Image-to-Image Translation of Synthetic Samples for Rare Classes

Edoardo Maria Lanzini[1] , **Robert-Jan Bruintjes**[1] , **Attila Lengyel**[1] ,
Jan van Gemert[1] , **Sara Beery**[2]
[1]TU Delft , [2]Caltech

Abstract

The natural world is long-tailed: rare classes are observed orders of magnitudes less frequently than common ones, leading to highly-imbalanced data where rare classes can have only handfuls of examples. Learning from few examples is a known challenge for deep learning based classification algorithms, and is the focus of the field of low-shot learning. One potential approach to increase the training data for these rare classes is to augment the limited real data with synthetic samples. This has been shown to help, but the domain shift between real and synthetic hinders the approaches' efficacy when tested on real data.

We explore the use of image-to-image translation methods to close the domain gap between synthetic and real imagery for animal species classification in data collected from camera traps: motion-activated static cameras used to monitor wildlife. We use low-level feature alignment between source and target domains to make synthetic data for a rare species generated using a graphics engine more "realistic". Compared against a system augmented with unaligned synthetic data, our experiments show a considerable decrease in classification error rates on a rare species.

1 Introduction

Accurately and scalably monitoring biodiversity is vital to our understanding of the changing world around us. Policymakers need near-real-time monitoring data to analyze the efficacy of conservation actions in the face of human encroachment and climate change. Camera traps and other static passive monitoring sensors provide vital monitoring data to ecologists, but as the size of these networks of sensors increase, the magnitude of data outpaces human processing capacity. Ecologists are increasingly turning to computer vision and machine learning approaches to help automate the detection and categorization of animal species, necessary in order to scale this critical assessment.

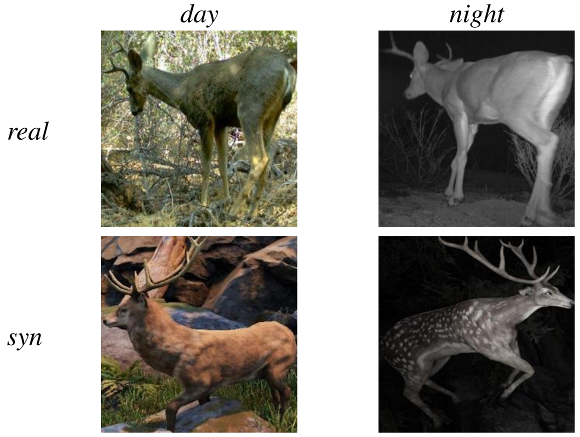

Figure 1: **Examples of real and synthetic images of deer.** The visual difference between the two domains is noticeable for both *day* and *night* examples.

Camera trap data introduces challenges beyond those addressed in traditional computer vision benchmark datasets like ImageNet [13]. These include long-tailed distributions [49] and a multitude of different sub-domains (locations) within the same dataset [8]. In particular, the classification of rare species of animals is notoriously troublesome due to the combined effect of scarcity in number of examples and the low sample efficiency of data from a given camera deployment.

To limit the bias toward well-represented classes, both algorithmic [15, 21, 22] and data solutions [10] have been proposed. Beery *et al.* explored the addition of synthetic samples for a single rare class and showed to improve classification accuracy [4]. However, despite the impressive capabilities of graphical engines, the synthetic samples are still perceived by the network as semantically distant compared to the real ones [4].

Beery *et al.* crafted a dataset starting from the Caltech Camera Traps (CCT) Dataset [8], artificially undersampling the deer class and training the classification model with synthetic renderings [4]. The same synthetic data is used as a starting point in this work to investigate the impact of synthetic-to-real image-to-image translation on the classification of the single rare class.

In this work we quantify the domain shift between synthetic and real camera trap data using color distribution, texture, and feature distance. We narrow the gap with unpaired image-to-image translation methods operating in a low-data regime with only a handful of real samples from the target domain. We show this results in higher efficacy when using synthetic data to augment limited real examples for a rare species, ultimately leading to an increase in classification performance for both seen and unseen locations.

2 Related Work

2.1 Domain Adaptation from Synthetic to Real

Domain adaptation techniques often operate in the feature space, seeking to close the distribution gap between samples from different domains [37]. Supervised and unsupervised techniques are used to align the features of the source (synthetic) and the target (real) [11, 24, 29, 36]. The gap is commonly bridged by either mapping the two domains to a domain-invariant representations [16, 17] or forcing the two learned distributions to be close [19, 44, 45]. Various metrics have been proposed to measure this domain gap, including maximum mean discrepancy [32], correlation distance [46], or adversarial discriminator accuracy [16, 48]. Hoffman *et al.* introduced Cycle-Consistent Adversarial Domain Adaptation (CyCADA), operating at both pixel and feature-level, showing significant improvements over previous methods [24].

2.2 Image-to-Image Translation

As an alternative to feature-level domain adaptation, image-to-image (I2I) translation attempts to directly increase the "realism" of synthetic data at the pixel-level. Paired I2I [25] maps an image from source to target domain using an adversarial loss [18], combined with a reconstruction loss between the result and target. In the unpaired setting, the samples from the two domains are not paired, and correspondence is enforced using cycle-consistency [26, 51, 52], learning the mapping in both directions and computing a loss on the reconstruction of the original input.

Early adaptations of Generative Adversarial Networks (GANs) [18] showed promising results in simple settings, with small images and minimal semantic difference between domains [11, 30, 43].

CycleGAN [52] uses a cycle consistency loss in its adversarial approach, training two different generators to translate in opposite directions, introducing a reconstructions loss. The architecture introduced by Isola *et al.* is often extended with context-specific loss terms that allows to enforce further constraints on the translation learned [42, 50].

UNIT [29], used in this work, is an I2I framework based on Coupled GANs [30]. Compared to CycleGAN, the network does not learn a direct mapping between the two domains but instead operates under the assumption of a common latent space, in which both domains can be mapped. This assumption also implies a cycle-consistency constraint between the two domains [29]. The adversarial setting of both UNIT and CycleGAN makes training complex.

Recent work by Park *et al.*, tackles the unpaired I2I problem using contrastive learning, operating at the level of patches, enforcing the constraint that corresponding patches in the two domains should have high mutual information [35]. This intuition is formulated using a multilayer, patchwise contrastive loss that allows to learn a one-sided translation.

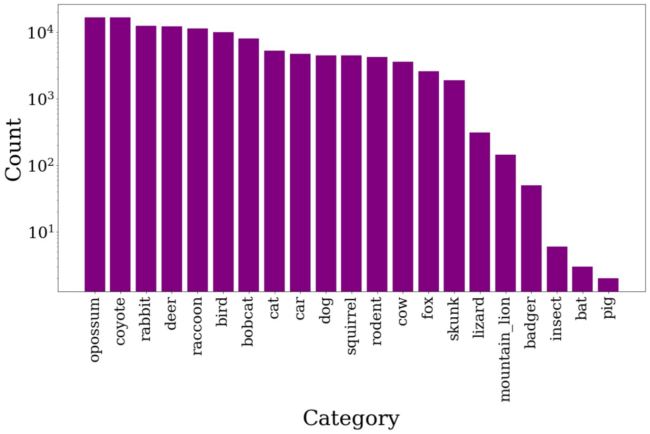

Figure 2: **Distribution of categories in CCT.** The number of samples across the different categories is long-tailed. The deer class is far from uncommon in CCT but it is artificially isolated as rare in the CCT-20 split. Note that the y axis is in log scale.

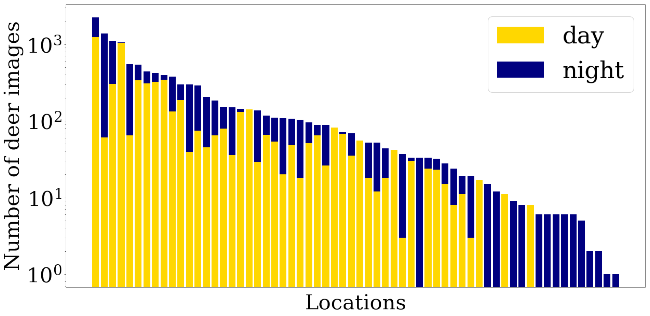

Figure 3: **Distribution of deer across CCT locations.** The number of deer seen at different camera locations is long-tailed and often combined with an uneven split between *day* and *night* within the same location. Note that the y axis is in log scale.

The application of I2I to translate from synthetic to real has improved performance in other real-world applications [1, 2, 14, 31, 43].

2.3 CV for Camera Trap Data

Camera traps are increasingly used by biologists to unobtrusively monitor wildlife. The use of deep learning to increase data processing speeds has been widely investigated in recent years [3, 5–7, 9, 33, 34, 39, 40, 47]. The static nature of camera traps, combined with the long-tailed distribution of species in the real world, leads to poor generalization performance in novel deployments and for rare species [8, 27, 39]. Recent works tackle these challenges directly, focusing on categorization of rare species or generalization to novel camera deployments. Beyond data augmentation approaches like the one explored in this work, architectures [38] and loss functions [12, 28] designed for long-tailed distributions have also shown promise.

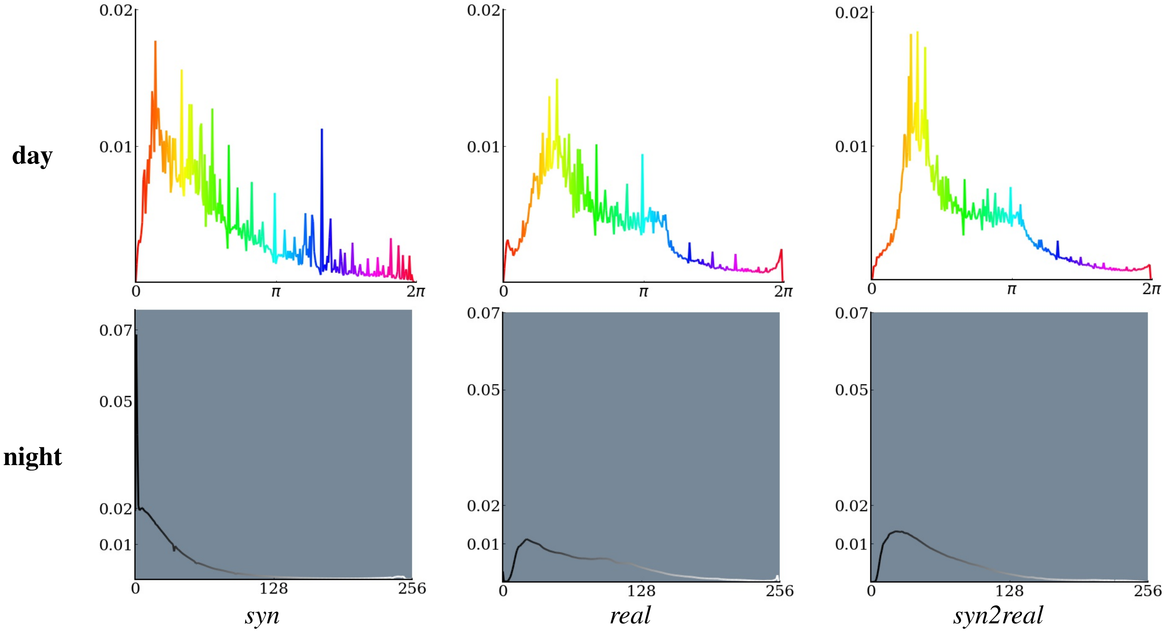

Figure 4: **Comparison between the *syn*, *real* and *syn2real* color distribution for day (top) and night (bottom).** The color distribution for the day is computed aggregating the discretized hue channel. For the night, samples are first converted to grayscale and then pixel values are aggregated. The resulting distribution of both is normalized. The *syn2real* distributions move closer to the *real* ones compared to the *syn*. The high intensity of *syn* night is due to the high saturation of the renderings.

3 Data

3.1 CCT

The Caltech Camera Traps (CCT) dataset contains 243,187 images from 140 camera trap locations covering 30 classes of animals, curated from data provided by the United States Geological Survey and the National Park Service [8]. CCT is used as a testbed for long-tailed distributions under real-world conditions, where the number of samples for each species is unbalanced (see Fig. 2). The distribution of samples per-sensor is also long-tailed, with an additional uneven split between day and night occurrences (see Fig. 3).

3.2 CCT-20

We use the same data split as [4], starting with the CCT-20 subset introduced in [8] and isolating deer as the single rare class of interest. The *real* training set is composed of 13,553 images from 9 camera locations, containing only 44 deer examples, and is used as the source of real samples for our I2I translation task. Our additional *synthetic* training data is also the same as [4], which is generated with Unity's 3D game development engine. To constrain the task of translation, we make use of the bounding box annotations for both real and synthetic data to build two sets of images that share similar framing (see Figure 1).

4 Experiments

First, we use the entire collection of deer samples from CCT (denoted **CCT-deer**) to evaluate the different I2I translation models. The data is split between day (2342 samples) and night (3132 samples), and models are trained separately to translate bounding box crops resized to 256x256 pixels from the synthetic to the real domain.

CCT-deer		
Correlation	day	night
correlation(syn, real)	0.73	0.36
correlation(syn2real, real)	**0.96**	**0.96**
correlation(syn2real, syn)	0.81	0.46

CCT-20		
Correlation	day	night
correlation(syn, real)	0.73	0.36
correlation(syn2real, real)	**0.94**	**0.95**
correlation(syn2real, syn)	0.70	0.29

Table 1: **Average color distribution correlations.** Measured between (i) the *syn* and *real* images, (ii) the *syn2real* and *real* images and (iii) the *syn2real* and *syn* images for both day and night. The model trained on CCT-20 (bottom) is performing similarly to the model trained on CCT-deer (top).

To bridge the domain gap, three different unpaired I2I translation methods are compared. Using the official implementations of UNIT[1], CycleGAN[2] and CUT[3], we trained each model with the default hyperparameters. CycleGAN starts from a generative adversarial setting and adds a cy-

[1] https://github.com/mingyuliutw/UNIT

[2] https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

[3] https://github.com/taesungp/contrastive-unpaired-translation

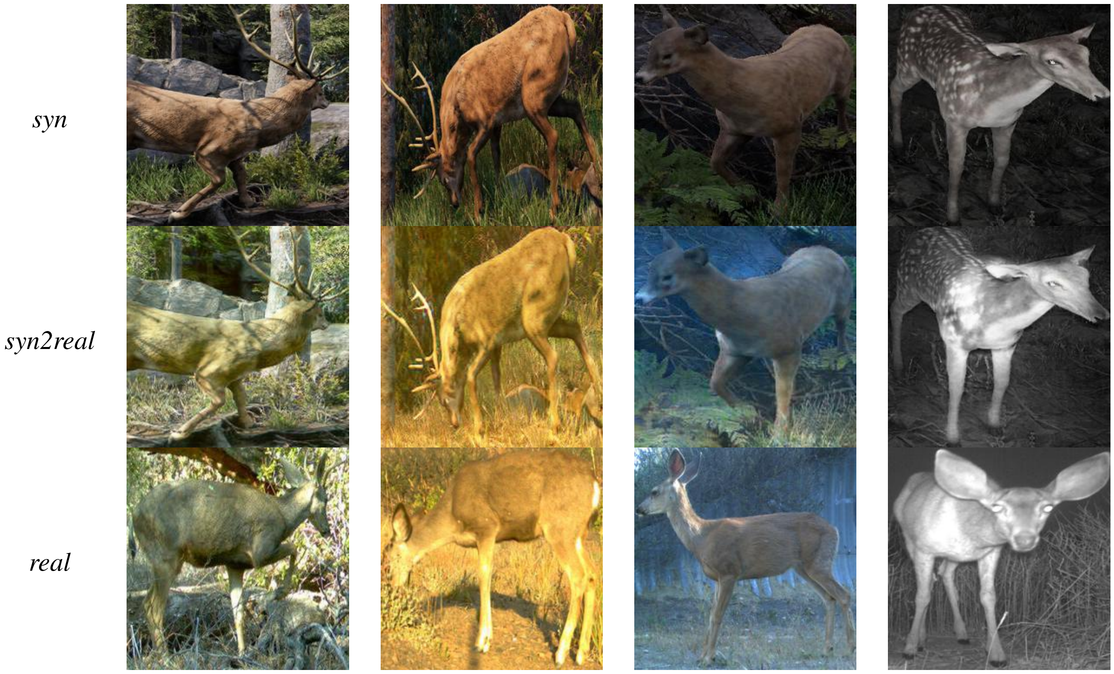

Figure 5: **Examples generated by UNIT trained on CCT-deer.** These hand-picked examples show deer in similar poses starting with the *syn* and comparing the two outputs of the models (*syn2real*) with the *real* sample. The translation learns to match the color distribution of the real imagery, while the texture appears unchanged.

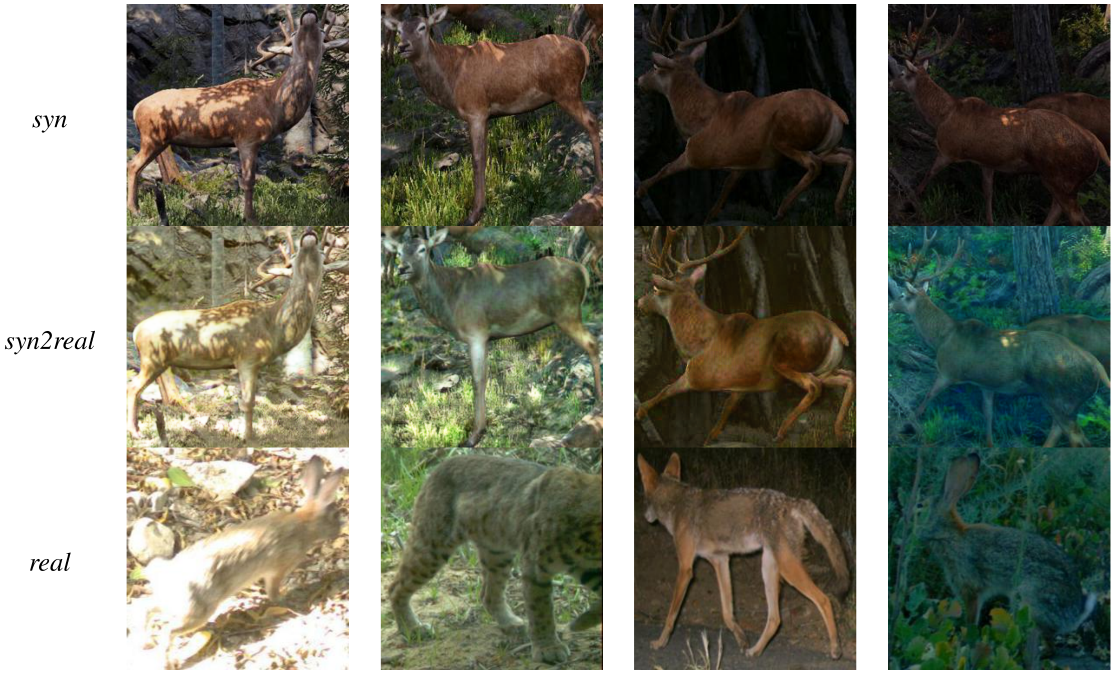

Figure 6: **Examples generated by UNIT trained on the entire real training set of CCT-20.** When trained with all the categories as a target, the model learns to imitate the chromatic distribution of different locations seen during training.

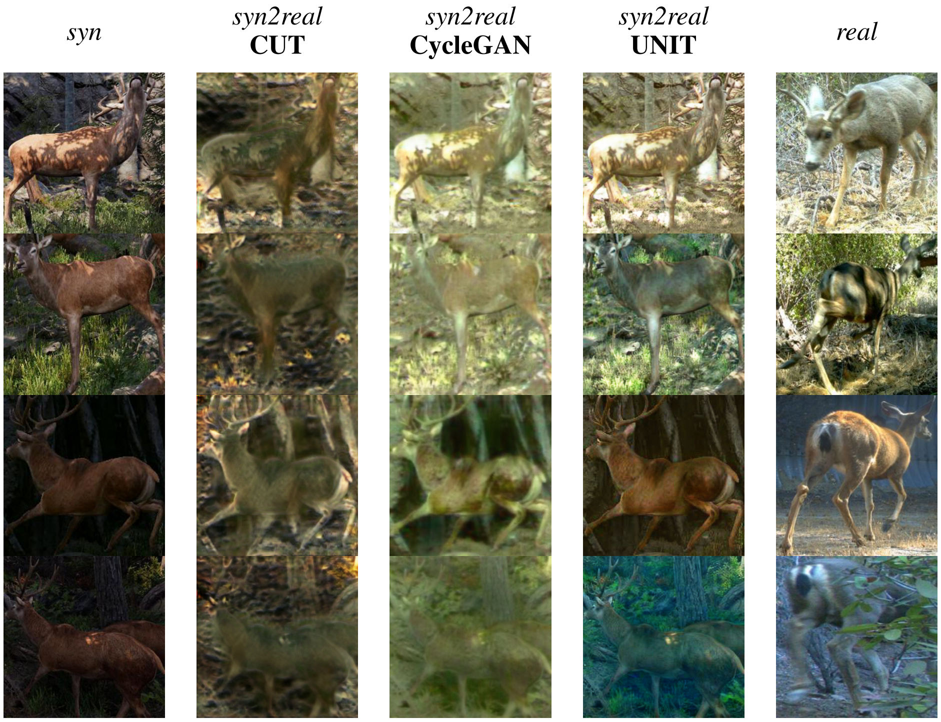

Figure 7: **Comparison of the different unpaired I2I models trained on the same data.** These hand-picked examples show different outputs coming from the translation learned by the different models. UNIT shows better results, with more variance in the color distribution and sharpness in the refinement compared to CycleGAN and CUT.

cle consistency loss, to constraint the learned mapping [52]. Similarly, UNIT uses a shared-latent space constraint, separating the generator into an encoder and decoder component, enforcing cycle-consistency between those [29]. CUT uses contrastive learning to encourage patches from the two domains to share mutual information [35]. With the same amount of training, UNIT produces qualitatively superior results (see Figure 7). From a qualitative inspection, CycleGAN and CUT models appear to show less variance in the learned translation and sometimes introduce artifacts in their outputs. Because of this, we chose to use UNIT for the remainder of our experiments. The UNIT model trained on CCT-deer appears to visually imitate the locations seen during training from the real samples, altering the look of the synthetic image to mimic the real locations (see Fig. 5). Qualitatively, the model seems to learn to alter the colors of the image but the texture appears to be untouched. To measure this effect quantitatively, we analyze the color distribution and texture of the real data as well as the synthetic data pre- and post-translation.

4.1 Color space

The most notable change in the translated samples is the shift in the color distribution of the synthetic samples, that appear to resemble the color scheme of the real samples. We consider day and night separately, as samples from the two are visually and statistically distinct.

Day

To evaluate the color difference for all samples obtained during the *day*, we look at the sample-normalized distribution of the Hue value from the HSI colorspace, representing the pure color at each pixel regardless of saturation and illumination. To measure the distance between the *real*, *syn* and *syn2real* distribution, we computed the Pearson correlation coefficient between each of them. The *syn2real* correlation improves from 0.73 to 0.96 with *real* samples and decreases from 1.0 to 0.81 with *syn* samples (see Table 1).

Night

The *night* samples are first converted to grayscale and their color features are captured by the sample-normalized distribution of pixel values. The *syn2real* correlation improves from 0.36 to 0.96 with *real* samples and decreases from 1.0 to 0.46 with *syn* samples (see Table 1).

These measurements suggest that the model is able to approximate the distribution of the real samples for both *day* and *night*. An important observation is that the model imitates the color distribution of the **locations** that is trained on, uniformly altering the color of all the pixels in the image.

4.2 Texture space

Another dimension through which we measure distance is texture space, quantifying the translation impact on the synthetic samples, compared to the real ones. To characterize textures, we use gray level co-occurrence matrix (GLCM)

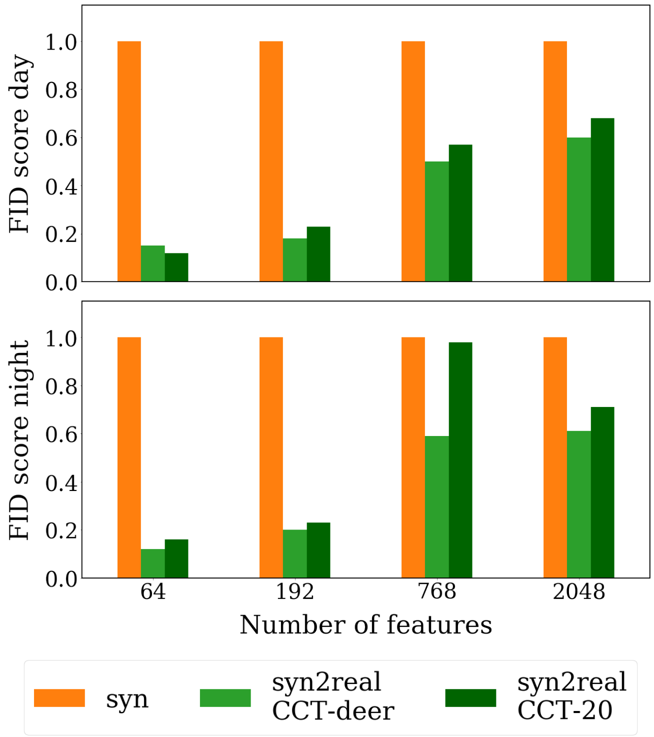

Figure 8: **Normalized FID score computed for both *day* and *night* at different depths in feature space.** The score quantifies the distance from *real* for both *syn* and *syn2real*. We express the *syn2real* score as a fraction of the *syn* at each architecture depth. We see that the gap is getting narrower at lower-level features.

features [20]. In particular, we extract *contrast*, *homogeneity*, *energy* and *entropy*.

The goal is to measure the difference between the texture of the fur of the animal across the two domains. To isolate this sub-experiment, a model is trained to map synthetic samples to a single location (location ID 34 in CCT) for which we have an abundant sample size during the day, allowing us to normalize the context in which textures are measured. This is due to the fact that texture changes across locations, due to lightning conditions and other factors. By translating to a single location, we can also manually pick real samples that were captured at a camera location with a similar look and confine the effect of the translation on the textures. Similar to [2], we manually crop 4 20x20 patches for 10 manually selected real and synthetic samples, compute and average GLCM features across the set.

The GLCM texture features measured on the *syn2real* samples show a negligible improvement, compared to *syn* samples. As confirmed by a qualitative inspection, the model is not considerably shifting the distribution of texture space. The positive delta introduced by the translation is small compared to the impact on the color distribution (see Section 4.1).

4.3 Exploring translation for a rare class

We have shown that it is possible to narrow the difference in visual appearance between *syn* and *real* camera trap data using CCT-deer as training set. Performing the same translation task for the deer class on the CCT-20 dataset becomes problematic due to the limited amount of target data (44 deer samples), but this represents a more realistic scenario for any rare species. That said, the previous experiment suggests that the mapping learned from the *syn* to the *real* data alters mostly the lower-level color features, with the textures being slightly changed. In other words, the model learns the appearance of the different **locations** presented during training. This suggests that the model could also learn a similar chromatic transformation using real images that do not necessarily correspond to the deer class, extending the target set from the 44 deer images to the 13,553 CCT-20 training images across all categories.

Using the entire CCT-20 training set as our target, the model replicates the chromatic distribution learned from the locations seen during training. As shown in Figure 6, those correspond to locations populated by categories outside of the *deer* class. Using the same procedure described in Section 4.1 to measure color distributions, we find a correlation of 0.94 (day) and 0.95 (night) (see Table 1) with the real imagery.

4.4 Feature space

To further evaluate the quality of the two translation models, we use Fréchet Inception Distance (FID) to quantitatively capture domain similarity [23]. To capture the semantic distance at different architecture depths, activations of 64, 192, 768 and 2048 are extracted from a pretrained Inception classifier [41].

Figure 8 shows the computed FID between source and target for the respective feature dimensions. For both *day* and *night*, the translation method appears to close the gap most significantly early in the network, with the largest decrease at the first max-pooling layer (64 features), encapsulating lower-level features. The CCT-20 model performs similarly to the CCT-deer model, suggesting that the features corresponding to realness can be learned and transferred from a target set containing multiple categories, bypassing the need for large amounts of real data of our rare class.

5 Classification

The ultimate goal of our method is to improve classification of the rare class of interest by making our synthetic data more “real”. To test this, we finetune an Inception V3 model, pretrained on ImageNet, to classify species in bounding box crops from CCT-20. We use the same training parameters - learning rate, optimizer and input transformations - as [4]. We compare classification results when training with (1) only the *real* data, (2) augmenting it with 10K *syn* samples (5K day, 5K night), and (3) augmenting with the same 10K synthetic samples post-translation (*syn2real*), using the model from 4.3.

Cis

The cis test set is made up of held out images from camera locations seen during training. The error rate on cis test set decreases by 16% from *real* to *syn* and improves by 37% from *real* to *syn2real*. The model trained on *syn2real* images

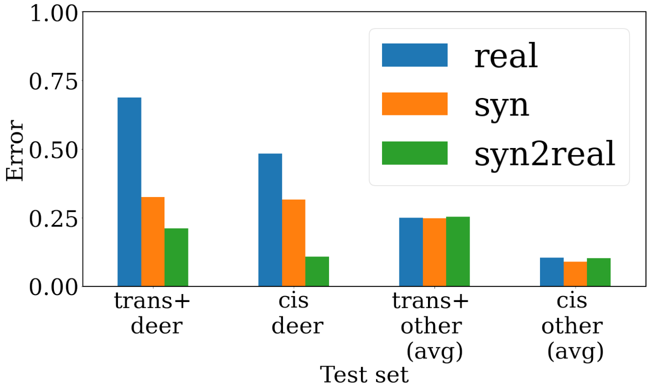

Figure 9: **Error rates measured on the classification of trans+ and cis test sets.** On both sets, the error rate for the ”rare” deer class is significantly decreasing when the model in trained *syn2real* data compared to just *syn* samples. The change in the classification error of the other classes is negligible.

improves the classification of the deer class on the cis test set by 21%. The considerable improvement in the cis test set may stem from the ability of the I2I model trained on CCT-20 to mimic the low-level statistics of each camera location.

Trans
The trans test set is composed of samples from camera locations not seen during training. This initial set is augmented with all the deer samples present in CCT (**trans+**). In the classification of deer for the trans+ test set, we see a 36% decrease in error rate from *real* to *syn* and a 48% decrease from *real* to *syn2real*. The 12% improvement in the trans+ test set may stem from the training that is performed on translated samples that resemble the style of classes different from deer. This might help the model to generalize the classification to unseen locations for the rare deer class.

In both testing scenarios, we see a negligible change in the average error rate of the other classes ($\pm < 1\%$) (see Fig. 9).

6 Responsible Research

To ensure the reproducibility of the experiments showed in this work, we outlined the methodologies to both create the different datasets used and hyperparameters set to train the different models.

6.1 Data

As detailed in Section 3, the CCT-20 split of CCT is created using the annotations provided in [8] and applying the same modifications, removing the *fox*, *car*, *car*, *badger* and *empty* classes, to isolate *deer* as the only rare class. The synthetic data used is randomly sampled from the 1.4M renderings provided in [4]. For both domains, bounding boxes annotations are used to crop samples from the starting images. The resizing is done using cubic interpolation.

The synthetic data is generated using human crafted models of deer. To make sure enough variability in the models is present, all 18 different available models are used to generate the 3D scenes and later cross them with a virtual camera. Nonetheless, humans' representation of deer is certainly biased towards a common representation and cannot reproduce the full variability that is found in nature.

6.2 Models

To train the different I2I models, the default hyperparameters, proposed in the official implementations are used to evaluate the quality of the learned translations. The three different models mentioned in Section 4 are all trained using the same datasets of real and synthetic data for *day* and *night*. To reproduce the same conditions for the classification as in [4], the same hyperparameters, data augmentation and stopping criteria are used.

7 Conclusion and Future Work

The domain shift present in the low-level features between real and synthetic images can be effectively narrowed by simply imitating the color distribution of the locations in the target samples. Our experiments show this I2I translation can be learned using the entire training set of real samples, including samples from other categories. This is particularly beneficial when dealing with real-world long-tailed distributions, where rare classes are underrepresented. It remains to be tested how different I2I models deal with a multitude of domains (locations), investigating the distribution of the locations that the model is able to reproduce, compared to the training data.

The improvements on classification from the enhancement in “realness” of the synthetic data is encouraging and could beneficially impact the wildlife monitoring of rare endangered species.

8 Acknowledgements

We would like to thank the USGS and NPS for providing data and Microsoft AI for Earth for providing compute resources.

References

[1] Charith Atapattu and Banafsheh Rekabdar. Improving the realism of synthetic images through a combination of adversarial and perceptual losses. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2019.

[2] Ruud Barth, Jochen Hemming, and Eldert J Van Henten. Optimising realism of synthetic images using cycle generative adversarial networks for improved part segmentation. *Computers and Electronics in Agriculture*, 173:105378, 2020.

[3] Sara Beery, Elijah Cole, and Arvi Gjoka. The iwildcam 2020 competition dataset. *arXiv preprint arXiv:2004.10340*, 2020.

[4] 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 (WACV)*, March 2020.

[5] Sara Beery, Dan Morris, and Pietro Perona. The iwildcam 2019 challenge dataset. *ArXiv*, abs/1907.07617, 2019.

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

[7] Sara Beery, Grant van Horn, Oisin MacAodha, and Pietro Perona. The iwildcam 2018 challenge dataset. *arXiv preprint arXiv:1904.05986*, 2019.

[8] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[9] Sara Beery, Guanhang Wu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Context r-cnn: Long term temporal context for per-camera object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13075–13085, 2020.

[10] Elizabeth Bondi, Debadeepta Dey, Ashish Kapoor, Jim Piavis, Shital Shah, Fei Fang, Bistra Dilkina, Robert Hannaford, Arvind Iyer, Lucas Joppa, et al. Airsim-w: A simulation environment for wildlife conservation with uavs. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 1–12, 2018.

[11] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017.

[12] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019.

[13] 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.

[14] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3d human pose estimation: motion to the rescue. *Advances in Neural Information Processing Systems*, 32:12949–12961, 2019.

[15] Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

[16] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[17] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[19] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[20] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973.

[21] H He, Y Bai, EA Garcia, and S ADASYN Li. adaptive synthetic sampling approach for imbalanced learning. ieee international joint conference on neural networks. In *2008 (IEEE World Congress On Computational Intelligence)*, 2008.

[22] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6629–6640, 2017.

[24] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor

Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018.

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[26] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, pages 1857–1865. PMLR, 2017.

[27] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, et al. Wilds: A benchmark of in-the-wild distribution shifts. *arXiv preprint arXiv:2012.07421*, 2020.

[28] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[29] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 700–708, 2017.

[30] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *NIPS*, 2016.

[31] Rui Liu, Chengxi Yang, Wenxiu Sun, Xiaogang Wang, and Hongsheng Li. Stereogan: Bridging synthetic-to-real domain gap by joint optimization of domain translation and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12757–12766, 2020.

[32] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

[33] Mohammad Sadegh Norouzzadeh, Dan Morris, Sara Beery, Neel Joshi, Nebojsa Jojic, and Jeff Clune. A deep active learning system for species identification and counting in camera trap images. *Methods in Ecology and Evolution*, 12(1):150–161, 2021.

[34] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25):E5716–E5725, 2018.

[35] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020.

[36] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

[37] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.

[38] Dvir Samuel, Yuval Atzmon, and Gal Chechik. From generalized zero-shot learning to long-tail with class descriptors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 286–295, January 2021.

[39] 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.

[40] Stefan Schneider, Graham W Taylor, Stefan Linquist, and Stefan C Kremer. Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods in Ecology and Evolution*, 10(4):461–470, 2019.

[41] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.1.1.

[42] Matan Sela, Pingmei Xu, Junfeng He, Vidhya Navalpakkam, and Dmitry Lagun. Gazegan-unpaired adversarial image generation for gaze estimation. *arXiv preprint arXiv:1711.09767*, 2017.

[43] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017.

[44] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[45] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

[46] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

[47] Michael A Tabak, Mohammad S Norouzzadeh, David W Wolfson, Erica J Newton, Raoul K Boughton, Jacob S Ivan, Eric A Odell, Eric S Newkirk, Reesa Y Conrey, Jennifer L Stenglein, et al. Improving the accessibility and transferability of machine learning algorithms for identification of animals in camera trap images: Mlwic2. *bioRxiv*, 2020.

[48] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[49] Grant Van Horn and Pietro Perona. The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450*, 2017.

[50] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018.

[51] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017.

[52] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.