A generative approach for Facial Expression Data Augmentation
Advanced Deep Learning Models and Methods

Francesco Azzoni
Politecnico di Milano
francesco.azzoni@mail.polimi.it

Corrado Fasana
Politecnico di Milano
corrado.fasana@mail.polimi.it

Samuele Pasini
Politecnico di Milano
samuele.pasini@mail.polimi.it

Abstract

Image-to-Image translation aims to transfer images from a source domain to a target one while preserving the content representations. It can be applied to a wide range of applications, such as style transfer, season transfer, and photo enhancement. To accomplish this task several architectures have been proposed, including CycleGANs which are composed of a pair of GANs, and their improved version Enhanced CycleGAN (ECycleGAN).

In the context of Deep Learning, one major problem is the need for huge datasets to effectively train a deep model. However, the amount of available images can be limited and dependent on the specific class, leading to unbalanced datasets. To solve this problem, a large amount of different data augmentation techniques have been developed during the last years. Our work will focus on exploiting the advancements in Image-to-Image translation to perform data augmentation of facial expression data, exploiting Enhanced CycleGANs. The effectiveness of the proposed method will be assessed analysing the classification performance on the unbalanced FER2013 facial expression dataset. Further attention will be devoted to preprocessing input data and qualitatively evaluating the generated results.

1. Introduction

Since their introduction [12], Deep Learning (DL) based models have contributed to a remarkable improvement of performances in a large amount of tasks such as Image Classification [19], Object Detection [28] and Instance Segmentation [7]. Despite their advantages, the architectures that are used to produce these models are extremely data hungry, *i.e.,* they require lots of data to be properly trained to avoid overfitting. However, a large amount of labeled data is not always available and thus, the problem of data scarcity becomes a real challenge for DL-based models. For this reason, it is important to design solutions that allow to effectively train DL architectures without the need of collecting large datasets.

During the last years, many different ways of dealing with the problem of data scarcity have been developed. For instance, an intuitive way of solving this issue is to increase the size and diversity of the training set using data augmentation techniques [23]. The basic idea behind these approaches is to exploit the original training dataset to produce new samples that can then be used to construct a bigger training set.

Besides this, in some applications such as medical ones [14] the data scarcity problem can concern not only the whole dataset but also the single classes. This leads to another important issue which is data imbalance, as there are classes that are less represented than others. Also in this case, data augmentation techniques can be used to increase the amount of data for specific classes.

Among the many available data augmentation techniques [23], generative models cover an important role. In particular, Generative Adversarial Networks (GANs) [4] are very representative methods. A GAN is trained to produce data that belongs to a target distribution, by learning a mapping from an input distribution to the target one. This is done by exploiting a generator-discriminator architecture where the generator can help create new data whereas the discriminator ensures that the gap between the newly generated data and the original ones is not too large.

The power of GANs can also be leveraged to perform Image-to-Image translation [9], learning a mapping from images of a source domain to images of another. The advancements in Image-to-Image translation can be exploited to perform data augmentation. It is possible to learn an association between a reference well-represented class and a target poorly represented one, resulting in the ability to generate new target samples starting from reference ones.

CycleGANs [26] belong to this family of approaches and possess the advantage of not requiring paired samples to be trained. In particular, given two domains *A* and *B*, the idea of CycleGAN is to use two generators, each of which is capable of performing Image-to-Image translation from one domain to the other one. As stated before, this kind of model can be used for data balancing and data augmentation. Thus, Zhu et al. [27] exploited this architecture to augment a Facial Expression Dataset named FER2013 [5], where the amount of samples belonging to some classes (*e.g., Disgust*) is extremely lower than the number of images of other classes (*e.g., Happy* and *Neutral*). In this work, the authors use a CycleGAN to map Neutral images to Disgusted ones, and the generator of the disgusted images is used to produce new samples and augment the dataset. In the end, the paper shows that the classification performances on the dataset improved after the introduction of the augmented images. The CycleGAN architecture was extended by Zhang et al. [24] with the introduction of Enhanced CycleGAN (**ECycleGAN**) to improve the effectiveness of domain translation, trying to obtain more realistic images than the ones generated by CycleGANs. The presented work exploits ECycleGANs to augment FER2013 dataset and compare the performance with the method proposed by Zhu et al. [27]. Further attention is devoted to preprocessing input data and qualitatively evaluating the generated results. Thus, the main contributions of this work are:

- Reproduction of the experiment proposed by Zhu et al. [27] and analysis of FER2013 dataset to improve data quality.
- Implementation of ECycleGANs for data augmentation and assessment of its performance on Facial Expression dataset.
- Analysis of different training procedures for GAN training.

2. Related work

2.1. Data Augmentation

Data augmentation is a process of artificially increasing the amount and diversity of data by generating new data samples from existing ones. The increasing interest in data augmentation is concerned with the possibility of reducing the operating costs related to data collection which can be very expensive. There are many different ways of performing data augmentation, ranging from altering the original data to using generative models to create new synthetic samples. Data Augmentation is crucial when the task is performed using a dataset composed of images, as in this work.

According to [23] there are several approaches for image data augmentation such as:

- **Image manipulation**: consists in deriving synthetic samples from original images by applying geometric transformations (*e.g.,* flipping, translation, rotation) or photometric transformations (*e.g.,* noise addition, contrast modification) to increase the diversity of the training set. These methods can only be applied if the transformed images belong to a distribution close to the actual one.
- **Image erasing**: consists in deleting one or more subregions in the image. The main idea is to replace the pixel values of these sub-regions with constant values or random values [2, 25, 13]. In this way, Convolutional Neural Networks can learn to focus not only on the most discriminative part of the image.
- **Deep Generative Models**: consists in learning a way to generate samples belonging to a target distribution starting from data belonging to another distribution [4, 27].

Between these techniques, Deep Generative Models have gained much interest in the last years [15].

Data augmentation is not an easy task and several challenges still need to be faced:

- Assuring high quality of the augmented samples (especially with synthetic data generated by GANs) to ensure that the generated samples are similar enough to the original ones.
- Finding an optimal augmentation strategy for the data: different data have different characteristics, so different data augmentation methods have different benefits.
- Reduce as much as possible the bias towards the original data, used as input for the augmentation procedure. This is still concerned with the fact that the generated data should be realistic but at the same time, they should not be a mere copy of the training set.

2.2. GAN

Generative adversarial network (GAN) is a class of machine learning frameworks designed by Goodfellow et al. [4]. Two neural networks compete with each other. In particular, a discriminator and a generator are trained at the same time, competing in a sort of two players' adversarial game. The Generator receives an input that belongs to a certain data distribution and produces a new sample that should belong to a target distribution. On the other hand, the Discriminator is fed with a sample that may or may not be produced by the generator and outputs the probability of the sample being real (not generated) or fake (generated). The goal of the generator is to fool the discriminator, producing realistic images belonging to the target domain, while the goal of the discriminator is to find a way to distinguish between real and fake samples. Traditionally, the input of the generator is sampled from a known distribution and it is referred to as noise. In case the input of the generator is an image of a specific (source) domain, that should be transformed into an image of a different (target) domain, the addressed task is **Image-to-Image translation** [9]. In particular, the aim is to learn the mapping between an input image and an output image, usually using a training set of aligned image pairs. This could be useful for instance to perform style transfer, season transfer, and photo enhancement. Recently, different works have been developed to solve this task. For instance, Isola et al. [11] proposed Pix2Pix, a conditional adversarial network that learns the mapping from the input images to output images. However, training this network requires paired input-output images. Therefore, Zhu et al. [26] proposed CycleGAN, a new architecture that learns to translate an image from a source domain X to a target domain Y in the absence of paired samples. CycleGAN makes use of two couples of GAN where each couple can transform images from one domain to the other. The goal is to learn a mapping $G : X \rightarrow Y$ such that the distribution of generated images $G(X)$ is indistinguishable from the distribution of images belonging to domain Y. At the same time, this should hold also for the other domain. However, there are countless mappings between the two domains. For this reason, the authors propose the addition of cycle consistency loss to enforce that $F(G(X)) \approx X$ and $G(F(Y)) \approx Y$. This means that given an image x belonging to domain X, that is transformed by G into an image $\tilde{x}$ belonging to domain Y, if $\tilde{x}$ is fed to the other generator F, x should be produced. An additional constraint is added since the authors show that it can provide better quality solutions. This constraint is called identity constraint and it enforces that $F(X) \approx X$ and $G(Y) \approx Y$. This means that the architecture should not modify an image if it already belongs to the target domain. To accomplish this task, the loss function is composed of multiple terms. First of all, as usual, the GANs are trained using an **adversarial loss** [4]. However, the original loss is replaced with a least square loss [16] since it ensures more stable training and higher quality results. Thus, the objective function to train the generator $G : X \rightarrow Y$ and its corresponding discriminator D_Y is:

$$\begin{aligned}\mathcal{L}_{\text{LSGAN}}(G,\ D_Y,\ X,\ Y) = \mathbb{E}_{y\sim p_{\text{data}}(Y)}[(D_Y(y)-1)^2] \\ + \mathbb{E}_{x\sim p_{\text{data}}(X)}[D_Y(G(x))^2]\end{aligned} \tag{1}$$

Another important term is the **cycle consistency loss** used to ensure both forward cycle-consistency, i.e., $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$, and backward cycle-consistency, i.e., $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$. This is done in a pixel-wise manner and is expressed as:

$$\begin{aligned}\mathcal{L}_{pixel-cyc}(G,F) = \mathbb{E}_{x\sim P_{\text{data}}(X)}[\|F(G(x)) - x\|_1] \\ + \mathbb{E}_{y\sim P_{\text{data}}(Y)}[\|G(F(y)) - y\|_1]\end{aligned} \tag{2}$$

Finally, the last term to be considered is the **identity loss** used to ensure the identity constraints for both the X and Y domains and expressed via the following pixel-wise computation:

$$\begin{aligned}\mathcal{L}_{pixel-idt}(G,F) = \mathbb{E}_{x\sim P_{\text{data}}(X)}[\|F(x) - x\|_1] \\ + \mathbb{E}_{y\sim P_{\text{data}}(Y)}[\|G(y) - y\|_1]\end{aligned} \tag{3}$$

Thus, the overall loss function is given by a weighted sum of these three terms:

$$\begin{aligned}\mathcal{L}(G,F,D_X,D_Y) = \mathcal{L}_{\text{LSGAN}}(G,\ D_Y,\ X,\ Y) \\ + \mathcal{L}_{\text{LSGAN}}(F,\ D_X,\ Y,\ X) \\ + \lambda_{cyc}\mathcal{L}_{pixel-cyc}(G,F) \\ + \lambda_{idt}\mathcal{L}_{pixel-idt}(G,F)\end{aligned} \tag{4}$$

Even though CycleGANs can produce quite impressive results, for the task of 2-domain translation, the details about texture and style are often accompanied by unpleasant artifacts as reported in [24]. Thus, to improve the effectiveness of domain translation, and obtain more realistic images, Zhang et al. [24] proposed to modify the architecture, leading to the definition of Enhanced CycleGANs (**ECycleGANs**).

One of the improvements suggested in the paper is to avoid relying only on pixel-wise losses to ensure cycle consistency since this could result in perceptually unsatisfying solutions with overly smooth textures. Thus, a loss function that takes into consideration the perceptual similarity is employed [3]. This perceptual loss function includes a term named feature loss (5) that is the euclidean distance between the high-level abstract feature representation of a cycle reconstructed image $F(G(x)$ and the original image x. The feature representations are extracted using a pre-trained 19 layer VGG network [18]. $\phi_{i,j}$ identifies the feature map obtained by the $j-th$ convolution (after activation) before the $i-th$ max-pooling layer. This term is defined as:

$$\begin{aligned}\mathcal{L}_{feature-cyc}(G,F) = \\ \mathbb{E}_{i,j,x\sim P_{\text{data}}(X)}[\phi_{i,j}(F(G(x))) - \phi_{i,j}(x)] \\ + \mathbb{E}_{i,j,y\sim P_{\text{data}}(Y)}[\phi_{i,j}(G(F(y))) - \phi_{i,j}(y)]\end{aligned} \tag{5}$$

The feature loss and the pixel-wise loss are combined into the perceptual loss (6) that is used as cycle consistency loss, where α and β are the coefficients to balance different loss terms.

$$\mathcal{L}_{perc-cyc}(G,F) = \alpha(\mathcal{L}_{feature-cyc}(G,F) \\ + \beta(\mathcal{L}_{pixel-cyc}(G,F)) \tag{6}$$

To further improve the quality of the images produced by CycleGAN, a major adjustment to the structure of the generator is conducted. More specifically, the original basic residual block used inside the generator architecture is replaced with a Residual Dense Normalization Block (RDNB), which consists of a multi-level residual network, dense connection [10], and instance normalization layers [22]. This modification is performed following the fact that more layers and connections can optimize the performance of neural networks. However, to prevent instability in training a very deep network, a scaling factor named **residual scaling** is used to scale down the residuals. Thus, the perceptual loss encourages the translated images to be more realistic, while the introduction of RDNB blocks allows to generate high-quality images.

2.3. Generative Data Augmentation

The advancements in generative methods lead to the possibility of exploiting these methods to perform data augmentation. Zhu et al. [27] proposed an effective way of adapting an Image-to-Image translation method to perform data augmentation on a facial expression dataset. In particular, given that some classes such as *Disgust* are less represented than others such as *Neutral*, the authors propose to use CycleGANs for generating disgusted images starting from neutral ones. It is shown that adding the generated samples to the original dataset allows to improve the performances. The main contribution of this work consists in following the same idea exploiting ECycleGANs to verify their effectiveness also in terms of image quality.

3. Proposed approach

Before exploiting ECycleGANs to perform data augmentation, an attempt was made to reproduce the results obtained by Zhu et al. [27] to have a model for comparison. However, the obtained results didn't match those published in the paper. By analyzing the original dataset it is evident that the samples used for training could not be a random subset of the whole dataset. For this reason, different ways of sampling the dataset were tested given that the authors do not specify how the sampling is performed.

3.1. Dataset preparation

FER2013 dataset [5] is characterized by high intra-class diversity, high inter-class similarity, and the presence of mislabeled samples and samples belonging to other domains. Thus, the following techniques were employed to try to mitigate these problems and check whether the performances reported in [27] could be replicated:

- **Gaussian likelihood**: the idea is to perform instance selection by removing those instances that lie in low-density regions of the data manifold. This is done by exploiting the method proposed in DeVries et al. [1] which computes the likelihood of each image using a Gaussian model fit on feature embeddings produced by a pre-trained Inceptionv3 classifier [20]. Then, only those samples with a likelihood greater than a certain threshold are kept.
- **Confidence Filtering**: the idea is that first a classifier is trained on the whole dataset. All the samples are then evaluated using the classifier and ranked according to the probability of belonging to their real class. Using as threshold a minimum confidence or a number of samples it is possible to obtain a filtered version of the dataset where most of the ambiguous samples are discarded.

However, even though both of these approaches seemed to correctly remove ambiguous samples, the performance reported in the paper [27] could not be reached.

3.2. ECycleGAN for data augmentation

Once the dataset has been filtered, the next step is the implementation and training of the EcycleGAN model. Starting from the existing implementation of CycleGAN provided by [26] several adjustments are made to match the architecture described in [24]:

- Instead of the original residual block [8] of the CycleGAN generator a deeper and more complex one is employed, called RDNB.
- The perceptual loss is implemented as described in Section 2.2 in place of the original L1-loss and is used both as cycle consistency loss and identity loss.
- The Wasserstein GAN objective with gradient penalty (WGAN-GP) [6] is implemented as an alternative to the LSGAN objective function [16] to stabilize the training. Both have been tested and compared in the experiments section.

Given the fact that the number of samples of some classes is very limited, the discriminator of such classes tends to overfit. Several GAN-specific techniques have been proposed to mitigate this problem:

- **One-sided Label smoothing** [17]: it is concerned with the addition of label noise. More precisely, the discriminator is trained on randomly flipped labels instead of real labels. This label noise is applied only to the samples of the less represented class when fed to the discriminator.

- **Instance Noise** [21]: in this case, the discriminator sees the correct labels, but its input sample is noisy. This avoids the saturation of the discriminator objective, reducing overfitting. The considered noise is a Gaussian one with zero mean and standard deviation decaying during training.
- **Alternate training**: when the training procedure is initiated (after a fixed number of epochs) the discriminator of the less represented class is not trained at every iteration. Since the number of samples is small the discriminator is prone to overfitting, thus it can discriminate between real and fake samples with very low uncertainty. Hence, training it less frequently than the generator gives the latter more time to learn.

4. Experiments

Classifier results replication. The first attempt that was performed is the replication of the results of [27] without augmentation. However, the results obtained with the classifier used in the paper were far worse than those reported by the authors. The main problem is that the authors did not specify how the selection of the training and testing subset is performed. For this reason, the methods described in Section 3.1 were employed. Confidence filtering provided the best results. However, the gap concerning the performance reported in the paper was still very high. For the following experiments, a lighter classifier (3 Convolution-ReLU-MaxPooling-BatchNorm blocks + 3 fully connected layers) has been adopted, allowing shorter training times while keeping similar performances. Class weights have also been used to tackle the high-class imbalance.

Used dataset. The original FER2013 dataset has 7 classes, namely: *Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral*. The total number of samples is 35887, and some classes are more represented then others (*e.g., Disgust* has 547 samples while *Happy* 8989). However, due to the previous observations, a filtered version of the dataset was used. The filtered dataset instead has 28941, keeping the same imbalanced nature of the classes (*e.g., Disgust* samples: 421, *Happy* samples: 7967). 100 images per class are kept apart to test the classifier, while the remaining ones are used for training.

Performed experiments. To compare the quality of the synthetic data generated using CycleGAN and especially ECycleGAN, several experiments were conducted using different parameters and finally compared. The final performance of the classifier on the original and augmented dataset are assessed using 3x10-fold cross-validation and averaging the results.

CycleGAN. For these first experiments, the employed CycleGAN architecture and parameters are the same as [27]. The paper does not consider the identity loss, so different experiments were performed weighting the identity loss in different ways ($\lambda_{idt} = [0.0; 0.3; 0.5]$). For each configuration, after training, 100 *Disgust* images are generated starting from a fixed set of *Neutral* images and added to the original dataset to assess the impact on the classifier performance. Initially, the cycle consistency loss weight was set to $\lambda_{cyc} = 10$ as in [27]. However, given that the generated images were extremely similar to the input ones, the λ_{cyc} was fixed to 1 for most of the following attempts, leading to a more significant impact on the image, while maintaining a convincing cycle consistency and a low reconstruction error. Finally, some generated samples are also manually selected to compare the different settings from a qualitative perspective.

ECycleGAN. The experiments of ECycleGAN started replicating the same base configuration reported in [24], which makes use of WGAN-GP loss function [6] and RDNBs with 3 dense-blocks, each composed of 5 convolutional densely connected sub-blocks. Unfortunately due to the limited available resources neither training such a powerful network nor performing a complete hyper-parameters tuning is feasible. Thus, the number of dense blocks and sub-blocks is decreased to 2 and 3 respectively. Given the poor results and the divergence problems that arose using the proposed WGAN-GP, the following experiments were performed using LSGAN [16] that resulted in a more stable training. The coefficient for the Cycle-loss λ_{cyc} is set to the same value used in the CycleGAN case, and the same holds for the λ_{idt} values. Regarding the perceptual loss, the balancing coefficients α and β are both set to 0.5 after some trial and error, and the feature map considered for the feature-loss component is the one generated by the last convolutional layer of the VGG19 feature-extractor. Finally, the residual scaling of the RDNBs is tuned by choosing between the values $\lambda_{idt} = [0.3; 0.5; 0.7]$, balancing the influence of the residual connection.

Given that the discriminator of the less represented class tends to overfit in the previous models, the model providing the more promising results was selected and enhanced with the techniques reported in Section 3.2 to try to mitigate this problem. More specifically, Label-smoothing was applied with a label flip probability of 1%. Instance-noise was implemented using a Gaussian noise with a standard deviation of 0.1 and 0.05, while in the case of Alternate-training the training ratio generator-discriminator was set to $10 : 1$ or $5 : 1$. As done for CycleGAN, for each trained model, 100

Disgust images are generated starting from the same fixed set of *Neutral* images for better comparison and added to the original dataset to assess the impact on the classifier performance. Then, the most promising model is used to generate also 200 and 500 images to check whether a bigger augmentation can further boost the classifier performance. Finally, a few experiments were performed to generate *Surprise* images from *Neutral* ones, to check the impact of the imbalance gap.

Results and discussion. The obtained results were evaluated both qualitatively and quantitatively. In particular, a qualitative evaluation of the images generated by each model setting was performed. Instead, concerning the classifier, the performances were assessed by measuring the per-class and mean *Precision*, and the per-class and mean *Receiver Operating Characteristic - Area Under the Curve (AUC)*. The idea is that the Precision can provide a first very intuitive indication of the classifier's capabilities, while the AUC tells how much the model is capable of distinguishing between classes (in a one-vs-all setting).

Qualitative. The main improvement from CycleGAN to ECycleGAN is related to the quality of the generated images. Also, the ratio between the number of images of the two domains, and the consequent loss behavior, have a huge impact on the quality of the generated samples. The smaller the gap (*e.g., Neutral-Surprise* translation), the better the quality. The techniques proposed in 3.2 were adopted to force the loss behavior of the Neutral-Disgust experiment in which the samples gap is far bigger, to be similar to the Neutral-Surprise one, wishing to increase the quality of the results. While the loss behavior was effectively changed as desired, the quality of the images does not show a convincing improvement.

Quantitative. The same classification setup is used for all the experiments, taking into consideration previously cited metrics for evaluation. The classification performance on the non-augmented filtered dataset is used as the baseline.

Among the CycleGAN experiments there is a noticeable improvement in the *Disgust* precision w.r.t. the baseline, but at the same time the other classes' precision decreases, resulting in a mean precision that is similar to the baseline one. This is because only the samples that are clearly disgusted are classified as such, reducing the false positives of class *Disgust*. The other less certain Disgust samples are assigned to other classes, increasing their false positives and reducing their precision. According to the AUC the quality of the augmentation depends on the value of λ_{idt} used: the higher the value the lower the AUC. Thus, it is possible to derive that synthetic samples that have features similar to real faces (more probable with higher λ_{idt}) are more easily misclassified since the stronger identity constraint does not allow to make them look disgusted. Thus, CycleGAN architecture is not powerful enough to allow better discrimination between classes, leading to a mean AUC which is similar to that of the baseline. Regarding the ECycleGAN the first experiments were performed to establish the best value of *residual scaling (α)* and *identity loss weight* λ_{idt} hyper-parameters. When using $\lambda_{idt} = 0.5$ and $\alpha = 0.5$ the best overall improvement is obtained for both metrics. Different values of these parameters cause input images to be modified too much or not enough. Using the best model, different techniques to avoid overfitting were employed, however even though there is an improvement concerning the generator loss convergence, no significant performance boost was observed. Increasing the number of augmented images to 200 or 500 decreases the performance due to the introduction of too many bad-quality samples. Finally, considering a more represented class such as *Surprise*, the ECycleGAN qualitative performance is surprising also w.r.t. CycleGAN. However, there is still not much improvement in the classifier performances since this class is already quite distinguishable from the others and thus, the introduction of some lower quality samples can even slightly decrease the performances w.r.t. the baseline.

5. Conclusion

The experiments report partial effectiveness of the proposed generative approach for Facial Expression Data Augmentation. Despite an acceptable qualitative result on some samples, the influence of bad generated samples is too high to augment the dataset significantly. While introducing a low number of generated samples (around 100) leads to a limited performance boost, introducing more samples (from 200 on), the major influence of the bad-quality ones results in a degradation of the classification performances since the intra-class diversity increases too much. However, the improvement w.r.t. the previously proposed CycleGAN model is present from both qualitative and quantitative viewpoints. Further experiments and studies could be conducted by adopting the complete ECycleGAN architecture and performing a finer hyper-parameter tuning if resources are available. Moreover, other filtering methods for the input dataset could be explored, as well as an instance selection algorithm to choose the best samples generated by ECycleGANs. Other techniques to avoid overfitting the discriminator could be considered. Finally, pretraining the VGG19 network used for the perceptual loss with a pretext task on human faces could help to extract more suitable features to be considered for the consistency losses, boosting the performance.

References

[1] T. DeVries, M. Drozdzal, and G. W. Taylor. Instance selection for gans. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13285–13296. Curran Associates, Inc., 2020. 4

[2] T. DeVries and G. W. Taylor. Improved regularization of convolutional neural networks with cutout, 2017. 2

[3] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 1, 2, 3

[5] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, pages 117–124. Springer, 2013. 2, 4

[6] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017. 4, 5

[7] A. M. Hafiz and G. M. Bhat. A survey on instance segmentation: state of the art. *International journal of multimedia information retrieval*, 9(3):171–189, 2020. 1

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015. 4

[9] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340, 2001. 1, 3

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 4

[11] 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. 3

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 1

[13] P. Li, X. Li, and X. Long. Fencemask: A data augmentation approach for pre-extracted image features. *arXiv preprint arXiv:2006.07877*, 2020. 2

[14] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017. 1

[15] Y. Lu, D. Chen, E. Olaniyi, and Y. Huang. Generative adversarial networks (gans) for image augmentation in agriculture: A systematic review. *Computers and Electronics in Agriculture*, 200:107208, 2022. 2

[16] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 3, 4, 5

[17] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans, 2016. 4

[18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[19] A. Singh and P. Singh. Image classification: A survey. *Journal of Informatics Electrical and Electronics Engineering (JIEEE)*, 1(2):1–9, 2020. 1

[20] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 4

[21] C. K. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Huszár. Amortised map inference for image super-resolution, 2016. 5

[22] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 4

[23] S. Yang, W. Xiao, M. Zhang, S. Guo, J. Zhao, and F. Shen. Image data augmentation for deep learning: A survey, 2022. 1, 2

[24] X. Zhang and C. Zhou. Ecyclegan: Enhanced cycle-consistent generative adversarial networks. In *Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence*, ICCAI '20, page 374–379, New York, NY, USA, 2020. Association for Computing Machinery. 2, 3, 4, 5

[25] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020. 2

[26] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2017. 2, 3, 4

[27] X. Zhu, Y. Liu, J. Li, T. Wan, and Z. Qin. Emotion classification with data augmentation using generative adversarial networks. In D. Phung, V. S. Tseng, G. I. Webb, B. Ho, M. Ganji, and L. Rashidi, editors, *Advances in Knowledge Discovery and Data Mining*, pages 349–360, Cham, 2018. Springer International Publishing. 2, 4, 5

[28] Z. Zou, Z. Shi, Y. Guo, and J. Ye. Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*, 2019. 1