an inverse trend in practice.
- The reliance on third-party data when building large datasets is needed, which induces redistribution limitations. Acknowledging the potential reproducibility limitations, we follow recent practices [71] and provide the image URLs to respect image rights.
- We tested pre-training and continual learning to mitigate temporal shifts. Other techniques can be considered to counter this shift, including (1) PETL methods [29] with adapters designed for temporal shifts, (2) domain adaptation methods [73] to better preserve past knowledge through time, and (3) imbalanced learning methods [24, 66] to rebalance performance when the number of samples per class varies within a period or across them.
- VCT-107 covers several general topics, enabling their analysis over time. However, the dataset would benefit from including additional topics and enriching existing ones to broaden the analysis. It would also be interesting to analyze the effects of time for finer-grained visual classes. These developments are left for future work, building on the proposed dataset creation pipeline.
- The images included in VCT-107 are labeled for a single class, following a protocol commonly used in image classification [14, 35, 88]. It would be interesting to add multi-label annotations to all dataset images to test the effect of class co-occurrences during classification.
- We fixed classes over time to facilitate comparisons across periods. An enriched version of the dataset could include classes that appear over time. This enrichment would be beneficial for fine-grained datasets.
- The dataset measures the effect of time at the year scale. Refining the temporal scale to enable stream learning would be interesting, as proposed in [9].

We hope this work will stimulate the community's interest in considering the temporal dimension of image classification. This research topic can increase the robustness of deep models, especially for classes whose visual representations change frequently over time.

References

[1] Jon Agar. *Constant touch: A global history of the mobile phone*. Icon Books Ltd, 2013. 4

[2] Ferruh Altun, Sezai Alper Tekin, Seyfettin Gürel, and Mihai Cernat. Design and optimization of electric cars. a review of technological advances. In *2019 8th International Conference on Renewable Energy Research and Applications (ICRERA)*, pages 645–650. IEEE, 2019. 4

[3] Sara Beery, Guanhang Wu, Trevor Edwards, Filip Pavetic, Bo Majewski, Shreyasee Mukherjee, Stanley Chan, John Morgan, Vivek Rathod, and Jonathan Huang. The auto arborist dataset: a large-scale benchmark for multiview urban forest monitoring under domain shift. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21294–21307, 2022. 2

[4] Eden Belouadah, Adrian Popescu, and Ioannis Kanellos. A comprehensive study of class incremental learning algorithms for visual tasks. *Neural Networks*, 135:38–54, 2021. 2

[5] Yuri Bizzoni, Stefania Degaetano-Ortlieb, Peter Fankhauser, and Elke Teich. Linguistic variation and change in 250 years of english scientific writing: A data-driven approach. *Frontiers in Artificial Intelligence*, 3:73, 2020. 2

[6] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 2

[7] Stevo Bozinovski. Reminder of the first paper on transfer learning in neural networks, 1976. *Informatica*, 44(3), 2020. 2

[8] Kyle Buettner, Sina Malakouti, Xiang Lorraine Li, and Adriana Kovashka. Incorporating geo-diverse knowledge into prompting for increased geographical robustness in object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13515–13524, 2024. 2

[9] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. Neil: Extracting visual knowledge from web data. In *Proceedings of the IEEE international conference on computer vision*, pages 1409–1416, 2013. 8

[10] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018. 2, 3

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[12] Andrea Cossu, Antonio Carta, Lucia Passaro, Vincenzo Lomonaco, Tinne Tuytelaars, and Davide Bacciu. Continual pre-training mitigates forgetting in language and vision. *Neural Networks*, page 106492, 2024. 8

[13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 7

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 8

[15] D.C. Dowson and B.V. Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982. 7

[16] Utku Evci, Vincent Dumoulin, Hugo Larochelle, and Michael C Mozer. Head2Toe: Utilizing intermediate representations for better transfer learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 6009–6033. PMLR, 17–23 Jul 2022. 2

[17] Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. A survey on bias in visual datasets. *Computer Vision and Image Understanding*, 223:103552, 2022. 2, 4

[18] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016. 2

[19] Shiry Ginosar, Kate Rakelly, Sarah Sachs, Brian Yin, and Alexei A Efros. A century of portraits: A visual historical record of american high school yearbooks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–7, 2015. 2, 3

[20] Dipam Goswami, Yuyang Liu, Bartłomiej Twardowski, and Joost van de Weijer. Fecam: Exploiting the heterogeneity of class distributions in exemplar-free continual learning. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 6

[21] Ruite Guo and Vic Patrangenaru. Testing for the equality of two distributions on high dimensional object spaces. *arXiv preprint arXiv:1703.07856*, 2017. 7

[22] Elie G Haddad, David Rifkind, and Ms Sarah Deyong. *A critical history of contemporary architecture: 1960-2010*. Ashgate Publishing, Ltd., 2014. 4

[23] William L Hamilton, Jure Leskovec, and Dan Jurafsky. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the conference on empirical methods in natural language processing. Conference on empirical methods in natural language processing*, volume 2016, page 2116. NIH Public Access, 2016. 2

[24] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.*, 21(9):1263–1284, 2009. 8

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, CVPR, 2016. 5

[26] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization.

In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021. 2

[27] Harold Herzog. Forty-two thousand and one dalmatians: Fads, social contagion, and dog breed popularity. *Society & animals*, 14(4):383–397, 2006. 4

[28] Daniel E Ho, Emily Black, Maneesh Agrwawala, and Fei-Fei Li. Domain shift and emerging questions in facial recognition technology. *policy brief, Stanford University Human-Centered Artificial Intelligence, https://hai. stanford. edu/sites/default/files/2020-11/HAI_FRT_WhitePaper_PolicyBrief_Nov2020. pdf*, 2020. 2

[29] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. 8

[30] Joachim Houyon, Anthony Cioppa, Yasir Ghunaim, Motasem Alfarra, Anaïs Halin, Maxim Henry, Bernard Ghanem, and Marc Van Droogenbroeck. Online distillation with continual learning for cyclic domain shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2023. 2

[31] Paul Janson, Wenxuan Zhang, Rahaf Aljundi, and Mohamed Elhoseiny. A simple baseline that questions the use of pretrained-models in continual learning. *arXiv preprint arXiv:2210.04428*, 2022. 6

[32] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020. 2

[33] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? *CoRR*, abs/1805.08974, 2018. 4

[34] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 2

[35] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1, 2, 8

[36] Ilya Krylov, Sergei Nosov, and Vladislav Sovrasov. Open images v5 text annotation and yet another mask text spotter. In Vineeth N. Balasubramanian and Ivor Tsang, editors, *Proceedings of The 13th Asian Conference on Machine Learning*, volume 157 of *Proceedings of Machine Learning Research*, pages 379–389. PMLR, 17–19 Nov 2021. 1

[37] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th international conference on world wide web*, pages 625–635, 2015. 2

[38] A Kutuzov, L Øvrelid, E Velldal, and T Szymanski. Diachronic word embeddings and semantic shifts: A survey. In *COLING 2018-27th International Conference on Computational Linguistics, Proceedings*, pages 1384–1397, 2018. 2

[39] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 1

[40] Christiaan Lamers, René Vidal, Nabil Belbachir, Niki van Stein, Thomas Bäeck, and Paris Giampouras. Clustering-based domain-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3384–3392, October 2023. 2

[41] Chun-Hsien Lin and Bing-Fei Wu. Mitigating domain mismatch in face recognition using style matching. *Neurocomputing*, 487:9–21, 2022. 2

[42] Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. In *Conference on Robot Learning*, pages 17–26. PMLR, 2017. 2, 3

[43] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost Van De Weijer. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, 2022. 2

[44] Mark D McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton van den Hengel. Ranpac: Random projections and pre-trained models for continual learning. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 6

[45] Arthur Mensch and Gabriel Peyré. Online sinkhorn: Optimal transport distances from sample streams. *Advances in Neural Information Processing Systems*, 33:1657–1667, 2020. 7

[46] Timo Milbich, Karsten Roth, Samarth Sinha, Ludwig Schmidt, Marzyeh Ghassemi, and Bjorn Ommer. Characterizing generalization under out-of-distribution shifts in deep metric learning. *Advances in Neural Information Processing Systems*, 34:25006–25018, 2021. 7

[47] M Jehanzeb Mirza, Marc Masana, Horst Possegger, and Horst Bischof. An efficient domain-incremental learning approach to drive in all weather conditions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3001–3011, 2022. 2

[48] Ingrid Moons and Patrick De Pelsmacker. Emotions as determinants of electric car usage intention. *Journal of marketing management*, 28(3-4):195–237, 2012. 4

[49] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, volume 2, page 5, 2017. 2, 3

[50] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017. 1

[51] Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of clip. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 21455–21469. Curran Associates, Inc., 2022. 2, 5

[52] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008. 2

[53] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*, pages 3456–3465, 2017. 1

[54] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*, 2:13, 2019. 4

[55] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2, 3, 6, 8

[56] Aristeidis Panos, Yuriko Kobe, Daniel Olmeda Reino, Rahaf Aljundi, and Richard E. Turner. First session adaptation: A strong replay-free baseline for class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18820–18830, October 2023. 2

[57] German Ignacio Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113, 2019. 5

[58] Chanwoo Park, Sangdoo Yun, and Sanghyuk Chun. A unified analysis of mixed sample data augmentation: A loss function perspective. *Advances in Neural Information Processing Systems*, 35:35504–35518, 2022. 2

[59] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in python. *CoRR*, abs/1201.0490, 2012. 3

[60] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 2

[61] Xingchao Peng, Ben Usman, Neela Kaushik, Dequan Wang, Judy Hoffman, and Kate Saenko. Visda: A synthetic-to-real benchmark for visual domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2021–2026, 2018. 2

[62] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 2

[63] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[64] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[65] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

[66] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020. 8

[67] Simon Reynolds. *Retromania: Pop culture's addiction to its own past*. Macmillan, 2011. 4

[68] William A Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2, 4

[69] Evgenia Rusak, Steffen Schneider, Peter Vincent Gehler, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Imagenet-d: A new challenging robustness dataset inspired by domain adaptation. In *ICML 2022 Shift Happens Workshop*, 2022. 2

[70] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 2, 4, 5

[71] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 3, 8

[72] Tianfeng Shi, Rong Huang, and Emine Sarigöllü. Consumer product use behavior throughout the product lifespan: A literature review and research agenda. *Journal of environmental management*, 302:114114, 2022. 4

[73] Peeyush Singhal, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. Domain adaptation: challenges, methods, datasets, and applications. *IEEE access*, 11:6973–7020, 2023. 2, 8

[74] Yisheng Song, Ting Wang, Puyu Cai, Subrota K Mondal, and Jyoti Prakash Sahoo. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Computing Surveys*, 55(13s):1–40, 2023. 5

[75] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021. 2

[76] Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. Shift: a synthetic driving dataset for continuous multi-task domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21371–21382, 2022. 2

[77] Gábor J Székely, Maria L Rizzo, et al. Testing for equal distributions in high dimension. *InterStat*, 5(16.10):1249–1272, 2004. 7

[78] Youssef Tamaazousti, Hervé Le Borgne, and Céline Hudelot. Mucale-net: Multi categorical-level networks to generate more discriminating features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6711–6720, 2017. 2

[79] Jin Tan, Taiping Zhang, Linchang Zhao, Xiaoliu Luo, and Yuan Yan Tang. A robust image representation method against illumination and occlusion variations. *Image and Vision Computing*, 112:104212, 2021. 4

[80] David Thiel. Identifying and eliminating csam in generative ml training data and models. Technical report, Technical Report. Stanford University, Palo Alto, CA. https://purl. stanford . . . , 2023. 3

[81] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 1, 3

[82] Mamatha Thota, Dewei Yi, and Georgios Leontidis. Lleda—lifelong self-supervised domain adaptation. *Knowledge-Based Systems*, 279:110959, 2023. 2

[83] Sebastian Thrun. A lifelong learning perspective for mobile robot control. In *Intelligent robots and systems*, pages 201–214. Elsevier, 1995. 2

[84] Songsong Tian, Weijun Li, Xin Ning, Hang Ran, Hong Qin, and Prayag Tiwari. Continuous transfer of neural network representational similarity for incremental learning. *Neurocomputing*, 545:126300, 2023. 2

[85] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, 2011. 2, 4, 5

[86] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 2, 6

[87] Gido M van de Ven, Tinne Tuytelaars, and Andreas S Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, 2022. 1, 2, 5

[88] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 8

[89] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[90] Rudi Volti. *Cars and culture: The life story of a technology*. JHU Press, 2006. 4

[91] Jie Wang, Rui Gao, and Yao Xie. Sinkhorn distributionally robust optimization. *arXiv preprint arXiv:2109.11926*, 2021. 7

[92] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584, 2020. 1

[93] Lu Yang, He Jiang, Qing Song, and Jun Guo. A survey on long-tailed visual recognition. *International Journal of Computer Vision*, 130(7):1837–1872, 2022. 4

[94] Huaxiu Yao, Caroline Choi, Bochuan Cao, Yoonho Lee, Pang Wei W Koh, and Chelsea Finn. Wild-time: A benchmark of in-the-wild distribution shift over time. *Advances in Neural Information Processing Systems*, 35:10309–10324, 2022. 1, 2

[95] Burak Yildiz, Seyran Khademi, Ronald Maria Siebes, and Jan Van Gemert. Amstertime: A visual place recognition benchmark dataset for severe domain shift. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2749–2755. IEEE, 2022. 2, 3