Getting the Bugs Out: Entomology Using Computer Vision

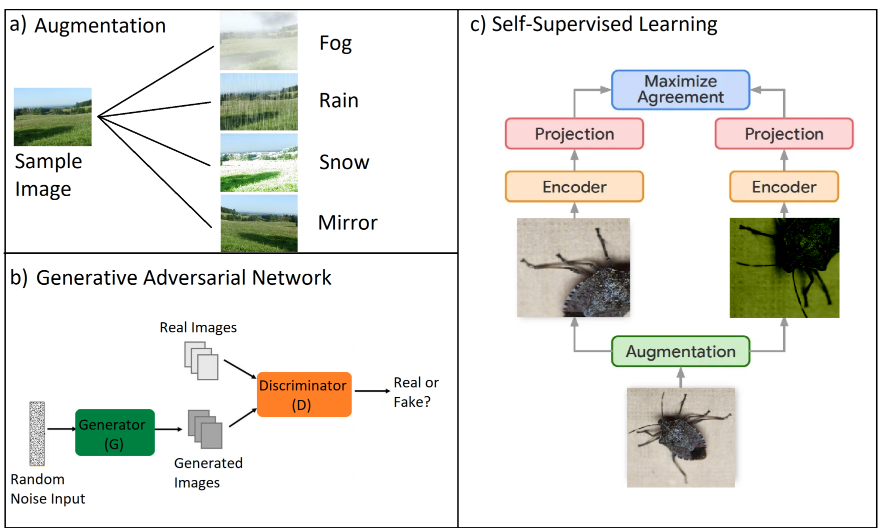
Stefan Schneider[1], Graham Taylor[1], Stefan Kremer[1], and John Fryxell[1]

[1]University of Guelph

October 13, 2022

Abstract

Deep learning for computer vision has shown promising results in the field of entomology. Deep learning performance is maximized primarily by bulk labeled data which, outside of rare circumstances, are limited in ecological studies. Currently, to utilize deep learning systems, ecologists undergo extensive data collection efforts, or limit their problem to niche tasks. These solutions do not scale to region agnostic models. There are solutions using data augmentation, simulators, generative models, and self-supervised learning that supplement limited data labels. Here, we highlight the success of deep learning for computer vision within entomology, discuss data collection efforts, provide methodologies for annotation efficient learning, and conclude with practical guidelines for how ecologists can empower accessible automated ecological monitoring on a global scale.



Getting the Bugs Out: Entomology Using Computer Vision

Stefan Schneider, Graham W. Taylor, Stefan C. Kremer, John M. Fryxell

Corresponding Author:

Stefan Schneider

50 Stone Rd E, Guelph, Ontario, Canada

sschne01@uoguelph.ca

October 12, 2022

Authorship Statement – Stefan Schneider was primary motivator of this work, responsible for the research, writing, and networking between authors. Graham Taylor and Stefan Kremer assisted in conceptualizing the deep learning components of the work. John Frxyell provided ecological insights and motivations for this work. All authors were responsible for revising the initial manuscript drafted by Stefan Schneider.

Data Accessibility – We confirm the data for this article will be archived in Zenodo and the data DOI link included in this document.

Keywords— arthropod, computer vision, deep learning, entomology, generative models, foundation model

Length Abstract – 120 words.

Length Document – 4360 words.

Number of Figures – 1.

Number of References – 92.

Number of Tables – 0.

Teaser – Reviewing the existing efforts of deep learning for entomology and organizing towards foundation models that generalize across taxa.

Abstract

Deep learning for computer vision has shown promising results in the field of entomology. Deep learning performance is maximized primarily by bulk labeled data which, outside of rare circumstances, are limited in ecological studies. Currently, to utilize deep learning systems, ecologists undergo extensive data collection efforts, or limit their problem to niche tasks. These solutions do not scale to region agnostic models. There are solutions using data augmentation, simulators, generative models, and self-supervised learning that supplement limited data labels. Here, we highlight the success of deep learning for computer vision within entomology, discuss data collection efforts, provide methodologies for annotation efficient learning, and conclude with practical guidelines for how ecologists can empower accessible automated ecological monitoring on a global scale.

1 Introduction

We live in a time of rapid global change where the pace at which we can collect and analyze ecological data makes it imperative to capture signals of ecosystem collapse. Insects and other arthropods play a crucial role in crop pollination [1], beneficial control of pests [2], and terrestrial food web dynamics [3]. Hallmann et al. [4]'s ground-breaking study demonstrated a 75% decrease in insect abundance across 63 conservation areas over a 30 year span. Subsequent work documents that this declining trend in insect abundance has been occurring across a wide variety of taxa and locations [5, 6, 7]. Drastic changes in arthropod population abundance and diversity have negative cascading effects on ecological stability and ecosystem resiliency [8, 9, 10]. To expedite and improve the analysis of these trends, the ecological field is currently developing deep learning methods to better understand this potential threat of food web collapse [11, 12, 13, 14, 15].

Deep learning systems for computer vision offer the predictive capabilities of an expert anywhere in the world at massive cost reduction. While computationally expensive to train, deployed deep learning systems can operate on average computers and modern mobile devices

[16]. van Klink et al. [17]'s 2022 review highlights the use of deep learning for computer vision, acoustic monitoring, radar, and molecular models for entomology. As a continuation of these recent successes, ecological deep learning methods would benefit from initiatives that focus on broad scale applications with a global perspective. Current approaches require building a dataset using experts with laboratory devices and training models on computing resources only available in first world countries [11, 15]. This approach creates a bias in trends analyzed and prevents less resourced labs from participating in the deep learning advance. To achieve a global initiative of ecological data collection, we believe there should be a focus on designing accessible and generalizable deep learning systems to process ecological data collected cheaply from rural environments, using only a net, camera, and possibly an internet connection [18]. This would empower those untrained to contribute to expert level analysis from remote locations anywhere in the world. This form of data collection effort would create an ethically fair data analysis pipeline capable of providing a dynamic feedback loop of year-over-year metrics related to abundance, biomass, and richness anywhere in the world.

In order for this global objective to succeed, there exist many technical challenges. A main challenge for deep learning models to perform in global settings is the availability of data that extend class labels beyond niche taxa groupings or confined geographic regions. Currently, the majority of models trained have been limited to narrow groupings, primarily due to limited labeled data availability [19, 20, 21, 22]. There exist deep learning methods related to annotation efficient learning that overcome this limitation that have been successfully utilized in other disciplines [23, 24, 25, 26, 27]. Here, we focus on methods that can empower ecologists to accomplishing the training of deep learning models with a global initiative, focusing specifically on computer vision. To do this, we highlight current successes, current limitations, technical solutions for how these limitations can be overcome, and lastly our perspective on future directions.

2 Computer Vision Entomologist AI Systems

The ongoing exploration of deep learning in the field of entomology continually makes strides to accomplish what previously required human experts [28, 29]. This is particularly true for computer vision and entomology as arthropod image data with fixed numbers of classifications is well-suited for deep learning models that have, in recent years, standardized around specific vision architectures (ResNet, DenseNet, Vision Transformer, etc.) [30, 31, 32]. There are alternative ways of approaching vision tasks depending on the input image and output label. These differences can be summarized into two main dichotomous pairs:

- Lab-based vs. field-based images. Lab-based results can utilize imaging with standardized/uniform conditions [28, 33, 34] while field-based images must generalize to variable backgrounds and lighting conditions [19, 35, 36]. If desired, lab based approaches can also take advantage of capturing multiple images per individual from a variety of angles.
- Single vs. multiple individuals per image. Images of single individuals typically assume that the subject is centered and occupies the majority of the image, thus they do not need a separate segmentation step [11, 37], while images with multiple individuals require a model with the ability to successfully crop, extract and classify specific regions of an image [19, 35, 36].

The use of deep learning for computer vision in entomology has been predominately in three disciplines: museum specimens, pest management, and ecological sampling. We briefly explore these here.

2.1 Museum Specimens

Images of museum specimens are often ideal: lab based, single individual, well-mounted, high resolution, and clear with little to no noise in the background. These conditions are optimal for maximizing machine learning performance. Marques et al. [33] demonstrated the potential success of deep learning systems when applied under museum conditions classifying 57 ant genera using 127,832 images, where head views provided the best prediction accuracy. Hansen et al. [28] demonstrated that deep learning systems can distinguish among 361 carabid beetle species considering 364 images taken from the British Isles. The breadth and diversity of museum specimens will provide rich source of training data for general entomologist AI systems.

2.2 Pest Management

Images used to detect and manage pests are often 'noisy' images with variable backgrounds and lighting conditions requiring a model's ability to generalize often beyond the training distribution. In addition, images may contain many individuals, requiring object detection models to localize individuals. Xia et al. [36] used deep learning systems to classify 24 pest insects from field crop images with non-uniform backgrounds. Ding and Taylor [19] expanded a limited dataset of 100s of images using data augmentation to localize and train a deep learning model to count the number of codling moths, a major pest to agricultural crops. Rustia et al. [35] collected data autonomously from greenhouse sticky traps using an object detector and series of sub-classification deep learning networks to localize insect individuals and re-train and improve the model over time. Expanding these works to consider a single model capable of generalizing across pests would aid farmers all over the world.

2.3 Ecological Sampling

Images taken in an ecological context are often either images from the field, or images of curated samples captured in a laboratory setting. In laboratory settings, imaging is traditionally, but not necessarily, done using a single individual per image. Motta et al. [37]'s deep learning classifier can distinguish mosquitoes by species and sex using images captured in a laboratory setting from a dataset of 4,000 images. Tuda and Luna-Maldonado [38] showed deep learning systems outperformed traditional computer vision methods for characterizing populations and species assemblages of the pest beetle *Callosobruchus chinensis* and 2 parasitic wasps: *Anisopteromalus* and *Heterospilu*. Gerovichev et al. [18] analyzed sticky traps placed in Eucalyptus forests to quantify the abundance of two hemipteran pests of eucalypts and a parasitoid wasp. Ärje et al. [11] quantified insect assemblage/diversity using the robotic system BIOSCAN which funnels single individuals into a tube where an image is captured. Similarly, Schneider et al. [15] utilized a white background to isolate arthropod individuals from bulk samples, classifying order, diversity, and order level biomass of 1000s of arthropod samples from a single photo. The use of a single model to generalize across taxa could automate ecological analyses anywhere in the world.

3 Big Data?

The above papers demonstrate the successful predictive capabilities of deep learning on ecological data. These studies, however, follow a trend where each are based on niche, limited ecological datasets that consider a small number of classes and are restricted to specific geographic regions. When considering broad ecological questions and the prospect of global ecological efforts, models need be more general, and operate beyond these niche subsets. This problem is exacerbated as we pursue finer-grained classification from order, down to species, where the number of required labels grows by several orders of magnitude.

It is common to see modern learning systems with millions to billions of parameters which are tuned during training to a given data distribution [39]. With such a large number of parameters, deep learning systems continually improve performance when presented with millions or more labeled examples, achieving spectacular results [39, 40]. One approach to expand the data availability and solve predictive tasks using deep learning in ecology is the massive data science effort to aggregate images from lab and field cameras around the world [41, 42, 43]. While we do encourage efforts to empower research groups around the world with standardized data releases, there are many challenges to overcome. These challenges include:

- Permissions - Often times multiple individuals and funding sources are involved in the collection of data. Ecological data collection efforts often span years, and even decades. Getting permissions from all parties involved in the formulation of data can be difficult to obtain.

- Standardizing labels - When assigning taxonomic labels there exists a hierarchy of label granularity, where samples may be labeled to any of the order, family, genus, or species level depending on the original research objective. When training models from combined data sources, one must be able to handle these intermittent hierarchical taxonomic labels.

- Human error - Different research labs have different levels of access to experts and equipment that improve the accuracy of taxonomic labels. A combine dataset would have varied levels of label accuracy.

- Image resolution - Images of arthropod samples will range wildly depending on how the data were collected considering the original task. One must determine how best to handle these variable image resolutions.

- Environmental setting - Across tasks, arthropods will be captured in a wide variety of environmental settings. Biases towards particular environments may impact performance when training models.

- Numbers of individuals - Ecological images can contain a variable number of individuals. One may need to maintain two datasets: one for object detection with location annotations, and another for standard classification.

- Data biases - When considering ecological sampling, there will be inevitably biases within the data. Arthropods of interest and frequent arthropods are often over-represented, while rare arthropods from underrepresented geographic locations will inevitably be under represented.

While not an exhaustive list, these challenges are examples of what must be overcome for each dataset. Dealing with these challenges will be primarily a manual process requiring an organization to monitor and govern the overall quality and usability of the data releases. While important, the data science approach will be slow and still require technical solutions like those described below to account for biases within the data.

In ecology, an additional consideration when utilizing deep learning systems is that, we often care about the rare, endangered, and unexpected over the common. Deep learning systems, in principle, are designed for the opposite, as they predict signals that are frequent within the realm of variation provided by a given data distribution [44]. In classification systems, this is known as class imbalance, where classes with frequent observations overwhelm the few examples of rare classes [45, 46, 47, 48]. Due to the urgently needed motivations of ecological research to observe the rare and under-represented, we have the opportunity to employ technical innovations that overcome such challenges in data collection efforts.

Ecological analyses will benefit from deep learning approaches focused on data efficiency where there is limited, and even no, labeled data. Here we outline three deep learning techniques, in combination with case studies, highlighting the method and providing data scenarios where the technique would help overcome their limitations. We group these techniques into three main forms: data augmentation [49, 50, 51], data generation [52, 53, 54, 55], and self-supervised learning [56, 57, 58, 59, 60, 61] (Fig 1). Each methodology has its own problem formulation, strengths and weaknesses, and ability to extract signal from limited observations. One encouraging trend within the deep learning community is a focus on reproducibility. This results in the rapid release of novel methods in the form of pre-prints and often associated example code, reporting new techniques as they are developed.

4 Improving Data Efficiency

4.1 Data Augmentation

Data Augmentation is a form of annotation efficient learning where one uses a series of predefined techniques to manipulate data samples to increase the input representations that correspond to a given label [51]. When considering computer vision, deep learning models learn to identify patterns within the numeric values represented as pixels. A simple example of augmentation to expand this representation is mirroring an image. When mirrored, the high-level concept of what is contained within the image remains unchanged, but the model sees an entirely new pixel representation. For computer vision, standardized image augmentation techniques include: translation, rotation, colour manipulation, additive Gaussian noise, random masking, light glare, even artificial weather conditions, among many others [51, 62, 63].

When training deep learning models, the parameters of a model are modified over multiple epochs. During each epoch, the model is fed each data sample. The key to the use of augmentation is that every time a data point is sampled, the series of augmentations used are randomly applied. In so doing, the model never sees identical images, forcing it to learn a general representation as opposed to memorizing the data. Deep learning models see the world by observing samples from a hypothetical "data generating distribution". Data augmentation intuitively can be viewed as a way of upweighting the tails of this distribution in a way that doesn't require collecting more data.

Data augmentation is primarily applied to scenarios where labeled data is limited, which is nearly all scenarios in ecology. Data augmentation is also applicable as a tool to mitigate class imbalance. When training, one can re-sample under-represented classes with a higher frequency while then applying aggressive augmentation [47]. An additional ecological boon is that, particular lighting and weather conditions augmentations can be applied to help models be robust to variable environmental conditions [64].

4.2 Simulators & Generative Models

When training deep learning models, it is often beneficial to provide addition data through synthetic means to inflate underrepresented classes, such as rare species. This data synthesis process can be performed through programmed simulators, or learned from data using a generative model. There are multiple forms of generative models including: Variational Autoencoders (VAEs), Flow-based models, Diffusion Models, and Generative Adversarial Networks (GANs) [65]. Below we focus on GANs because of their recent success and popularity.

Simulation is a form of generating additional data using human-coded programmatic rules. Simulated data can take many forms depending on the problem formulation. One problem common within ecology is domain shift, which includes scenarios in which classes and their background are correlated, biasing future predictions to behave the same [47, 66]. One can simulate example data by training a model to crop objects of interest from images, and paste these cutouts on new locations before, or during training [67]. More generally, to obtain individuals in new poses, researchers have used rendering engines to create synthetic examples of the classes of interest. Using these renders, one can then programatically manipulate the pose, environment, or general appearance [53, 54]. Creating renders can be expensive in terms of time and effort, however, if these renders or the engine that created them are released to the public domain the overhead of creating the model only needs to occur once for all to use, and the process becomes much more feasible.

Alternatively, GANs are a deep learning approach where, in computer vision, models are trained to create novel lifelike images conditioned on the domain of the training data. GANs train two models in competition with one another, a generator and discriminator. The generator is trained to create novel images conditioned from random noise, while the discriminator is trained to detect if the generator's images are real or fake. After training, the result is a model that can generate lifelike images of a desired domain [68, 69]. Using this approach, one can generate nearly endless novel images from limited datasets and under-represented classes [26]. One promising area of research is the use of GANs to generate not only the image, but corresponding labels as well. The end result is a 'labeled data factory' which can be applied to rare classes within a dataset [70].

For enhancing ecological data, generative models should be used as a tool to grow limited datasets, supplement under-represented classes, or in the case of labeled data factory, provide data and their annotations in bulk. This is not an exclusive list, but a subset of problems that may be overcome using data generation when data is limited for the use of deep learning systems.

4.3 Self-Supervised Learning

When referring to deep learning systems to this point, we have been primarily referring to traditional classifiers which produce a class label from an image considering a predefined list of possible options - a multiple choice question of which arthropod is the dominant subject of an image. To train these systems, the approach requires human annotators to provide a class label for every image within the data. For niche ecological problems, this is feasible only when considering a small number classes and only if one has the availability of experts to label the data.

When training traditional deep learning models with a softmax, multiple choice output, it is often thought that one requires class labels for all data samples. Due to the expensive nature of obtaining labels, this is sometimes infeasible, especially when requiring an expert to provide labels, as in ecology. One approach to utilize all of a partially labeled dataset is known as semi-supervised learning [71]. Semi-supervised learning exploits both labeled and unlabeled data for learning, usually in the setting where labeled data is restricted and unlabeled data is plentiful. One popular form of semi-supervised learning known as "pseudo-labeling" is a simple technique in which one first trains a model on the labeled data subset, followed then by using this model to predict the labels of the remaining unlabeled data. For each unlabeled input, deep learning models provide a predicted label as well as a confidence. Using these confidences, one then adds the predictions with high confidence to the training data along with the predicted "pseudo-labels" and repeats the process. While the model may make prediction errors, the overall process has been found to improve performance in comparison to considering only the labeled subset of data [71, 72].

Models limited to detect only expected classes, like supervised and semi-supervised, have a number of vulnerabilities. Such models are unable to expect unanticipated classes, such as invasive species, and cannot be used in different regions where other classes exist. For global initiatives, as we aim to be region agnostic and eventually increase the resolution of taxa beyond order, the labeling efforts required to train traditional classifiers quickly become infeasible. This is due to the number of fine-grained classes, geographic data imbalance, and the inevitable human error leading to label noise. Considering the extreme case of species, there are estimated to be millions of insect species in the world, all of which would require hundreds of expert labeled images [73]. Supervised deep learning models trained with human labels to answer a multiple choice question with millions of possible choices will not be the large scale solution to species-level entomology.

Self-supervised learning is an alternative approach that can generalize to classes not present in the original training data. To do this, self-supervised models operate on a proxy task, such as distinguishing if two input images are the same or different considering the domain from which the model was trained [59, 74]. How these two input images are selected depends on the availability of data labels. In the case of entomology, if one has taxa labels, one can select the same or different taxa, while if one has no labels, one can select a single image and apply two unique forms of augmentation to create two distinct samples [57, 75]. The result is a model trained to learn to distinguish if *any* input images of arthropods are the same or different taxa, extending to those never before seen in the training data [61]. This model then becomes agnostic to geographic region, capable of detecting invasive species, and does not require a library of labeled images. In practice, one would train a model for each taxa: order, family, genus, and species, and use the model appropriate for the task's granularity requirement. By training a performant comparison of taxa this way, the model becomes universal to data biases related to rarity and is applicable to comparisons from any geographic region in the world.

Self-supervised learning should be a tool used when: data labeling is unattainable, the data are bountiful but 'noisy' and difficult to label, the data do not contain a large representation of all the classes one would like to identify, or one would like their model to be robust to geographic region.

4.4 Real World Practicalities

The urgency of insect collapse falls back to one main motivation. What is the shortest path to improving the speed and accuracy of ecological predictions on a global scale? When we consider a global scale, this implies that machine learning methods be universal and used to empower data analyses in remote locations of the world. As attractive as machine learning approaches may be in their current form, as we outline above, there are still serious obstacles to overcome to achieve this objective of generality.

To offer pragmatic solutions in pursuit of a global arthropod deep learning system, the first general approach would be to aggregate as large of a universal dataset as possible and limit the scope of classifying arthropods to the order level. Using these data, one would then train a model with either the traditional classification or self-supervised approach, using data augmentation with synthetic data from a renderer or generative model. To measure model generality, one could then divide the data into training and testing relative to geographic regions, reporting performance classifying arthropod individuals from the withheld regions.

The result of training a model performant at the general task of order level arthropod classification would be the origin of a *foundation model* for entomology [76, 77]. Foundation models are models recognized as a tool that universally solve a particular task. Examples include: GPT-3 [78] for text generation, DALL-E 2 [79] for text-to-image generation, and the Megadetector for animal localization from camera trap images [80]. The creation of such tools have benefits that ripple beyond academic disciplines to institutional frameworks in need of efficient arthropod detection, such as the Food and Agriculture Organization (FAO) [81] and Institute for Nature and Environmental Protection (INEP) [82]. This comes at a time when there is a critical shortage of taxonomists in the world, especially in remote locations [83]. Even in its early stages, generalized deep learning models can be used to ease this shortage by allowing deep learning models to complement parataxonomists in remote locations of the world.

5 Focus on AI and Ecology Moving Forward

While we detail methods to improve the implementations of universal computer vision systems for entomology, there still exist a number of research challenges in computer vision that are required to be overcome. One scenario without a current solution is the separation of species that evolved to mimic the phenology of another [84]. Other scenarios that pose problems are taxa with variable appearances when the training data of these variations are underrepresented. Some of these scenarios include: wildly variable colourings across sex, species that undergo large phenotypic transformations over the course of their lifespan, such as *Lepidoptera* from caterpillars to butterflies, or images of taxa that have undergone some form of injury.

One area of rapid research is the use of cross-modality data. van Klink et al. [17] recently highlighted how deep learning for ecology has been well represented in four distinct modalities: computer vision, acoustics, radar, and molecular methods. Recent successes in deep learning research have shown training models that utilize a combination of these representations can improve performances over a single modality, especially for fine-grained classification tasks [85, 86, 87]. We believe there are vast numbers of research directions to explore considering multimodal ecological data. One area we believe has particular potential is to use DNA similarity as the measure of distance for self-supervised computer vision models [88, 89]. The result would be a model that can predict the genetic distance of two arthropods from their corresponding input images. Alternatively, there is an exciting area of research training generative models to create images of species considering only the DNA sequence as a prior. This problem formulation would follow the same text-to-image approach used to train DALL-E 2 [79]. Lastly, there has been success in combining DNA and image representations to predict class labels that exist in one modality that are not present in the other [90]. For example when training a model on complementary DNA and image data, while having robust DNA class labels but having only a subset of the total number of classes as images, models have been shown to predict the class of an image that was only represented as DNA during training [90]. This approach is known as zero-shot learning [91].

Lastly, while the approaches discussed here are have been largely focused around entomology, the annotation efficient and multi-modal learning techniques described are all general. These are applicable to nearly all data domains relevant to ecology and beyond. For example, the methods described can be used to inflate under-represented classes when considering camera trap data [47, 53]. Or, the multi-modal combinations of acoustics and vision could help identify species, such as birds with the task of bird classification [92].

At a high level, we are at an inflection point where accelerated methodological development is revolutionizing the approaches and discoveries of academic disciplines. Ecology is well-suited to benefit from this boom, as the ecological process of drawing trends from noisy data is a well-suited task for deep learning systems. The current limiting factor is providing the massive amount of labeled data required. To fully utilize deep learning systems, it will require a multi-faceted approach of data sharing, data organization, but also annotation efficient learning approaches. Here, we provided practical guidelines of such efforts to help overcome the limitations that face ecologists. The combination of all these approaches will allow ecologists to utilize ecological data to produce more general deep learning systems in pursuit of a general purpose foundation model of taxa classification. The future we are quickly approaching urgently needs the creation of a universal, region agnostic computer vision tool capable of identifying a globally broad range of taxa, including those rare and unexpected.

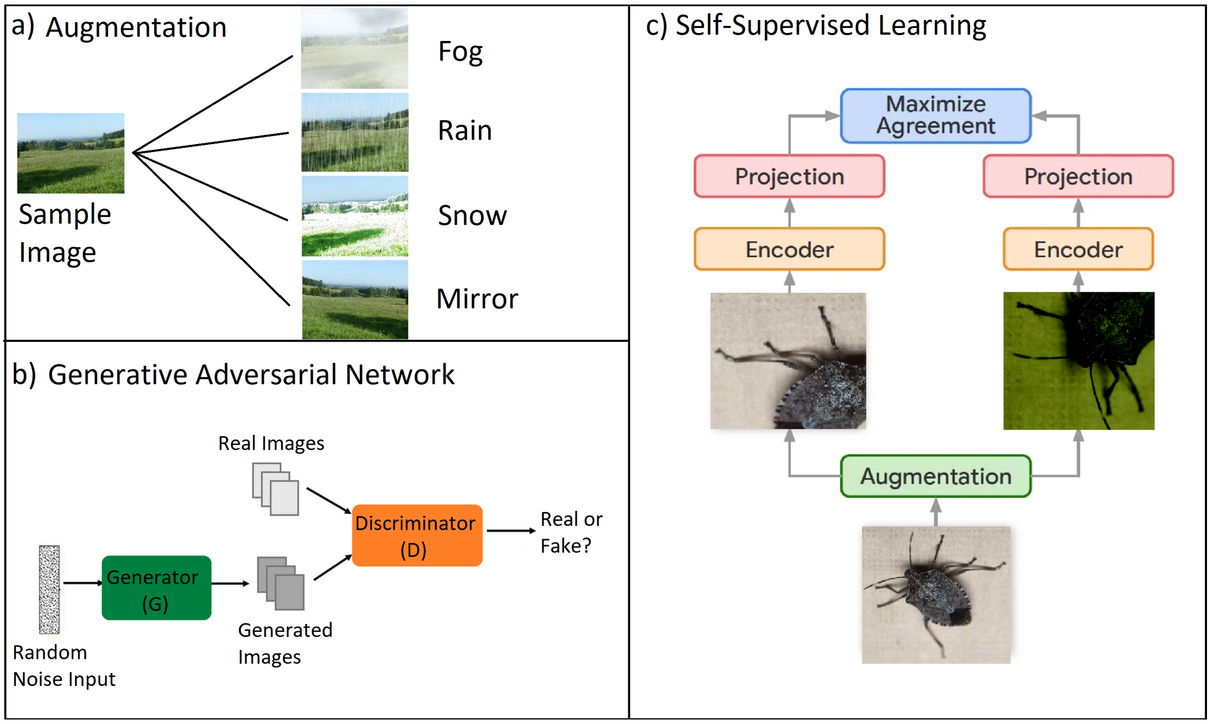


Figure 1: Visual summary of annotation efficient learning methods. a) Example augmentations. Exponentially increases the amount of data by randomly varying an image each time it is sampled. b) Example framework of a generative adversarial network. The trained generator is used to create additional images for training classifiers. c) Example framework for self-supervised learning. Images are sampled and randomly applied augmentation. The system learns similarity by predicting these images are still the same

References

[1] John Brand Free et al. *Insect pollination of crops*. Number Ed. 2. Academic press, 1993.

[2] Sarina Macfadyen, Rachel Gibson, Andrew Polaszek, Rebecca J Morris, Paul G Craze, Robert Planqué, William OC Symondson, and Jane Memmott. Do differences in food web structure between organic and conventional farms affect the ecosystem service of pest control? *Ecology letters*, 12(3):229–238, 2009.

[3] Shigeru Nakano, Hitoshi Miyasaka, and Naotoshi Kuhara. Terrestrial–aquatic linkages: riparian arthropod inputs alter trophic cascades in a stream food web. *Ecology*, 80(7): 2435–2441, 1999.

[4] Caspar A Hallmann, Martin Sorg, Eelke Jongejans, Henk Siepel, Nick Hofland, Heinz Schwan, Werner Stenmans, Andreas Müller, Hubert Sumser, Thomas Hörren, et al. More

than 75 percent decline over 27 years in total flying insect biomass in protected areas. *PloS one*, 12(10):e0185809, 2017.

[5] Francisco Sánchez-Bayo and Kris AG Wyckhuys. Worldwide decline of the entomofauna: A review of its drivers. *Biological conservation*, 232:8–27, 2019.

[6] Sebastian Seibold, Martin M Gossner, Nadja K Simons, Nico Blüthgen, Jörg Müller, Didem Ambarlı, Christian Ammer, Jürgen Bauhus, Markus Fischer, Jan C Habel, et al. Arthropod decline in grasslands and forests is associated with landscape-level drivers. *Nature*, 574 (7780):671–674, 2019.

[7] David L Wagner. Insect declines in the anthropocene. *Annual review of entomology*, 65: 457–480, 2020.

[8] C Kremen, RK Colwell, TL Erwin, DD Murphy, RF Noss, , and MA Sanjayan. Terrestrial arthropod assemblages: their use in conservation planning. *Conservation biology*, pages 796–808, 1993.

[9] Elizabeth T Borer, Eric W Seabloom, and David Tilman. Plant diversity controls arthropod biomass and temporal stability. *Ecology letters*, 15(12):1457–1464, 2012.

[10] Teja Tscharntke, Jason M Tylianakis, Tatyana A Rand, Raphael K Didham, Lenore Fahrig, Péter Batáry, Janne Bengtsson, Yann Clough, Thomas O Crist, Carsten F Dormann, et al. Landscape moderation of biodiversity patterns and processes-eight hypotheses. *Biological reviews*, 87(3):661–685, 2012.

[11] Johanna Ärje, Claus Melvad, Mads Rosenhøj Jeppesen, Sigurd Agerskov Madsen, Jenni Raitoharju, Maria Strandgård Rasmussen, Alexandros Iosifidis, Ville Tirronen, Moncef Gabbouj, Kristian Meissner, et al. Automatic image-based identification and biomass estimation of invertebrates. *Methods in Ecology and Evolution*, 11(8):922–931, 2020.

[12] Paul Tresson, Dominique Carval, Philippe Tixier, and William Puech. Hierarchical classification of very small objects: Application to the detection of arthropod species. *IEEE Access*, 9:63925–63932, 2021.

[13] Duhita Wani and Tomas Maul. Image super-resolution for arthropod identification. In *2021 4th International Conference on Computer Science and Software Engineering (CSSE 2021)*, pages 317–324, 2021.

[14] Pierce Helton, Khoa Luu, and Ashley Dowling. Artificial intelligence system for automatic imaging, quantification, and identification of arthropods in leaf litter and pitfall samples. *Inquiry: The University of Arkansas Undergraduate Research Journal*, 21(1):5, 2022.

[15] Stefan Schneider, Graham W Taylor, Stefan C Kremer, Patrick Burgess, Jillian McGroarty, Kyomi Mitsui, Alex Zhuang, Jeremy R deWaard, and John M Fryxell. Bulk arthropod abundance, biomass and diversity estimation using deep learning for computer vision. *Methods in Ecology and Evolution*, 13(2):346–357, 2022.

[16] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[17] Roel van Klink, Tom August, Yves Bas, Paul Bodesheim, Aletta Bonn, Frode Fossøy, Toke T. Høye, Eelke Jongejans, Myles H.M. Menz, Andreia Miraldo, Tomas Roslin, Helen E. Roy, Ireneusz Ruczyński, Dmitry Schigel, Livia Schäffler, Julie K. Sheard, Cecilie Svenningsen, Georg F. Tschan, Jana Wäldchen, Vera M.A. Zizka, Jens Åström, and Diana E. Bowler. Emerging technologies revolutionise insect ecology and monitoring. *Trends in Ecology & Evolution*, 2022. ISSN 0169-5347. doi: https://doi.org/10.1016/j.tree.2022.06.001. URL `https://www.sciencedirect.com/science/article/pii/S0169534722001343`.

[18] Alexander Gerovichev, Achiad Sadeh, Vlad Winter, Avi Bar-Massada, Tamar Keasar, and Chen Keasar. High throughput data acquisition and deep learning for insect ecoinformatics. *Frontiers in Ecology and Evolution*, 9:309, 2021.

[19] Weiguang Ding and Graham Taylor. Automatic moth detection from trap images for pest management. *Computers and Electronics in Agriculture*, 123:17–28, 2016.

[20] Le-Qing Zhu, Meng-Yuan Ma, Zhen Zhang, Pei-Yi Zhang, Wei Wu, Da-Dong Wang, Da-Xing Zhang, Xun Wang, and Hui-Yan Wang. Hybrid deep learning for automated lepidopteran insect image classification. *Oriental Insects*, 51(2):79–91, 2017.

[21] Everton Castelão Tetila, Bruno Brandoli Machado, Geazy Vilharva Menezes, Nicolas Alessandro de Souza Belete, Gilberto Astolfi, and Hemerson Pistori. A deep-learning approach for automatic counting of soybean insect pests. *IEEE Geoscience and Remote Sensing Letters*, 17(10):1837–1841, 2019.

[22] Dimitri Korsch, Paul Bodesheim, and Joachim Denzler. Deep learning pipeline for automated visual moth monitoring: insect localization and species classification. *INFORMATIK 2021*, 2021.

[23] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE international conference on computer vision*, pages 3754–3762, 2017.

[24] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 289–293. IEEE, 2018.

[25] Changhee Han, Hideaki Hayashi, Leonardo Rundo, Ryosuke Araki, Wataru Shimoda, Shinichi Muramatsu, Yujiro Furukawa, Giancarlo Mauri, and Hideki Nakayama. Gan-based synthetic brain mr image generation. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 734–738. IEEE, 2018.

[26] Xu Cao, Ziyi Wei, Yinjie Gao, and Yingqiu Huo. Recognition of common insect in field based on deep learning. In *Journal of Physics: Conference Series*, volume 1634, page 012034. IOP Publishing, 2020.

[27] Sefik Emre Eskimez, Dimitrios Dimitriadis, Robert Gmyr, and Kenichi Kumanati. Gan-based data generation for speech emotion recognition. In *INTERSPEECH*, pages 3446–3450, 2020.

[28] Oskar LP Hansen, Jens-Christian Svenning, Kent Olsen, Steen Dupont, Beulah H Garner, Alexandros Iosifidis, Benjamin W Price, and Toke T Høye. Species-level image classification with convolutional neural network enables insect identification from habitus images. *Ecology and evolution*, 10(2):737–747, 2020.

[29] Dongjun Xin, Yen-Wei Chen, and Jianjun Li. Fine-grained butterfly classification in ecological images using squeeze-and-excitation and spatial attention modules. *Applied Sciences*, 10(5):1681, 2020.

[30] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[31] Gao Huang, Shichen Liu, Laurens Van der Maaten, and Kilian Q Weinberger. Condensenet: An efficient densenet using learned group convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2752–2761, 2018.

[32] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[33] Alan Caio R Marques, Marcos M. Raimundo, Ellen Marianne B. Cavalheiro, Luis FP Salles, Christiano Lyra, and Fernando J. Von Zuben. Ant genera identification using an ensemble of convolutional neural networks. *Plos one*, 13(1):e0192011, 2018.

[34] Johanna Ärje, Jenni Raitoharju, Alexandros Iosifidis, Ville Tirronen, Kristian Meissner, Moncef Gabbouj, Serkan Kiranyaz, and Salme Kärkkäinen. Human experts vs. machines in taxa recognition. *Signal Processing: Image Communication*, 87:115917, 2020.

[35] Dan Jeric Arcega Rustia, Jun-Jee Chao, Jui-Yung Chung, and Ta-Te Lin. An online unsupervised deep learning approach for an automated pest insect monitoring system. In *2019 ASABE Annual International Meeting*, page 1. American Society of Agricultural and Biological Engineers, 2019.

[36] Denan Xia, Peng Chen, Bing Wang, Jun Zhang, and Chengjun Xie. Insect detection and classification based on an improved convolutional neural network. *Sensors*, 18(12):4169, 2018.

[37] Daniel Motta, Alex Álisson Bandeira Santos, Ingrid Winkler, Bruna Aparecida Souza Machado, Daniel André Dias Imperial Pereira, Alexandre Morais Cavalcanti, Eduardo Oyama Lins Fonseca, Frank Kirchner, and Roberto Badaró. Application of convolutional neural networks for classification of adult mosquitoes in the field. *PloS one*, 14(1):e0210829, 2019.

[38] Midori Tuda and Alejandro Isabel Luna-Maldonado. Image-based insect species and gender classification by trained supervised machine learning algorithms. *Ecological Informatics*, 60:101135, 2020.

[39] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[40] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

[41] Miklos Balint, Markus Pfenninger, Hans-Peter Grossart, Pierre Taberlet, Mark Vellend, Mathew A Leibold, Göran Englund, and Diana Bowler. Environmental dna time series in ecology. *Trends in Ecology & Evolution*, 33(12):945–957, 2018.

[42] Toke T Høye, Johanna Ärje, Kim Bjerge, Oskar LP Hansen, Alexandros Iosifidis, Florian Leese, Hjalte MR Mann, Kristian Meissner, Claus Melvad, and Jenni Raitoharju. Deep learning and computer vision will transform entomology. *Proceedings of the National Academy of Sciences*, 118(2), 2021.

[43] Marie I Tosa, Emily H Dziedzic, Cara L Appel, Jenny Urbina, Aimee Massey, Joel Ruprecht, Charlotte E Eriksson, Jane E Dolliver, Damon B Lesmeister, Matthew G Betts, et al. The rapid rise of next-generation natural history. *Frontiers in Ecology and Evolution*, 9:698131, 2021.

[44] Jianqing Fan, Cong Ma, and Yiqiao Zhong. A selective overview of deep learning. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 36(2):264, 2021.

[45] Joffrey L Leevy, Taghi M Khoshgoftaar, Richard A Bauder, and Naeem Seliya. A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5(1):1–30, 2018.

[46] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

[47] Stefan Schneider, Saul Greenberg, Graham W Taylor, and Stefan C Kremer. Three critical factors affecting automated image species recognition performance for camera traps. *Ecology and evolution*, 10(7):3503–3517, 2020.

[48] Deng-Qi Yang, Tao Li, Meng-Tao Liu, Xiao-Wei Li, and Ben-Hui Chen. A systematic study of the class imbalance problem: Automatically identifying empty camera trap images using convolutional neural networks. *Ecological Informatics*, 64:101350, 2021.

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

[50] Agnieszka Mikołajczyk and Michał Grochowski. Data augmentation for improving deep learning in image classification problem. In *2018 international interdisciplinary PhD workshop (IIPhDW)*, pages 117–122. IEEE, 2018.

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

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

[53] Sara Beery, Yang Liu, Dan Morris, Jim Piavis, Ashish Kapoor, Neel Joshi, Markus Meister, and Pietro Perona. Synthetic examples improve generalization for rare classes. In

Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 863–873, 2020.

[54] Sergey I Nikolenko. *Synthetic data for deep learning*, volume 174. Springer, 2021.

[55] Subhajit Chatterjee, Debapriya Hazra, Yung-Cheol Byun, and Yong-Woon Kim. Enhancement of image classification using transfer learning and gan-based synthetic data augmentation. *Mathematics*, 10(9):1541, 2022.

[56] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1476–1485, 2019.

[57] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

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

[59] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

[60] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43 (11):4037–4058, 2020.

[61] Stefan Schneider, Graham W Taylor, and Stefan C Kremer. Similarity learning networks for animal individual re-identification-beyond the capabilities of a human observer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 44–52, 2020.

[62] Ilya Kostrikov, Denis Yarats, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. *arXiv preprint arXiv:2004.13649*, 2020.

[63] Alexander B. Jung. imgaug. `https://github.com/aleju/imgaug`, 2018. [Online].

[64] Quoc-Viet Hoang, Trung-Hieu Le, and Shih-Chia Huang. Data augmentation for improving ssd performance in rainy weather conditions. In *2020 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan)*, pages 1–2. IEEE, 2020.

[65] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *arXiv preprint arXiv:2103.04922*, 2021.

[66] Michael A Tabak, Mohammad S Norouzzadeh, David W Wolfson, Steven J Sweeney, Kurt C VerCauteren, Nathan P Snow, Joseph M Halseth, Paul A Di Salvo, Jesse S Lewis, Michael D White, et al. Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, 10(4):585–590, 2019.

[67] Stefan Schneider and Alex Zhuang. Counting fish and dolphins in sonar images using deep learning. *arXiv preprint arXiv:2007.12808*, 2020.

[68] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[69] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019.

[70] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10145–10155, 2021.

[71] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

[72] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[73] Paul Eggleton. The state of the world's insects. *Annu. Rev. Environ. Resour*, 45:61–82, 2020.

[74] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[75] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[76] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[77] Alexandre Lacoste, Evan David Sherwin, Hannah Kerner, Hamed Alemohammad, Björn Lütjens, Jeremy Irvin, David Dao, Alex Chang, Mehmet Gunturkun, Alexandre Drouin, et al. Toward foundation models for earth monitoring: Proposal for a climate change benchmark. *arXiv preprint arXiv:2112.00570*, 2021.

[78] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[79] OpenAI. Dall-E 2. `https://openai.com/dall-e-2/`, 2022. [Online].

[80] Sara Beery, Dan Morris, and Siyu Yang. Efficient pipeline for camera trap image review. *arXiv preprint arXiv:1907.06772*, 2019.

[81] Food and Agricultural Organization of the United Nations. `https://www.fao.org/home/en`, 2022. [Online].

[82] Institute of Nature and Environmental Conservation. https://www.inecgh.org/, 2022. [Online].

[83] Michael S Engel, Luis MP Ceríaco, Gimo M Daniel, Pablo M Dellapé, Ivan Löbl, Milen Marinov, Roberto E Reis, Mark T Young, Alain Dubois, Ishan Agarwal, et al. The taxonomic impediment: a shortage of taxonomists, not the lack of technical approaches, 2021.

[84] Camille Garcin, Alexis Joly, Pierre Bonnet, Jean-Christophe Lombardo, Antoine Affouard, Mathias Chouet, Maximilien Servajean, Joseph Salmon, and Titouan Lorieul. Pl@ ntnet-300k: a plant image dataset with high label ambiguity and a long-tailed distribution. In *NeurIPS 2021-35th Conference on Neural Information Processing Systems*, 2021.

[85] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12486, 2021.

[86] Jabeen Summaira, Xi Li, Amin Muhammad Shoib, Songyuan Li, and Jabbar Abdul. Recent advances and trends in multimodal deep learning: A review. *arXiv preprint arXiv:2105.11087*, 2021.

[87] Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. Multimodal deep learning for biomedical data fusion: a review. *Briefings in Bioinformatics*, 23(2):bbab569, 2022.

[88] Xin Jin, Qian Jiang, Yanyan Chen, Shin-Jye Lee, Rencan Nie, Shaowen Yao, Dongming Zhou, and Kangjian He. Similarity/dissimilarity calculation methods of dna sequences: a survey. *Journal of Molecular Graphics and Modelling*, 76:342–355, 2017.

[89] Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020.

[90] Sarkhan Badirli, Zeynep Akata, George Mohler, Christine Picard, and Mehmet M Dundar. Fine-grained zero-shot learning with dna as side information. *Advances in Neural Information Processing Systems*, 34:19352–19362, 2021.

[91] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

[92] Dan Stowell, Michael D Wood, Hanna Pamuła, Yannis Stylianou, and Hervé Glotin. Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge. *Methods in Ecology and Evolution*, 10(3):368–380, 2019.