IMAGE DATA AUGMENTATION WITH UNPAIRED IMAGE-TO-IMAGE CAMERA MODEL TRANSLATION

Chi Fa Foo[⋆], *Stefan Winkler*[⋆†]

[⋆] School of Computing, National University of Singapore
[†] ASUS Intelligent Cloud Services (AICS), Singapore

ABSTRACT

Many image datasets are built from web searches, with images taken by various cameras. The variance of camera sources can lead to different camera signals and colors within images of the same class, which may impede neural networks from fitting the data. To generalize neural networks to different camera sources, we propose an augmentation method using unpaired image-to-image translation to transfer training images into another camera model domain. Our approach utilizes CycleGAN to create a translation mapping between two different camera models. We show that such a mapping can be applied to any image as a form of data augmentation and is able to outperform traditional color-based transformations. Additionally, this approach can be further enhanced with geometric transformations.

Index Terms— CycleGAN, Camera style transfer, Image classification

1. INTRODUCTION

Deep neural networks achieve state-of-the-art performance in many computer vision tasks such as image classification and object detection. Unfortunately, neural networks require large amounts of diverse data in order to train a useful model that can generalize to unseen images [1]. Image data augmentation is a way to possibly achieve better generalization and prevent overfitting on a given dataset. Most common approaches employ an array of geometric and color-based transformations for this purpose [2, 3]. Additionally, with the growth of generative adversarial networks (GAN), synthetic image generation has become more widely used [1].

Many common image classification datasets construct their database through web searches from various sources [4, 5]. Such approaches introduce differences in camera signals and color between images of the same class, which may affect the ability of a neural network to fit to the data [6]. Variance in source cameras will result in different images despite similar scenes as a consequence of varied sensor types, demosaicing algorithms, and other proprietary processing done on the images [7, 8]. Such variance may cause the neural network to equate certain camera model signals to a specific class, overfitting the color noise. This is particularly problematic if the dataset is small, which will have minimal variety of camera models in each class. There are some recent studies done on creating camera-invariant data augmentation [6, 9], however, they are not extendable to a general image classification problem.

Our proposed approach focuses on using unpaired image-to-image translation with CycleGAN [10] to learn the feature space of two different camera models and construct the translation from one to another. This emulates the idea of taking a picture of the same scene using different cameras. The translation mapping will be used in a general case, as our approach will translate the source camera model of any original image in a direction towards another camera model. This can be seen as a GAN-based method of color and signal transformation on the source image. We show that our image augmentation approach performs significantly better than traditional color-based transformations in image classification problems.

2. RELATED WORK

Image data augmentation is a useful tool to increase the accuracy of image classification tasks [1]. It reduces the distance between training and validation sets by representing a more complete set of possible data points [1]. Recently, more studies have attempted to create an automatic learning strategy for image augmentation [2, 3]. These methods employ a search space among probability and magnitudes of common geometric and color-based transformations to augment the dataset.

In addition to geometric and color-based transformation methods, some have explored synthetic image generation through GANs [1]. This is especially interesting in fields where images are scarce, such as medical imaging [11]. However, creating a GAN that generates high-resolution images from noise requires a significant amount of training data, which leads to some research focusing on using image-to-image translation to create reliable high-resolution image data [12]. Image-to-image translation is the problem of translating the domain of an image from one to another [13]. CycleGAN has been a widely used translation mechanism by various researchers for medical imaging and emotional classification problems [1]. This is partly because CycleGAN

is able to translate domains from unpaired images, as paired images are difficult or sometimes impossible to acquire [14].

Other research focused on the camera invariant aspects of a neural network. [6] used CycleGAN to implicitly learn the difference in camera views on person re-identification. CycleGAN was able to reliably translate images from one source camera to another, which improved the results significantly. [9] modeled different camera effects such as chromatic aberration, blur, exposure, and noise. Unlike [6] which learned camera views from a set of training data, these camera effects were made into a search space of parameters to augment synthetic data. The research showed promising results for object detection in urban driving. Following the usage of synthetic camera images, [15] explored ray-tracing for the simulation of a multispectral camera pipeline to generalize neural networks for driving. Their research showed that synthetic images were able to emulate the camera invariance obtained from different cameras.

These efforts show promise in the idea of introducing camera invariance in neural networks. Furthermore, they highlight the possibility of using synthetic images or translation to build such invariance. However, they require a specific problem or dataset, and are difficult to extend to a general classification problem.

3. CAMERA MODEL TRANSLATION

3.1. CycleGAN

Our method employs CycleGAN [10] to synthetically map images from one camera model domain to another. The translated images are then used to augment the original dataset.

CycleGAN employs two sets of GANs, creating mappings between two image domains X and Y: X→Y and Y→X. Each set of GANs contains its own generator and discriminator network competing through adversarial training. An additional cycle consistency loss is added to the objective, which is built on the idea of transitivity between two sets of images. Cycle consistency is defined as the ability to translate back to the original domain after the first translation [10]. Therefore, if an image is translated from domain X to domain Y, it should be possible to translate it back to its original image in domain X.

Cycle consistency is an important tool to ensure that the primary features of the original image will not become distorted as a result of the translation. This is critical as we want to preserve the objects from the original image to emulate the same scene taken from a different camera.

3.2. Pre-processing Training Data

In order to train the GAN, we need to reduce the image dimensions of training images. Normally, resizing and random cropping are used for pre-processing images, however, [8] reported that the accuracy of identifying camera models through image features and color filter array drops if the images were compressed. [16] found cropping images to be a reliable method to reduce image dimension for camera model identification with neural networks. Drawing inspiration from pre-processing methods from camera model identification problems, we choose to use only cropping on the images to reduce their dimension to 256x256.

This method of pre-processing brings two primary benefits: reliable construction of color-based translation, and increased training data size. The first comes from the preservation of the unique color-based signals from the CCD, color filter array, and demosaicing algorithm of a source camera [8]. The second benefit is important because of the scarce amount of data for each specific camera model.

3.3. Source Camera Datasets

Our CycleGAN training dataset is created by combining three different source camera identification datasets: VISION [17], Forchheim [18] and Kaggle Camera Model Identification Competition [19]. By compiling images from these datasets, we were able to collate a total of 763 iPhone5c images and 521 Huawei P9Lite images. The image dimension of the classes are 3264x2448 and 4160x3120 respectively. These two camera models were selected as they are from a different phone brands and camera makers, whilst having the largest amount of data. After cropping, each image is converted to 130 and 221 cropped images for iPhone5c and P9Lite respectively. The total amount of cropped images are 99,190 for iPhone5c and 115,141 for P9Lite.

3.4. Training and Augmentation

Upon pre-processing, we map the images to two domains based on their source camera model. These images are sourced from various databases and contain random scenes with little relation to one another. This is to enable the CycleGAN to automatically learn the source camera features of both models and construct a mapping from one model to the other. Our CycleGAN is trained with a 3-layer discriminator and ResNet generator, with a dimension of 16 for both the first layer of the discriminator and last layer of the generator. We utilize these parameters in line with the findings that source camera features are easily identified through lower-level layers [20]. We trained for 60 epochs with a learning rate of 0.002, with the learning rate decaying from the last 30 epochs, on 30,000 images for each camera model.

The trained CycleGAN is employed as a data augmentation tool in the next step. For each image to be augmented, we apply the CycleGAN translation for both directions (X→Y, Y→X). The translated images are combined with the original images to create the augmented dataset, effectively tripling the dataset size.

4. IMAGE CLASSIFICATION EXPERIMENTS

4.1. Datasets

To verify the effectiveness of our approach in image classification tasks, we utilize the CIFAR-100 [5] and Oxford-IIIT Pet [21] datasets.

CIFAR-100 contains 100 classes with 500 training images and 100 test images each with a resolution of 32x32 pixels.

Oxford-IIIT Pet dataset contains 37 species of cats and dogs with roughly 200 images each, with a median resolution of 500x375. This is a considerably harder dataset as compared to CIFAR-100, because of the limited amount of data and the similarity between its classes. Some species of cats and dogs are difficult to differentiate correctly, as they share many similar features. This makes it harder to fit a neural network as the data points are generally closer together.

We designate 10% of each class as test images, and remove them from our data augmentation and training process.

4.2. Reference Network and Methods

We use the ResNet50 model from Torchvision [22] for both datasets. The model is optimized on SGD with a learning rate of 0.001 and momentum of 0.9, and all images are resized to 224x224 with batch size of 32. Each of the test runs are performed until the model converges and has no additional accuracy gains for the last 10 epochs.

The baseline augmentation method for both datasets follows the convention for most image classification problems, which includes random cropping, image normalization, and random horizontal flips with 50% chance [2].

Additionally, we compare our approach to a color-based transformation method. The color transformation is performed using Torchvision's colorjitter method. Colorjitter contains various color transformations, including saturation, brightness, contrast, and hue. Following [23], we utilize the findings of [2] on CIFAR-100 sub-policies for color transformations as the parameters for colorjitter for both datasets. The colorjitter is applied twice at every epoch and concatenated with the baseline dataset to be in-line with our approach, with an addition of 2 times the total number of training images.

Furthermore, as noted by [23], color transformations are known to perform best when paired with geometric transformations for CIFAR-100. Therefore, we perform an additional geometric transformation for our test on the CIFAR-100 dataset. We combine our approach with the random application of a geometric transformation from the set of: translation, rotation, and scaling. The transformation is applied randomly with a 30% chance on each of our training images. We compare our results with RandAugment [3], which contains a mixture of geometric and color-based transformation. To keep the RandAugment approach similar to our approach, we maintain the number of transformed images at 2.

4.3. Image Data Augmentation

In order to fit the images of CIFAR-100 to our CycleGAN, we have to increase the dimensions to 256x256. Despite the possibility of distorting the original source camera signals with the upscaling of image size, we show that our CycleGAN is still able to perform reasonably well in both the look and performance of the images as seen in Figure 1. Transformations A→B and B→A are consistent for both examples.

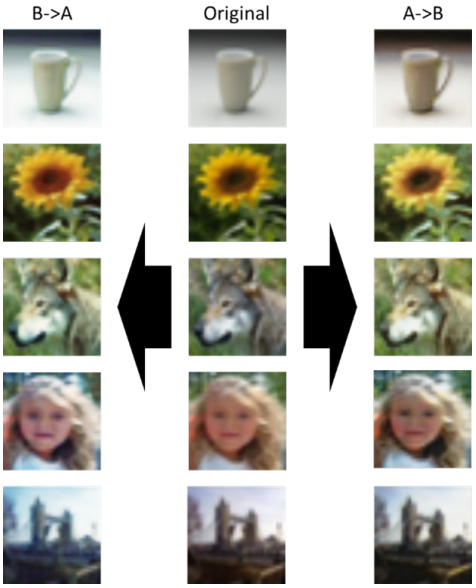

Fig. 1. Sample translation of CIFAR-100 images using our CycleGAN. Domains A and B correspond to iPhone5c and P9Lite respectively.

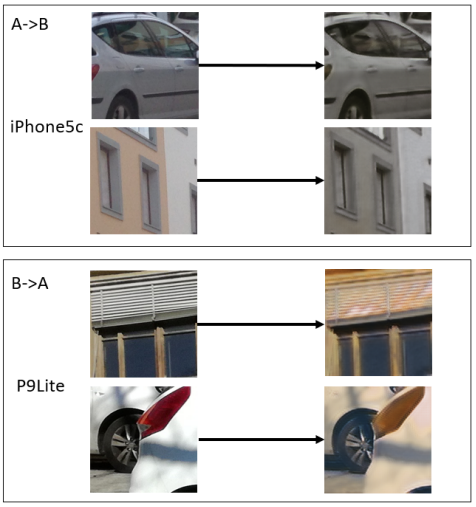

Fig. 2. Translation using an image from the original domain.

For the translation on Oxford-IIIT Pet, we use a patched approach, wherein we crop each image to patches of 256x256. As the image dimensions of Oxford-IIIT Pet are varied, images with dimension smaller than 256x256 are scaled up before being separated into patches. These patches are translated using our CycleGAN, and then formed back together into the dimensions of the original image.

Because of the larger image dimensions of the Oxford-IIIT Pet dataset, color transformation may end up creating too much color noise in the images. Comparatively, our method has less color noise, as seen in Figure 2.

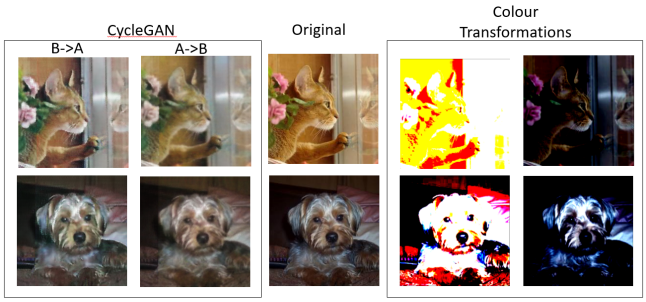

Fig. 3. Images from the Oxford-IIIT Pet dataset modified using our approach (left) vs. the traditional color-based approach (right), which in this case is a selection from saturation, contrast, and brightness transformation.

4.4. Results

Table 1 shows the results of our experiments with our approach against the baseline, colorjitter, and RandAugment methods. For CIFAR-100, we tested the baseline in two ways, imitating our approach of scaling from 32x32 to 256x256 down to 224x224 and directly scaling from 32x32 to 224x224. We found no large discrepancies between the results of the two scaling methods.

As seen from the results, we are able to achieve performance than regular color transformations by colorjitter. In addition, we are able to further improve the results of our approach by applying geometric transformations.

5. DISCUSSION

Reflecting the discovery by [23], pure color-based transformations will net low gains in accuracy even for CIFAR-100, which has generally shown larger improvements from color-based transformations [2]. In comparison, our CycleGAN augmentation method was able to improve the accuracy of our classification network significantly compared to the colorjitter method.

For Oxford-IIIT Pet, color-based transformations performed even more poorly than in CIFAR-100. This is likely because of the larger image dimensions, resulting in more color noise from traditional color-based transformations as seen in Figure 2. However, our method was still able to provide significant gains in accuracy while minimizing color noise.

Method	Loss	Accuracy (%)
CIFAR-100		
Baseline	2.8107	30.87 / 60.43
Baseline + Colorjitter	2.6674	33.48 / 63.38
Baseline + RandAugment	2.4689	37.50 / 67.83
Baseline + our approach	2.0570	46.81 / 77.05
Baseline + our approach + geometric	2.0016	56.83 / 83.57
Oxford-IIIT Pet		
Baseline	3.1516	15.00 / 43.11
Baseline + Colorjitter	3.0786	15.95 / 45.41
Baseline + our approach	2.7649	24.05 / 56.90

Table 1. Validation loss and accuracy (Top 1/ Top 5) on baseline, colorjitter, RandAugment, our approach alone, and combined with geometric transformations on CIFAR-100 and Oxford-IIIT Pet.

It is likely that our transformation method was able to perform better than traditional color-based method because of the use of deep learning. We utilized CycleGAN to learn from images with different source camera model and color signals. As such we were able to learn translation in the camera model space which imitates actual scenes taken by different cameras. Such an approach shows promise in helping the neural network fit datasets with different source cameras [6, 15].

6. CONCLUSION AND FUTURE WORKS

We experimented with a novel image augmentation approach using translation between camera model domains. By applying image-to-image translation on the camera model space, we were able to transform the training images. Our experiments showed that this approach performs better than traditional color-based augmentation methods. The images created from our approach can be further augmented using geometric transformation to achieve better results.

Further improvements to our current approach could be envisaged by using multi-domain translations. Such methods will allow us to translate each images to multiple different camera models. Additionally, the model will be able to better learn the latent camera model distribution and likely translate the camera specific signals better. This will allow our method to become more scalable and provide a larger search space to robustly train a camera-invariant neural network.

7. REFERENCES

[1] 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.

[2] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *arXiv preprint arXiv:1805.09501*, 2018.

[3] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 702–703.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[5] A. Krizhevsky, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009.

[6] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camstyle: A novel data augmentation method for person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1176–1190, 2018.

[7] A. S. Orozco, D. A. González, J. R. Corripio, L. G. Villalba, and J. Hernandez-Castro, "Techniques for source camera identification," in *Proc. 6th International Conference on Information Technology*, 2013, pp. 1–9.

[8] T. Van Lanh, K.-S. Chong, S. Emmanuel, and M. S. Kankanhalli, "A survey on digital camera image forensic methods," in *Proc. International Conference on Multimedia and Expo (ICME)*. IEEE, 2007, pp. 16–19.

[9] A. Carlson, K. A. Skinner, R. Vasudevan, and M. Johnson-Roberson, "Modeling camera effects to improve visual learning from synthetic data," in *Proc. European Conference on Computer Vision (ECCV) Workshops*, 2018.

[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2223–2232.

[11] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[12] A. Gupta, S. Venkatesh, S. Chopra, and C. Ledig, "Generative image translation for data augmentation of bone lesion pathology," in *Proc. International Conference on Medical Imaging with Deep Learning*. PMLR, 2019, pp. 225–235.

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1125–1134.

[14] R. Zhang, T. Pfister, and J. Li, "Harmonic unpaired image-to-image translation," *arXiv preprint arXiv:1902.09727*, 2019.

[15] Z. Liu, T. Lian, J. Farrell, and B. A. Wandell, "Neural network generalization: The impact of camera parameters," *IEEE Access*, vol. 8, pp. 10 443–10 454, 2020.

[16] A. Kuzin, A. Fattakhov, I. Kibardin, V. I. Iglovikov, and R. Dautov, "Camera model identification using convolutional neural networks," in *Proc. International Conference on Big Data*. IEEE, 2018, pp. 3107–3110.

[17] D. Shullani, M. Fontani, M. Iuliani, O. Al Shaya, and A. Piva, "Vision: A video and image dataset for source identification," *EURASIP Journal on Information Security*, vol. 2017, no. 1, pp. 1–16, 2017.

[18] B. Hadwiger and C. Riess, "The Forchheim image database for camera identification in the wild," in *Proc. International Conference on Pattern Recognition (ICPR)*. Springer, 2021, pp. 500–515.

[19] IEEE Signal Processing Society, "Camera model identification," https://www.kaggle.com/c/sp-society-camera-model-identification/data.

[20] D. Freire-Obregón, F. Narducci, S. Barra, and M. Castrillón-Santana, "Deep learning for source camera identification on mobile devices," *Pattern Recognition Letters*, vol. 126, pp. 86–91, 2019.

[21] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3498–3505.

[22] S. Marcel and Y. Rodriguez, "Torchvision the machine-vision package of torch," in *Proc. 18th ACM International Conference on Multimedia*, 2010, pp. 1485–1488.

[23] S. Mounsaveng, I. Laradji, I. Ben Ayed, D. Vazquez, and M. Pedersoli, "Learning data augmentation with online bilevel optimization for image classification," in *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1691–1700.