Machine Learning Challenges of Biological Factors in Insect Image Data

Nicholas Pellegrino Vision and Image Processing Group, System Design Engineering, University of Waterloo
Zahra Gharaee Vision and Image Processing Group, System Design Engineering, University of Waterloo
Paul Fieguth Vision and Image Processing Group, System Design Engineering, University of Waterloo
Email: {npellegr, zgharaee, pfieguth}@uwaterloo.ca

Abstract

The BIOSCAN project, led by the *International Barcode of Life Consortium*, seeks to study changes in biodiversity on a global scale. One component of the project is focused on studying the species interaction and dynamics of all insects. In addition to genetically barcoding insects, over 1.5 *million* images per *year* will be collected, each needing taxonomic classification. With the immense volume of incoming images, relying solely on expert taxonomists to label the images would be impossible; however, artificial intelligence and computer vision technology may offer a viable high-throughput solution. Additional tasks including manually weighing individual insects to determine biomass, remain tedious and costly. Here again, computer vision may offer an efficient and compelling alternative. While the use of computer vision methods is appealing for addressing these problems, significant challenges resulting from biological factors present themselves. These challenges are formulated in the context of machine learning in this paper.

1 Introduction

The BIOSCAN project [1], led by the *International Barcode of Life Consortium* (iBOL) at the University of Guelph, is a large-scale, multinational research project with three main areas of study and corresponding objectives:

1. **Species Discovery**: Generate genetic barcode [2] coverage for *two million* species.
2. **Species Interactions**: Reveal species *interactions* by targeting the symbiome.
3. **Species Dynamics**: Monitor *biodiversity* over time at 2,000 sites.

To achieve this, at least 10 million life-specimens will be collected throughout the course of this project. One particular component of the project focuses on studying insects, whereby insects from around the world will be both genetically barcoded and imaged. In fact, the plan is to collect over 1.5 *million* high-resolution images per *year*, with each one needing taxonomic classification. With the immense volume of incoming images, relying solely on expert taxonomists to label the images would be impossible; however, artificial intelligence (AI) and computer vision (CV) technology may offer a viable high-throughput solution.

In addition to taxonomic classification, the use of CV / AI may enable further information such as insect biomass and insect orientation and pose to be inferred, useful for downstream tasks. Furthermore, it may even be possible to correct DNA sequencing errors where the measured sequence differs substantially from what is expected based on the image-based taxonomic assignment. Figure 1 graphically illustrates the information we wish to extract / estimate from insect images to support the BIOSCAN project. The Centre for Biodiversity and Genomics, through BOLD [3], has provided a preliminary data set of images with included taxonomic annotations; Figure 2 shows six example insect images from this data set. Across these images, large variation in insect size (scale), colour, transparency, orientation, pose and illumination is seen, behaviours that any estimator or information extraction tool would need to contend with.

2 Proposed Scientific Outcomes

This project is motivated by the opportunity for amazing scientific discoveries to be made that may be enabled through the effective application of machine learning and computer vision technology. This section briefly highlights several of such possible scientific outcomes.

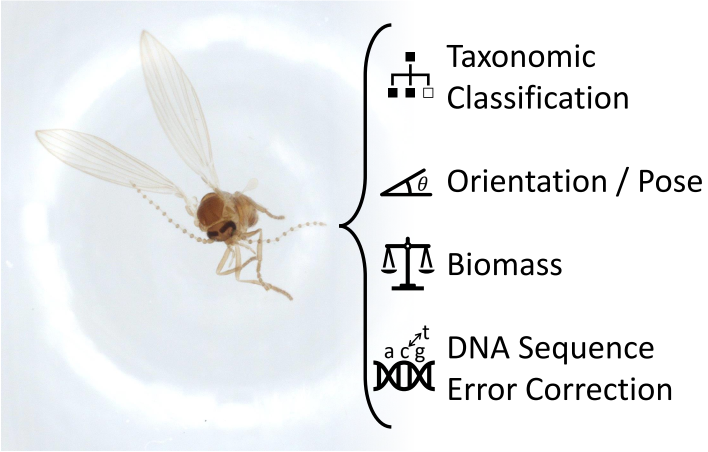

Fig. 1: Valuable information may be inferred from images of single insects. AI methods have the potential to automatically extract such information. Insect imaged by the Centre for Biodiversity Genomics.

2.1 Biomass Estimation from Images

Many reports world-wide indicate a grave decline in insect biomass [4, 5] over long-term periods, where not only a significant loss of in insect life in bulk is projected to continue into the future, but a strongly correlated catastrophic drop in biodiversity (i.e., species loss).

The use of insect traps (malaise traps) remains a common approach for monitoring insect abundance and biomass [4, 6]; however, the individual weighing of insects is still required. By replacing the tedious task of weighing insects manually with the automatic process of computer-vision-based biomass estimation, the broader task of processing insects in the lab is streamlined, thus saving valuable time / resources of the BIOSCAN project for more impactful duties.

2.2 Individual Species Identification through Bulk Insect Metabarcoding

Bulk metabarcoding uses polymerase chain reaction (PCR) amplification technology to identify biological samples comprising hundreds or thousands of specimens [7–9]. In particular, this approach has been used for bulk barcoding of arthropods (the taxon containing insects).

The availability of this method means that insect traps can be set up throughout forests and other ecosystems, collecting thousands of insects, and then through bulk processing, barcode information can (remarkably!) be produced for each *individual* species present in the bulk sample.

The main reason for using this approach is to drive down the *unit cost* of genetic barcoding. Previously, insects were carefully handled on an individual basis, meticulously being placed into separate test-tubes, before being genetically barcoded. The cost of labour to take on such an endeavour is enormous. However, by barcoding large groups of insects all at once, the cost is dramatically reduced and makes such a project all the more feasible.

Although bulk metabarcoding is already used today, the main contribution of the project in this area is the enormous *amount* of samples that are being collected and which can efficiently be barcoded using this method.

2.3 Species Interaction and Dynamics

Beyond studying insects, iBOL seeks to barcode *all* multicellular life. By doing so, it may become possible to, for example, sample an ecosystem by picking a leaf from a tree, extracting the DNA in

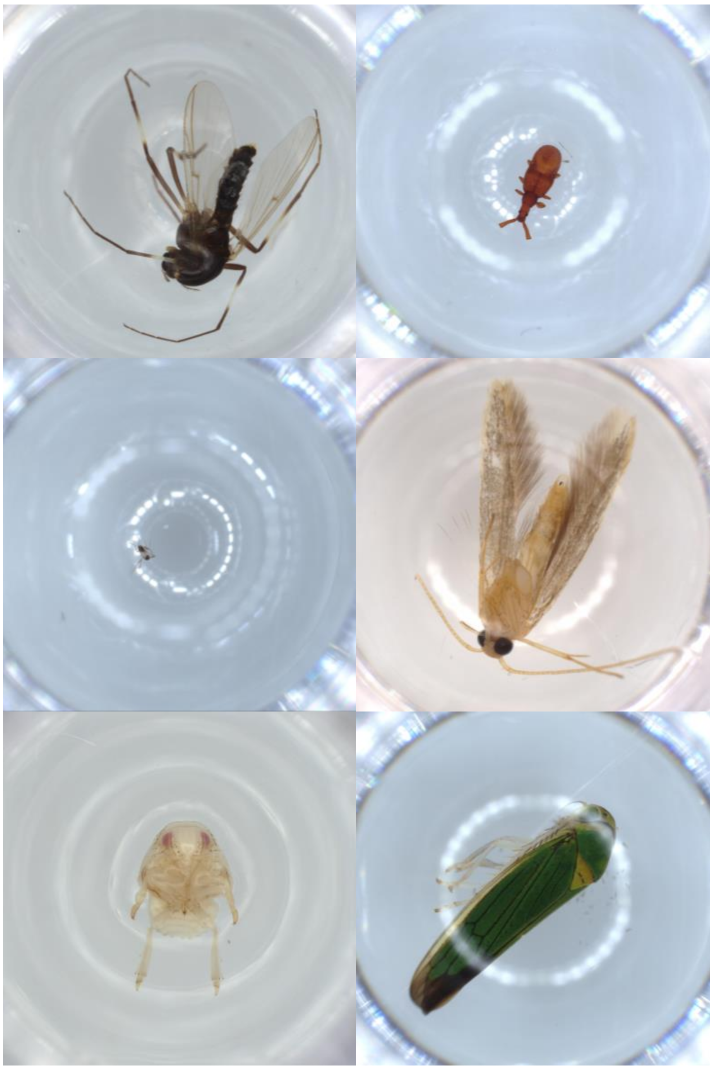

Fig. 2: Composite of six insect images included in the data set. Observe the variety in insect size (scale), colour, transparency, orientation, pose, and illumination. Insects imaged by the Centre for Biodiversity Genomics.

bulk from all organisms on that leaf, and finally processing through barcode analysis to identify all (multicellular) life on that leaf. Such an approach would reveal the species interaction between the plant (host) and everything else living on it, perhaps ranging from lichens, algae, and fungi, to nematodes and perhaps even insects. Such a sampling and analysis technique offers far greater geographical resolution / precision than relatively non-discriminatory insect traps placed throughout biomes, allowing for the study of *micro*-ecosystems.

Species interaction can be studied and made use of in other areas as well. Many insects feed on or parasitize other larger plants or animals within their ecosystems and thus contain some DNA from those organisms within their digestive tracts / guts. One such insect are mosquitoes, which regularly feed on animal blood. By catching and performing barcode analysis on mosquitoes found within a given ecosystem, much can be learned about the presence and perhaps even abundance of the animals from which they draw blood. This sort of analysis may give valuable insights into species *interactions* and, if repeated over spans of time, would illuminate details of species and ecosystem *dynamics*.

3 Machine Learning Challenges & Directions

3.1 Limited Available Training Data for ML Tasks

Data annotation is both a costly and tedious process, necessary to create *labelled* training data for downstream ML tasks. In particular, individually imaging, taxonomically classifying (by an expert), and weighing insects is an immense undertaking. As a result, while iBOL has the ability to acquire millions of images annually, it *does not* have the ability to label them at a similar rate while remaining cost-effective. Gold standard data sets within the CV community, in particular object detection and recognition data sets, often include several million images, e.g., ImageNet [10] with ∼14M images, Open Images [11] with ∼9M images, and Microsoft Common Objects in Context (COCO) [12] with ∼2.5M images. The great abundance of labelled data makes these data sets conducive to ML training tasks. While the data set currently available for this project is composed of roughly 1 million labelled images, many of these images are unusable as a result of processing issues whereby contaminants from other insects have compromised the associated barcode information, leaving closer to 200,000 vetted records. As a result of the having relatively limited training data, conventional methods or network architectures, which are normally effective when vast amounts of training data are available, are likely to perform poorly. Beyond training, the lack of labelled data creates a complication for model testing (i.e., determining whether the model is able to generalize to new, unseen data), since the limited data set may not adequately represent the variation present within the entire broad taxonomic class of insects.

One might imagine augmenting the available data with labelled images from other data sets. In principle, this would help; however, to the best of the authors' knowledge, the largest available insect data set, IP102 [13] contains only 75,000 images, and of which only 19,000 are annotated. Furthermore, the IP102 data set is quite limited in scope, containing only images of 102 insect species. Because of its relatively small size in comparison to the images already available through BOLD, differences in formatting, and image quality / staging, integrating the two data sets would be difficult while unfortunately providing little benefit.

3.2 Intrinsic Challenges Associated with Insect Data

Beyond the *quantity* of data available, as was discussed in Section 3.1, several complicating factors, intrinsic to the data set, exist. These are listed here, each with some discussion.

1. **Class imbalance**:
   An imbalance exists between classes, exhibiting a long-tailed distribution, characteristic of real-world biological data. This causes difficulty when training classification models because relatively few classes comprise the majority of the data set. In fact, at the taxonomic order-level, the order diptera (flies) accounts for the majority of all available training data, comprising roughly 70% of the entire data set. Figure 3 shows the relative frequency distribution of samples at across all taxonomic orders present. The long-tailed distribution is quite apparent. This degree of imbalance means that classifiers are able to achieve considerable accuracy (∼70%) on the *available* data set simply by *always* predicting diptera and without actually *learning* the characteristics of diptera vs. other orders [14]! Furthermore, there are many taxonomic orders with *very few* training images available, posing additional challenge.
2. **Variation in insect Size**:
   Insects that are especially small appear quite small in the images, whereas other larger insects may occupy the entire frame, as seen in Figure 2. This means that the sorts of resolvable physical features of the insects vary considerably with their size. Furthermore, by virtue of occupying fewer image pixels, less information is actually captured in images of smaller insects.
3. **Visual similarity of distinct insect species**:
   Distinct insect species may appear visually similar for a variety of reasons, most common of which is high genetic / taxonomic similarity. For example, many beetles appear visually similar and are characterized by traits such as their rounded shell, often covering a pair of slightly transparent wings. Other reasons for visual similarity between distinct species include mimicry / imitation [15–17], or more broadly, convergent evolution [18].
4. **Visual differences based on pose and orientation of the insect samples**:
   Insects exhibit bilateral symmetry, however, that symmetry is immediately broken in images where insects are not viewed parallel to their sagittal plane, a result of their *orientation* in the image. Furthermore, this symmetry is broken by differences between left and right in the insects' *pose*, whereby left and right limbs or joints may be at differing angles. Beyond bodily symmetry, it is clear that insects appear differ-

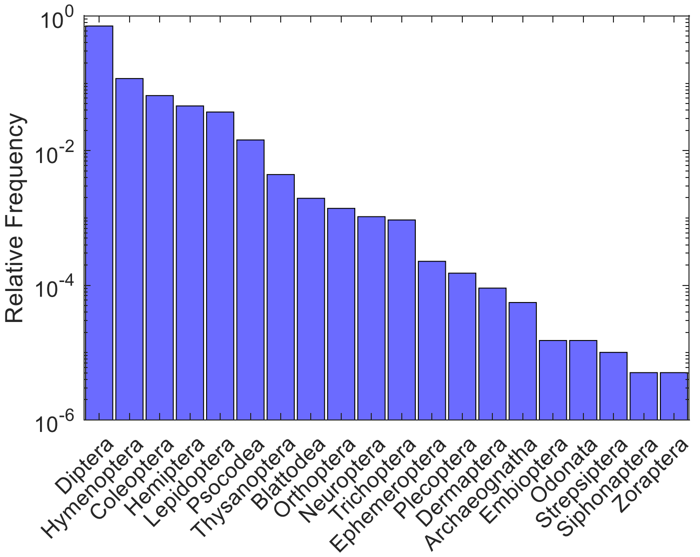

Fig. 3: Histogram showing relative frequency of each insect taxonomic order present within the data set, forming a long-tailed distribution. Notice the log-scaled vertical axis. The order diptera (flies) accounts for roughly 70% of all samples.

ently when viewed from different angles (front vs. back, top vs. bottom, etc.) and when limbs are differently positioned.

5. **Metamorphosis and variation associated with age / stage within life cycle**:
   Insects undergo enormous physical transformations as they transition through various stages of life, from eggs, to larvae or nymphs, to pupae, and finally adults. Consider the order lepidoptera, consisting of butterflies and moths, which transform dramatically throughout their life — transitioning from caterpillar, to chrysalis, and finally to adult butterfly/moth — all the while remaining taxonomically identical! These changes throughout the life cycle of insects result in significant intra-class variation, but also may increase inter-class similarity, given that many insect larvae are "worm-like", and even worse, all insect eggs can be described as oblong elliptic structures, though considerable variations in size (8 orders of magnitude) and aspect ratio do exist [19].
6. **Sexual dimorphism**:
   Sexual dimorphism, *meaning two forms*, is the physical difference between taxonomically identical insects (i.e., same species) of opposite sex. Examples range from difference in overall size, to proportions (e.g., limbs or antennae of differing lengths), to colouration and patterning [20], and even whether an insect poses certain body parts (e.g., cephalic and/or pronotal horns in male scarab beetles [21]). As in the preceding example, such differences contribute to the intra-class variance.

Many of these problematic biological challenges can be understood and succinctly summarized as factors that contribute to causing significant intra-class variance, while at the same time decreasing inter-class separation (i.e., increasing inter-class similarity). Such challenges often present themselves in the domain of fine-grained classification, a category of ML / CV problems whereby classes must be discriminated based on subtle or localized differences [22, 23].

3.3 Possible Directions Forward

Several potential and overlapping paths forward exist to address the challenges associated with the lack of data, high intra-class variance, and high inter-class similarity. These include the following.

1. **Data augmentation and re-sampling**:
   Since *some*, roughly 200,000, images do have labels already assigned to them, perhaps data augmentation [24, 25] can be used to artificially inflate the size of the available training data set, thus enabling the use of conventional / tried-and-true CV methods. Random re-sampling [26] is another approach that can be applied to simultaneously address the between-class and within-class imbalances problems.
2. **Sparse models**:
   In the absence of a suitably large training data set, careful selection of DNN/CNN model architectures may be necessary. Specifically, sparse models — those with relatively *few* weights — may offer a suitably generalized solution such that over-fitting to the small training data set can be avoided [27].
3. **Alternative learning techniques**:
   A further alternative would be to employ techniques such as domain adaptation [28, 29], transfer learning [30, 31], domain generalization through feature representation [32, 33] or meta-learning [34, 35], perhaps allowing a network model with sufficient object — possibly specifically *animal* — classification ability to be fine-tuned using the limited insect data set such that it is able to differentiate and adequately classify insects.
4. **Alternative loss functions**:
   Certain loss functions are designed in ways that may mitigate the impact of various inadequacies in the training data. One such example is the Focal Loss [36], which is designed specifically to address issues of class imbalance, and does so by weighting the loss associated with *"difficult"* training examples — those likely to be misclassified — more than training samples that are "easy" or likely to be classified correctly.

4 Conclusions

While numerous challenges prevent the simple application of existing computer vision techniques to assist in the BIOSCAN project, addressing these challenges would be a contribution to the ML / AI field, while also being highly meaningful to the broader international community of BIOSCAN collaborators. Future works will involve implementations and evaluations of the the proposed ML strategies for overcoming the challenges formulated in this paper. Such contributions may enable the tracking of changes in insect abundance, biomass, and diversity at the species level over time and across a multitude of geographical locations. This information is critical to understanding the impacts of ecological change and for formulating strategies for mitigating further catastrophic damage to the global ecosystem.

Acknowledgments

The authors thank the Centre for Biodiversity Genomics at the University of Guelph for access to the insect data within BOLD [3].

References

[1] "Bioscan," Jun 2022. [Online]. Available: https://ibol.org/programs/bioscan/

[2] P. D. Hebert, A. Cywinska, S. L. Ball, and J. R. DeWaard, "Biological identifications through dna barcodes," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 270, no. 1512, pp. 313–321, 2003.

[3] "Barcode of life data system." [Online]. Available: https://boldsystems.org/

[4] C. A. Hallmann, M. Sorg, E. Jongejans, H. Siepel, N. Hofland, H. Schwan, W. Stenmans, A. Müller, H. Sumser, T. Hörren *et al.*, "More than 75 percent decline over 27 years in total flying insect biomass in protected areas," *PloS one*, vol. 12, no. 10, p. e0185809, 2017.

[5] C. A. Hallmann, A. Ssymank, M. Sorg, H. de Kroon, and E. Jongejans, "Insect biomass decline scaled to species diversity: General patterns derived from a hoverfly community," *Proceedings of the National Academy of Sciences*, vol. 118, no. 2, p. e2002554117, 2021.

[6] M. J. Skvarla, J. L. Larson, J. R. Fisher, and A. P. Dowling, "A review of terrestrial and canopy malaise traps," *Annals of the Entomological Society of America*, vol. 114, no. 1, pp. 27–47, 2021.

[7] P. Taberlet, E. Coissac, F. Pompanon, C. Brochmann, and E. Willerslev, "Towards next-generation biodiversity assessment using dna metabarcoding," *Molecular ecology*, vol. 21, no. 8, pp. 2045–2050, 2012.

[8] J. Gibson, S. Shokralla, T. M. Porter, I. King, S. van Konynenburg, D. H. Janzen, W. Hallwachs, and M. Hajibabaei, "Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through dna metasystematics," *Proceedings of the National Academy of Sciences*, vol. 111, no. 22, pp. 8007–8012, 2014.

[9] C. Lynggaard, M. Nielsen, L. Santos-Bay, M. Gastauer, G. Oliveira, and K. Bohmann, "Vertebrate diversity revealed by metabarcoding of bulk arthropod samples from tropical forests," *Environmental DNA*, vol. 1, no. 4, pp. 329–341, 2019.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[11] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, A. Veit *et al.*, "Openimages: A public dataset for large-scale multi-label and multi-class image classification," *Dataset available from https://github. com/openimages*, vol. 2, no. 3, p. 18, 2017.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[13] X. Wu, C. Zhan, Y.-K. Lai, M.-M. Cheng, and J. Yang, "Ip102: A large-scale benchmark dataset for insect pest recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8787–8796.

[14] J. Harvie, "Computer vision based taxonomic classification of insecta using deep learning models," Ph.D. dissertation, University of Guelph, 2022.

[15] C. W. Rettenmeyer, "Insect mimicry," *Annual review of entomology*, vol. 15, no. 1, pp. 43–74, 1970.

[16] J. van Zandt Brower, "Experimental studies of mimicry in some north american butterflies: Part i. the monarch, danaus plexippus, and viceroy, limenitis archippus archippus," *Evolution*, pp. 32–47, 1958.

[17] D. B. Ritland and L. P. Brower, "The viceroy butterfly is not a batesian mimic," *Nature*, vol. 350, no. 6318, pp. 497–498, 1991.

[18] M. Maruyama and J. Parker, "Deep-time convergence in rove beetle symbionts of army ants," *Current Biology*, vol. 27, no. 6, pp. 920–926, 2017.

[19] S. H. Church, S. Donoughe, B. A. de Medeiros, and C. G. Extavour, "A dataset of egg size and shape from more than 6,700 insect species," *Scientific data*, vol. 6, no. 1, pp. 1–11, 2019.

[20] T. Vendl, P. Šípek, O. Kouklík, and L. Kratochvíl, "Hidden complexity in the ontogeny of sexual size dimorphism in male-larger beetles," *Scientific reports*, vol. 8, no. 1, pp. 1–10, 2018.

[21] D. Ahrens, J. Schwarzer, and A. P. Vogler, "The evolution of scarab beetles tracks the sequential rise of angiosperms and mammals," *Proceedings of the Royal Society B: Biological Sciences*, vol. 281, no. 1791, p. 20141470, 2014.

[22] Y. Peng, X. He, and J. Zhao, "Object-part attention model for fine-grained image classification," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1487–1500, 2017.

[23] P. Shroff, T. Chen, Y. Wei, and Z. Wang, "Focus longer to see better: Recursively refined attention for fine-grained image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 868–869.

[24] D. A. Van Dyk and X.-L. Meng, "The art of data augmentation," *Journal of Computational and Graphical Statistics*, vol. 10, no. 1, pp. 1–50, 2001.

[25] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[26] N. Japkowicz, "Concept-learning in the presence of between-class and within-class imbalances," in *Conference of the Canadian society for computational studies of intelligence*. Springer, 2001, pp. 67–77.

[27] Q. Xu, M. Zhang, Z. Gu, and G. Pan, "Overfitting remedy by sparsifying regularization on fully-connected layers of cnns," *Neurocomputing*, vol. 328, pp. 69–74, 2019.

[28] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," *Advances in data science and information engineering*, pp. 877–894, 2021.

[29] G. Csurka, "A comprehensive survey on domain adaptation for visual applications," *Domain adaptation in computer vision applications*, pp. 1–35, 2017.

[30] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[31] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[32] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *International Conference on Machine Learning*. PMLR, 2013, pp. 10–18.

[33] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5400–5409.

[34] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[35] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, "Metareg: Towards domain generalization using meta-regularization," *Advances in neural information processing systems*, vol. 31, 2018.

[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.