https://doi.org/10.2352/issn.2694-118X.2021.LIM-16


# Generative inter-class transformations for imbalanced data weather classification

***Apostolia Tsirikoglou, Marcus Gladh, Daniel Sahlin, Gabriel Eilertsen, Jonas Unger, Linköping University, Sweden.***

## Abstract

*This paper presents an evaluation of how data augmentation and inter-class transformations can be used to synthesize training data in low-data scenarios for single-image weather classification. In such scenarios, augmentations is a critical component, but there is a limit to how much improvements can be gained using classical augmentation strategies. Generative adversarial networks (GAN) have been demonstrated to generate impressive results, and have also been successful as a tool for data augmentation, but mostly for images of limited diversity, such as in medical applications. We investigate the possibilities in using generative augmentations for balancing a small weather classification dataset, where one class has a reduced number of images. We compare intra-class augmentations by means of classical transformations as well as noise-to-image GANs, to inter-class augmentations where images from another class are transformed to the underrepresented class. The results show that it is possible to take advantage of GANs for inter-class augmentations to balance a small dataset for weather classification. This opens up for future work on GAN-based augmentations in scenarios where data is both diverse and scarce.*

## Introduction

Although machine learning (ML), and in particular deep learning, over the last decade have shown great success and potential, it is becoming more and more apparent that one of the most pressing challenges is the data used in the training and evaluation processes. It has been shown that solutions based on deep neural networks, [15], can solve computer visions tasks with high accuracy and performance, outperforming traditional algorithms. However, their performance is limited by the training data used in the learning process. The fundamental problem is that there is a lack of both: the availability of training data with accurate ground truth annotations, as well as robust methods for capture and generation of such data in most application scenarios. Access to unbiased data relevant to the task is recognized as one of the central challenges in ML and promising approaches for data synthesis have recently been proposed [26].

A common difficulty is that the training data is imbalanced, i.e., the number of data points for a specific class is significantly lower compared to other classes due to, e.g., difficulties in collecting or annotating the data for a specific class or a scenario.

This paper focuses on class-imbalanced weather classification, in which we train a supervised classifier to determine which weather condition an image represents, {*Sunny, Foggy, Rainy, Snowy*}, in a low data availability scenario. The goal is to investigate how different augmentation and data synthesis strategies can be applied to improve classifier performance when one of the classes is represented by a significantly lower number of training samples compared to the others, see Figure 1. Although synthesis using direct computer graphics simulation is possible [30, 16], the complexity in the scenes makes it an intractable approach as it would require highly sophisticated modelling and rendering

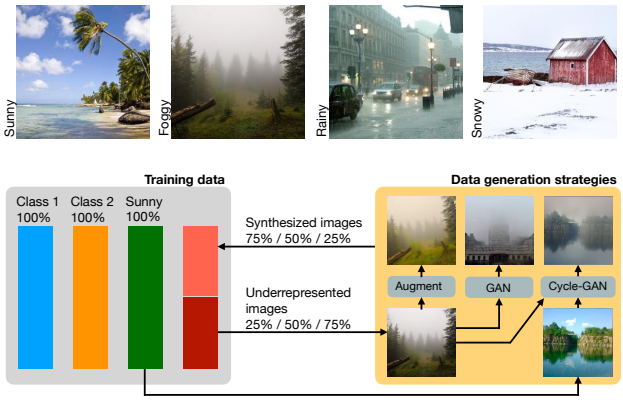


Figure 1: shows **(top)** example images illustrating the four different classes {*Sunny, Foggy, Rainy, Snowy*}, and **(bottom)** the augmentation strategies considered for class-balancing.

techniques and skills; only accessible to the very high-end visual effects, computer games and image production companies. Inspired by recent work in style transfer for training data, [20], we instead turn to generative models, in particular generative adversarial networks (GANs) [7, 11, 34], and evaluate their performance as a tool for data synthesis. Generating new training images using a noise-to-image GAN has been demonstrated useful in medical applications [1]. However, for natural images, where the image diversity is significantly more complex, it is difficult to generate new images that are of sufficient quality for training, especially when data is scarce. Instead, we consider inter-class transfer by means of Cycle-GANs [34], mapping, e.g., sunny images to the underrepresented foggy class. This approach has two distinct advantages over a noise-to-image GAN: 1) it is a simpler task to modify an already existing image instead of generating a completely new image, and 2) diversity can be increased by considering images from other classes, i.e., where image generation is not restricted to using the images of one particular class.

The contribution of this work is a systematic evaluation of Cycle-GANs as a tool for inter-class transfer for data synthesis in low data scenarios. Inter-class augmentations are compared to intra-class augmentations using noise-to-image GANs, as well as conventional augmentation operations. We also introduce a modification of Cycle-GAN, replacing instance normalization with the weight demodulation proposed in Style-GAN-2 [14], which can alleviate problems with artifacts also in Cycle-GAN. We evaluate the different augmentation strategies in a range of scenarios, with different inter-class transformations and with different availability of training data.

## Related work

Our approach builds upon a large body of work. The ongoing data for ML challenge has spurred research and development of techniques for synthetic data generation based on direct simulation using computer graphics techniques [22, 21, 30, 16], as well as using generative models, typically GANs, for direct data

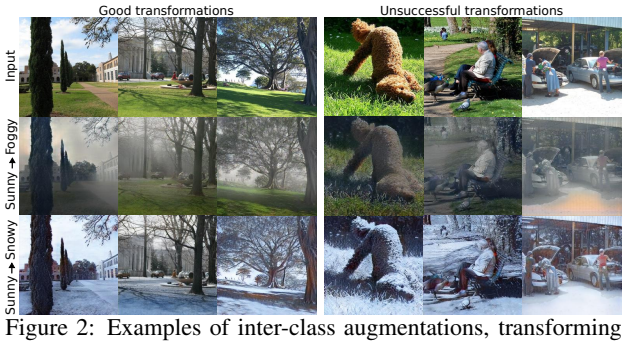

Figure 2: Examples of inter-class augmentations, transforming from the sunny class (top) to foggy (middle) and snowy (bottom). The examples on the left side have been selected as more successful transformations, while the right images demonstrate failure cases.

generation [4] and data domain transfer [20] for different applications. For an in-depth overview of synthetic data for ML see the survey by Tsirikoglou et al. [26].

GANs are trained in an adversarial mini-max game, where a data-generating neural network takes noise as input and produces, e.g., images as its output, while a discriminator network is trained with the objective to separate between generated and real images. The generator's objective is to produce images that the discriminator will mistake as real. From the introduction of GANs [7], intensive research has resulted in models that can generate photo-realistic images for sufficiently narrow data distributions [13, 2, 19, 32]. Adversarial training has also been extended to image-to-image problems, learning to transform images between different domains, both in supervised [10] and unsupervised settings [34].

Single-image weather classification is a challenging computer vision task, and there are only a few methods not based on deep learning, which extract features and apply conventional ML techniques [23, 18, 33]. However, neural networks have been demonstrated to improve state-of-the-art by a large margin [5], and there is a number of recent methods that have tested different variations of CNNs [36, 17, 8]. Methods are usually formulated to distinguish between 2-5 different weather phenomena, such as *sunny*, *cloudy*, *rainy*, *foggy* and *snowy*, but slight variations in problem formulation and the need for data, in general, has resulted in that most of the different works have also constructed new datasets, e.g., by gathering and labeling images from online sources [18, 36, 17, 8]. Given the different needs, and the still limited research for single-image weather classification, there is still a lack of large-scale weather datasets of high quality. Thus, it is crucial to make the most of the data at hand.

GANs for augmentation of training data has mostly been considered in medical imaging [1, 6, 9, 24, 31, 27, 28]. Medical images are typically represented by a narrower data distribution as compared to natural images, which makes it possible to apply adversarial image generation using a low number of images. In applications for natural images, there are only a few attempts at GAN-based augmentations [35, 29], and learning-based augmentations have instead been more focused on creating optimal transformations of already existing images [3, 25].

## Methodology

This papers focuses on the problem of augmenting a small-scale imbalanced single-image weather classification dataset by means of adversarial image generation, and in particular an evaluation of different augmentation/transformation methods. The downstream task considered is to train a classifier that can determine the weather condition from different classes under the constraint that one of the classes is underrepresented in the training dataset. This is a very challenging task, as natural images present a large degree of diversity, which means that it is problematic to train a noise-to-image GAN to produce new image samples (see Figure 5). Instead, we opt to transform already existing images from classes that are more well-represented, using a Cycle-GAN, which is both a simpler problem and can also increase the diversity of an underrepresented class. The difference between classical augmentations, intra-class GAN-augmentations and inter-class Cycle-GAN augmentations is illustrated in Figure 1, pointing to how inter-class augmentations can feed information from a source class to the target class.

**Dataset -** In this study we use the publicly available weather classification dataset published in [8]. We define two sets of experiments of intra- and inter-class augmentations: 1) for the four classes of $\{$*Sunny, Foggy, Rainy, Snowy*$\}$, where class *Cloudy* is removed due to possible overlap with *Rainy*, and *Foggy*, and 2) for the three classes $\{$*Sunny, Foggy, Snowy*$\}$ where $\{$*Rainy*$\}$ is additionally removed due to possible overlap with $\{$*Foggy*$\}$. The classes $\{$*Sunny, Foggy, Rainy, Snowy*$\}$ each consists of 1100 labeled images. The 4400 images are first divided 80:20 into class balanced training and test sets, such that the training set contains 880 images and the test set 220 images from each class. The training sets were further split to also provide a validation set using a 90:10 ratio and the images were cropped to a $256\times256$ tile around the center. Examples of the represented classes are illustrated in Figure 1. For the evaluation experiments, we randomly remove images for one of the classes, so that the underrepresented class in the imbalanced dataset contains 25%, 50%, or 75% of the original images. This is done for the classes $\{$*Foggy, Snowy*$\}$ respectively. To balance the dataset for training the weather classifier, the underrepresented class is filled in with images synthesized using different GAN architectures. The GANs are trained to transfer images from the source domain $\{$Sunny$\}$ to one of the target domains $\{$*Foggy, Snowy*$\}$, and the resulting synthesized images are inserted into the training set. We also include a reference scenario where we copy images directly from the underrepresented class either as they are, i.e., duplicates, or transformed through geometric (zooming, rotation, shifting, shearing, and flipping), and pixel transformations (brightness, noise, and color).

**GAN data synthesis -** We compare three different GANs: Progressively Growing GAN [11] (denoted **PG-GAN**), Cycle-GAN [34] (denoted **C-GAN**), and Cycle-GAN extended with weight demodulation inspired by Karras et al. [14] (denoted **C-GAN-WD**). The PG-GAN (mapping random numbers to image) and C-GAN (mapping image to image) are the vanilla architectures as described in [11, 34] respectively. For the C-GAN-WD we extend C-GAN by removing instance normalization, and replacing the convolution operation in the generator block with a custom one that includes weight demodulation. This modification is suggested to potentially improve problematic areas in images generated, identified as noise or the droplet effect [14].

For the C-GAN and C-GAN-WD generated images we investigate two selection protocols. The first one is a random selection, while the second one is based on the discriminator score (D-score). In this case, the complementary images, needed to balance the underrepresented class, are selected in a descending D-score order. This aims to ensure that the most convincing as real images to the network are included in the augmentation set.

**Weather condition classifier -** The classifier is, for efficiency, built around a simple architecture consisting of three con-

Figure 3: Inter-class augmentations with varying number of images in the target domain.

Figure 4: Differences between vanilla C-GAN and C-GAN-WD, for transforming between sunny and foggy images.

volutional layers and two fully connected layers. The classifiers are trained with Adam optimizer for 50 epochs, out of which the best model is selected. To avoid additional complexity in the analysis we do not perform on-line augmentations.

## Evaluation and results

We first discuss the characteristics of the GAN generated images, followed by an evaluation on the classification performance with varying data scenarios and augmentation strategies.

**Image synthesis results -** Figure 2 shows some examples of images generated using inter-class transforming C-GANs. While there are many examples where the transformations work well, there are also cases with less successful transformation. For the snowy case, the C-GAN often resorts to producing images without colors. The foggy transformation has a higher success rate, but fails for close-up shots where fog is a less pronounced phenomenon. Figure 3 shows the differences when training with different amount of images in the target domain. The foggy transformation can be learned with very small amount of training data, while the snowy transformation is more likely to show failure cases in the low-data scenarios. The differences between the vanilla C-GAN and C-GAN-WD are most often subtle, but the weight demodulation can help in reducing the amount of droplet or color smearing artifacts, as demonstrated in Figure 4. The noise-to-image GAN, in our case using PG-GAN, is very difficult to train on the low amount of diverse images presented by our scenarios. In most cases, it resorts to trying to reproduce the training images, as demonstrate in Figure 5, and there are severe problems with mode collapse. However, even if PG-GAN mostly produces distorted versions of the training images, these could still be beneficial for augmentations, so it is of interest to compare this approach to the inter-class transformations.

**Evaluation -** Table 1 shows the results from experiments where the underrepresented class has been extended with duplicates, augmented duplicates, and images generated using the GAN methods described above. Each classifier was trained 20 times and the table reports the mean accuracy and its standard deviation for the individual classes and the overall mean accuracy computed over all three or four classes; a total of 1960 trained models. We show the results for the $\{Foggy, Snowy\}$ classes where the first is easier for the GANs to learn.

A first observation is that the accuracy for the fully represented classes increases in many cases when the underrepresented class is performing poorly, see for example the imbalanced trainings (first row in each section) or when only duplicates are used. One can also see that this pushes up the overall, total mean accuracy so that it in some cases meets and even slightly exceeds training with the full dataset. This can be explained by that the variation, or feature richness, is low in the underrepresented class and that this leads to a stronger emphasis on the full classes during training. Studying the mean, total accuracy it is evident that all transformation and augmentation methods improve the overall performance significantly for all scenarios. Starting from a large relative improvement compared to the imbalanced case at 25% data the relative performance increase is gradually decreasing as more data is added to 50% and 75% data, approaching the performance of the full dataset (top row).

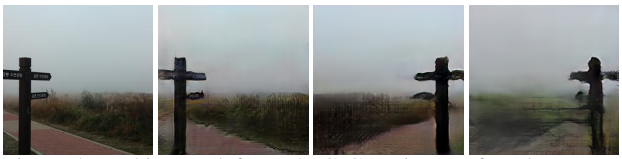
Figure 5: Real image (left), and PG-GAN images found to reproduce this when trained on 25%, 50% and 75% foggy data, respectively. Mirror augmentations have been applied during training, which explains why the replicated images can be mirrored.

The impact of the different GAN methods is most evident looking at individual classes and in particular the underrepresented class. For both the $\{Foggy, Snowy\}$ classes, the GAN based inter-class transformation methods consistently produce better results compared to duplicates and augmented duplicates. The best performing C-GAN reaches 85% and 83% accuracy already in both 25% cases in the 4 classes experiment and remains consistent also in the scenarios with 50% and 75%, and overall also in the 3 classes experiment. The performance of the inter-class GAN methods is overall comparable, but consistently better as compared to the intra-class augmentations. For scenarios with more data, e.g., 75% the duplicates with classical augmentation performs almost as good. This is expected due to the overall increase in available training data and the augmentation strategy used. PG-GAN is overall the worst performing GAN in both the 4 and 3 classes experiments. The reason is likely that it is trained to map random numbers to images in *the PG-GAN training set only*. This leads to a smaller feature variation as compared to the C-GANs which transform images from the source domain $\{Sunny\}$ to the target domains and thus introduces a larger variation in the training data. Although the C-GAN-WD can reduce some of the artifacts produced by the vanilla C-GAN, as illustrated in Figure 4, it does not improve the classifier performance significantly. This is most likely due to the fact that the GAN images produced contain artifacts that are more severe than the noise and droplet artifacts salvaged by the weight demodulation. Similarly, the selection of GAN produced images based on the D-Score does not improve the performance in comparison to random selection for the experiments conducted. Further investigations are necessary in both directions.

## Conclusion and future work

This paper presented an investigation of data augmentation and inter-class transfer for generating class-balanced training datasets for image-based weather classification. In addition to classical augmentation, the study included three different GAN approaches, PG-GAN [11], C-GAN [34], and C-GAN-WD in which the C-GAN has been extended with the weight demodulation described in [14]. The evaluation showed that the classi-

| | Experiment 1: 4 classes — Mean Accuracy±stddev (%) | | | | | Experiment 2: 3 classes | | | |
|---|---|---|---|---|---|---|---|---|---|
| Experiment | Foggy | Rainy | Snowy | Sunny | Total | Foggy | Snowy | Sunny | Total |
| **Full training set** | 85.9± 1.2 | 73.8± 2.9 | 85.0± 2.8 | 79.2±3.6 | 81.0±1.2 | 91.4±1.5 | 90.7±1.5 | 87.6±1.7 | 89.9±0.7 |
| **Foggy 25%** Imbalanced | 0.0±0.0 | 82.3± 2.9 | 89.3± 1.7 | 83.1± 2.5 | 63.7±1.0 | 53.2±36.0 | 94.2±1.7 | 90.3±2.4 | 79.2±11.6 |
| Duplicates | 75.9±3.1 | 76.6±3.0 | 86.5±2.1 | 80.5±2.7 | 79.9±1.1 | 85.0±2.6 | 92.1±1.8 | 90.0±1.8 | 89.1±1.0 |
| Dupl.+Augmentation | 78.0±3.9 | 69.7±3.3 | 83.7±2.3 | 79.0±3.4 | 77.6±1.2 | 86.1±3.3 | 88.5±1.7 | 86.9±3.3 | 87.1±1.4 |
| C-GAN (rand.) | 85.0±2.2 | 71.8±3.2 | 84.7±3.1 | 80.4±2.8 | 80.5±0.9 | **90.1±1.6** | 90.1±2.6 | 88.1±2.6 | 89.4±0.7 |
| C-GAN (D-scrore) | **85.6±2.4** | 71.4±2.8 | 85.2±2.0 | 81.5±1.8 | **80.9±0.8** | 90.0±1.6 | 89.3±2.4 | 88.7±2.5 | 89.3±0.8 |
| C-GAN-WD (rand.) | 80.1±2.3 | 72.8±3.7 | 87.8±1.7 | 82.5±2.1 | 80.8±0.8 | 87.2±1.6 | 91.9±1.9 | 89.7±1.9 | **89.6±0.7** |
| C-GAN-WD (D-scrore) | 80.7±2.4 | 72.9±3.6 | 87.5±1.8 | 81.7±2.9 | 80.7±1.1 | 86.0±3.2 | 92.4±1.7 | 89.0±1.8 | 89.1±0.9 |
| PG-GAN | 79.2±1.9 | 76.5±2.9 | 86.5±2.0 | 80.9±2.4 | 80.8±1.1 | 86.8±1.9 | 92.3±1.4 | 88.9±2.3 | 89.3±0.9 |
| **Foggy 50%** Imbalanced | 64.3±3.3 | 77.1±4.9 | 88.6±2.5 | 80.6±3.9 | 77.7±7.0 | 86.1±2.9 | 93.2±1.4 | 88.8±1.9 | 89.4±1.2 |
| Duplicates | 82.1±2.2 | 75.8±3.7 | 86.1± 2.4 | 75.9±1.8 | 80.0± 4.3 | 89.6±1.9 | 91.8±1.9 | 87.9±1.6 | 89.8±0.7 |
| Dupl.+Augmentation | 84.7±1.9 | 69.2±3.4 | 82.1±2.8 | 80.7±3.5 | 79.1±0.8 | **91.2±1.6** | 86.7±3.0 | 86.8±2.6 | 88.2±0.8 |
| C-GAN (rand.) | **85.8±1.2** | 71.3±3.7 | 84.8±4.1 | 81.2±3.4 | 80.8±1.2 | 91.0±1.6 | 90.0±2.2 | 88.6±2.1 | 89.9±0.6 |
| C-GAN (D-scrore) | 84.1±1.9 | 70.1±3.0 | 87.5±2.5 | 80.9±2.8 | 80.6±0.8 | 91.1±1.8 | 89.1±2.2 | 87.7±1.9 | 89.3±0.8 |
| C-GAN-WD (rand.) | 84.2±1.6 | 73.5±2.7 | 85.6±2.5 | 82.3±1.4 | **81.4±1.0** | 88.9±1.7 | 90.2±1.8 | 89.1±1.7 | 89.4±0.6 |
| C-GAN-WD (D-scrore) | 84.2±1.8 | 73.3±2.3 | 85.9±2.2 | 80.8±2.3 | 81.1±0.9 | 89.3±1.9 | 90.4±1.7 | 89.3±1.9 | 89.7±0.7 |
| PG-GAN | 82.7±2.5 | 75.8±2.9 | 85.7±2.4 | 80.0±3.7 | 81.0±0.7 | 90.1±1.5 | 91.9±1.4 | 87.9±2.0 | **90.0±0.8** |
| **Foggy 75%** Imbalanced | 82.5±2.1 | 74.4±4.1 | 86.1±1.9 | 81.7±3.1 | 81.2±1.0 | 87.8±1.8 | 91.5±1.9 | 87.7±2.4 | 89.0±1.0 |
| Duplicates | 84.0±2.0 | 74.9±2.6 | 85.3±2.1 | 80.0±3.7 | 81.0±1.0 | 90.2±2.1 | 91.8±1.7 | 87.3±3.0 | 89.8±0.7 |
| Dupl.+Augmentation | 86.1±2.2 | 69.1±3.9 | 85.0±2.3 | 80.9±2.7 | 80.3±1.0 | 90.7±2.3 | 88.0±2.0 | 87.1±2.6 | 88.6±1.0 |
| C-GAN (rand.) | 85.8±1.8 | 72.8±2.9 | 85.3±2.6 | 81.3± 3.2 | **81.3±0.9** | 91.6±1.5 | 91.2±1.6 | 86.9±1.7 | 89.9±0.6 |
| C-GAN (D-scrore) | **86.3±1.7** | 71.4±3.3 | 85.1±2.4 | 81.6±3.0 | 81.1±0.7 | 92.0±1.6 | 90.5±1.6 | 87.7±2.1 | **90.1±0.8** |
| C-GAN-WD (rand.) | 84.8±2.5 | 73.4±3.2 | 84.9±3.2 | 80.4±3.0 | 80.9±0.9 | **91.8±1.5** | 90.1±1.4 | 87.9±2.3 | 90.0±0.9 |
| C-GAN-WD (D-scrore) | 84.5±2.2 | 72.5±3.0 | 86.4±2.6 | 81.5±2.4 | 81.2±0.7 | 91.3±1.5 | 90.5±1.8 | 87.7±2.2 | 89.8±0.8 |
| PG-GAN | 84.2±1.7 | 73.9±3.1 | 85.2±3.1 | 81.0±2.7 | 81.1±1.1 | 90.9±1.8 | 91.0±1.8 | 87.8±1.7 | 89.9±0.8 |
| **Snowy 25%** Imbalanced | 89.3±1.9 | 79.2±3.8 | 0.0±0.0 | 84.9±3.2 | 63.4±1.0 | 97.6±1.9 | 00.0±0.0 | 96.3±1.1 | 64.6±0.5 |
| Duplicates | 85.8±2.0 | 77.3±3.2 | 64.7±3.6 | 85.3±2.2 | 78.3±1.1 | 93.3±1.7 | 77.0±4.3 | 93.3±1.9 | 87.9±1.5 |
| Dupl.+Augmentation | 87.1±1.8 | 73.8±4.0 | 76.5±4.0 | 82.3±1.7 | 79.9±1.2 | 93.3±1.0 | 83.1±1.9 | 89.2±1.4 | 88.5±0.5 |
| C-GAN (rand.) | 85.0±1.7 | 72.1±2.5 | 83.3±2.8 | 82.4±2.2 | **80.7±0.8** | 90.2±1.6 | **89.5±2.2** | 88.7±2.0 | 89.5±0.8 |
| C-GAN (D-scrore) | 84.5±2.1 | 69.7±2.0 | **83.7±2.2** | 81.9±2.9 | 80.0±0.8 | 90.2±1.4 | 88.8±2.6 | 88.3±1.7 | 89.1±0.8 |
| C-GAN-WD (rand.) | 84.6±1.5 | 71.3±3.2 | 72.8±3.6 | 85.4±2.9 | 78.5±1.0 | 89.7±1.6 | 80.9±2.5 | 92.7±1.4 | 87.8±0.9 |
| C-GAN-WD (D-scrore) | 84.4±2.3 | 70.9±3.9 | 70.4±4.3 | 86.0±3.7 | 77.9±1.0 | 90.6±1.6 | 79.4±1.8 | 92.3±1.5 | 87.4±1.0 |
| PG-GAN | 84.4±1.9 | 74.8±4.2 | 80.1±3.2 | 83.1±2.7 | 80.6±1.0 | 90.6±2.8 | 87.3±2.6 | 90.8±1.6 | **89.6±1.2** |
| **Snowy 50%** Imbalanced | 89.3±2.3 | 78.9±2.7 | 15.2±3.1 | 84.2±3.0 | 66.9±7.1 | 96.0±2.9 | 38.0±43.1 | 93.9±2.9 | 76.0±12.7 |
| Duplicates | 85.8±1.4 | 76.3±2.5 | 79.2±3.5 | 83.0±1.8 | 81.1±1.0 | 92.5±1.5 | 86.1±2.8 | 90.1±2.4 | 89.6±0.7 |
| Dupl.+Augmentation | 86.1±1.5 | 73.5±2.9 | 83.1±1.9 | 82.0±2.3 | 81.2±0.8 | 92.4±1.8 | 87.7±2.7 | 88.0±2.7 | 89.3±0.8 |
| C-GAN (rand.) | 84.4±1.8 | 71.5±3.1 | **85.7±2.0** | 81.2±3.0 | 80.7±1.0 | 89.6±1.8 | 91.0±1.6 | 87.2±2.0 | 89.3±0.7 |
| C-GAN (D-scrore) | 84.5±1.6 | 73.3±3.6 | 84.7±3.1 | 80.1±2.7 | 80.6±0.9 | 89.5±2.0 | **91.6±1.5** | 87.0±2.4 | 89.4±1.0 |
| C-GAN-WD (rand.) | 86.9±1.5 | 71.4±4.1 | 82.3±3.0 | 82.4±3.3 | 80.8±1.0 | 90.2±2.3 | 88.4±2.1 | 90.1±2.0 | 89.6±0.7 |
| C-GAN-WD (D-scrore) | 86.6±1.9 | 70.8±3.2 | 83.1±2.3 | 83.8±1.8 | 81.1±0.8 | 90.9±2.0 | 87.7±3.8 | 90.8±2.2 | 89.8±1.0 |
| PG-GAN | 85.4±1.5 | 75.9±3.4 | 84.0±2.6 | 81.5±2.2 | **81.7±0.7** | 91.4±3.4 | 88.8±2.0 | 89.8±1.6 | **90±1.0** |
| **Snowy 75%** Imbalanced | 87.5±1.6 | 75.1±2.7 | 82.4±2.9 | 80.2±3.8 | 81.3±1.0 | 89.8±4.5 | 86.6±4.7 | 88.0±2.6 | 88.1±1.8 |
| Duplicates | 85.9±1.7 | 75.0±3.7 | 82.3±3.1 | 81.1±1.7 | 81.1±1.0 | 91.2±1.3 | 89.1±1.7 | 88.9±1.4 | 89.7±0.7 |
| Dupl.+Augmentation | 85.1±2.1 | 73.0±4.6 | 84.6±2.3 | 81.9±2.0 | 81.2±1.1 | 91.1±1.6 | 90.0±1.7 | 87.5±2.1 | 89.6±0.9 |
| C-GAN (rand.) | 85.0±1.7 | 72.5±3.6 | **85.8±2.6** | 80.9±2.4 | 81.0±1.1 | 90.5±2.0 | 91.0±1.2 | 87.5±1.8 | 89.7±0.7 |
| C-GAN (D-scrore) | 85.8±2.5 | 72.6±3.6 | 85.4±1.9 | 81.8±2.4 | **81.4±0.9** | 90.8±1.6 | 90.4±2.0 | 88.0±2.9 | **89.8±0.8** |
| C-GAN-WD (rand.) | 85.7±1.6 | 72.3±3.0 | 85.0±2.8 | 78.3±3.8 | 80.3±0.8 | 91.8±1.8 | **91.1±1.8** | 85.5±2.3 | 89.5±0.9 |
| C-GAN-WD (D-scrore) | 86.2±1.3 | 73.5±2.8 | 84.9±2.8 | 78.9±3.1 | 80.9±1.0 | 91.7±1.3 | **91.1±1.9** | 85.1±2.4 | 89.3±0.9 |
| PG-GAN | 85.0±1.5 | 75.2±2.8 | 83.1±3.1 | 81.8±2.5 | 81.3±1.0 | 90.5±1.6 | 90.5±1.8 | 88.5±2.1 | **89.8±0.8** |

Table 1: The mean accuracy and standard deviation (%) for 20 runs per experiment. The classifier is trained on **(left)** 4 classes and **(right)** 3 classes. The GAN and augmentation strategies are evaluated with 25%, 50%, and 75% of the {*Foggy, Snowy*} images.

fier trained on data from the GAN methods performed better than classical augmentation in general and especially for low data scenarios. Although the GAN methods mostly performed on-par, the C-GAN consistently performed slightly better compared to the others. Based on the evaluation results, we believe that inter-class transformations using generative models have the potential to be developed into a tool for data synthesis.

What is encouraging is how it is possible to successfully use GAN-based training data synthesis in a low-data scenario on natural images of high diversity, which can be considered very challenging for GANs. The concept of inter-class generation is a promising tool for facilitating this problem, by utilizing already existing information from other classes to augment an underrepresented class. There are, however, several venues for future work. As generated images by visual inspection can be deemed to be of highly varying quality, one interesting problem would be to attempt at measuring this quality and use it for weighting of the loss function in the classification, or for sampling the best generated images. From our results, a direct use of the discriminator score does not seem to provide an adequate notion of quality. There are also recent promising techniques for improving the quality and diversity of GAN-generated images in low-data scenarios [12], and it would be interesting to investigate if this could be incorporated in a Cycle-GAN. Finally, as it is possible to generate images with different strategies, it would be of interest to investigate how these can be combined in the best way possible, i.e., to have synthetic data generated as a combination of inter- and intra-class augmentation strategies.

## Acknowledgements

This project was funded by Knut and Alice Wallenberg Foundation, Wallenberg Autonomous Systems and Software Program, the strategic research environment ELLIIT, and 'AI for Climate Adaptation' through VINNOVA grant 2020-03388.

## References


[1] C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernández, J. Wardlaw, and D. Rueckert. GAN augmentation: Augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863*, 2018.

[2] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019.

[3] 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.

[4] G. Eilertsen, A. Tsirikoglou, C. Lundström, and J. Unger. Ensembles of GANs for synthetic training data generation. In *ICLR 2021 workshop on Synthetic Data Generation: Quality, Privacy, Bias*, 2021.

[5] M. Elhoseiny, S. Huang, and A. Elgammal. Weather classification with deep convolutional neural networks. In *IEEE International Conference on Image Processing*, pages 3349–3353, 2015.

[6] 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*, 321:321–331, 2018.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[8] J. C. V. Guerra, Z. Khanam, S. Ehsan, R. Stolkin, and K. McDonald-Maier. Weather classification: A new multi-class dataset, data augmentation approach and comprehensive evaluations of convolutional neural networks. In *NASA/ESA Conference on Adaptive Hardware and Systems*, pages 305–310, 2018.

[9] C. Han, H. Hayashi, L. Rundo, R. Araki, W. Shimoda, S. Muramatsu, Y. Furukawa, G. Mauri, and H. Nakayama. GAN-based synthetic brain mr image generation. In *2018 IEEE 15th International Symposium on Biomedical Imaging*, pages 734–738, 2018.

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. 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.

[11] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *Proceedings of the International Conference on Learning Representations, (ICLR 2018)*, 2018.

[12] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila. Training generative adversarial networks with limited data. In *Advances in Neural Information Processing Systems*, 2020.

[13] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE CVPR*, pages 4401–4410, 2019.

[14] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE CVPR*, June 2020.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

[16] Z. Li, T.-W. Yu, S. Sang, S. Wang, M. Song, Y. Liu, Y.-Y. Yeh, R. Zhu, N. Gundavarapu, J. Shi, S. Bi, H.-X. Yu, Z. Xu, K. Sunkavalli, M. Hasan, R. Ramamoorthi, and M. Chandraker. Openrooms: An open framework for photorealistic indoor scene datasets. In *Proceedings of the IEEE CVPR*, pages 7190–7199, June 2021.

[17] D. Lin, C. Lu, H. Huang, and J. Jia. RSCM: Region selection and concurrency model for multi-class weather recognition. *IEEE Transactions on Image Processing*, 26(9):4154–4167, 2017.

[18] C. Lu, D. Lin, J. Jia, and C.-K. Tang. Two-class weather classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3718–3725, 2014.

[19] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018.

[20] S. R. Richter, H. A. AlHaija, and V. Koltun. Enhancing photorealism enhancement. *arXiv:2105.04619*, 2021.

[21] S. R. Richter, Z. Hayder, and V. Koltun. Playing for benchmarks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2232–2241, 2017.

[22] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision (ECCV)*, volume 9906 of *LNCS*, pages 102–118. Springer International Publishing, 2016.

[23] M. Roser and F. Moosmann. Classification of weather situations on single color images. In *2008 IEEE Intelligent Vehicles Symposium*, pages 798–803, 2008.

[24] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers. Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. *Scientific reports*, 9(1):1–9, 2019.

[25] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

[26] A. Tsirikoglou, G. Eilertsen, and J. Unger. A Survey of Image Synthesis Methods for Visual Machine Learning. *Computer Graphics Forum*, 2020.

[27] A. Tsirikoglou, K. Stacke, G. Eilertsen, M. Lindvall, and J. Unger. A study of deep learning colon cancer detection in limited data access scenarios. In *ICLR Workshop on AI for Overcoming Global Disparities in Cancer Care (AI4CC)*, 2020.

[28] A. Tsirikoglou, K. Stacke, G. Eilertsen, and J. Unger. Primary tumor and inter-organ augmentations for supervised lymph node colon adenocarcinoma metastasis detection. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Springer International Publishing, 2021.

[29] J. Wang, L. Perez, et al. The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*, 11:1–8, 2017.

[30] M. Wrenninge and J. Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing, 2018.

[31] X. Yi, E. Walia, and P. Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552, 2019.

[32] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 97, pages 7354–7363, 2019.

[33] Z. Zhang, H. Ma, H. Fu, and C. Zhang. Scene-free multi-class weather classification on single images. *Neurocomputing*, 207:365–373, 2016.

[34] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017.

[35] X. Zhu, Y. Liu, J. Li, T. Wan, and Z. Qin. Emotion classification with data augmentation using generative adversarial networks. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 349–360, 2018.

[36] Z. Zhu, L. Zhuo, P. Qu, K. Zhou, and J. Zhang. Extreme weather recognition using convolutional neural networks. In *2016 IEEE International Symposium on Multimedia (ISM)*, pages 621–625, 2016.