Conditional normalizing flows for nonlinear remote sensing image augmentation and classification

Victor Enescu, Hichem Sahbi

CONDITIONAL NORMALIZING FLOWS FOR NONLINEAR REMOTE SENSING IMAGE AUGMENTATION AND CLASSIFICATION

Victor Enescu *Hichem Sahbi*

Sorbonne University, CNRS, LIP6, F-75005, Paris, France

ABSTRACT

Deep neural networks have recently shown outstanding performances in remote sensing image classification. The success of these models is highly reliant on the availability of large collections of hand-labeled training images which are usually scarce. Data augmentation mitigates this scarcity by enriching labeled training sets using different geometric and photometric transformations, or by relying on deep generative models. In this paper, we investigate the potential of generative models, and particularly normalizing flows (NFs), in remote sensing image augmentation and classification. The main contribution relies on a novel conditional NF model that achieves a bidirectional mapping of images between ambient and latent spaces with the particularity of learning disentangled multi-modal distributions through image classes. The proposed NF also achieves nonlinear augmentations in highly intricate ambient spaces by mapping images to latent spaces where augmentations become linear and more tractable. Extensive experiments conducted on the EuroSAT benchmark show the benefit of our NF-based augmentation when learning vision transformers.

Index Terms— Generative Models, Normalizing Flows, Transformers, Remote Sensing Image Augmentation and Classification

1. INTRODUCTION

Remote sensing image classification seeks to automatically assign labels to the visual content of aerial and satellite images [1]. This task is challenging as observed scenes are subject to different sources of variability due to eclectic contents and sensors. This variability can either be attenuated using different normalization techniques (such as registration, radiometric corrections, etc), or considered as a part of scene appearance modeling. The latter is particularly successful and relies on different machine learning models. Among these models, deep neural networks [2–4] are particularly interesting but their success is tributary to the availability of large hand-labeled image collections. For some tasks, including remote sensing image classification, labeled training collections are difficult to obtain and their hand-labeling is cumbersome [5, 6]. Alternative and more effective approaches rely on data augmentation (DA) in order to mitigate the scarcity of labeled data.

Data augmentation is nowadays becoming mainstream, in training deep neural networks, and its purpose is to create artificial data by leveraging different transformations. Staple augmentation methods rely on geometric and photometric transformations (mirroring, etc.) while more sophisticated DA methods combine images through different operations such as mixing [7] and interpolation [8]. Alternative augmentation methods — producing more realistic images — are based on deep generative models. Their principle consists in mapping images from *ambient* to *latent* spaces, achieving augmentation in the latent spaces, prior to reconstructing images in the ambient spaces. In particular, variational autoencoders (VAEs) [9–12] follow this principle and proceed either by (i) interpolating images in the latent spaces while adversarially learning class-dependent masks that make the resulting interpolated images realistic, or (ii) by injecting noise in the latent image representations.

Existing generative models, including VAEs [9–12] and generative adversarial networks (GANs) [2, 13, 14], map and reconstruct input data through low dimensional bottlenecks. However, these models suffer either from (i) a significant drop in their generative properties due to these bottlenecks (which make these mappings non-bijective and thereby data reconstruction challenging), or (ii) instabilities when learning these models. Other generative approaches, known as normalizing flows (NFs) [15–18] are rather more appropriate, and consist in learning bijective mappings between equidimensional (latent and ambient) spaces leading to exact density estimation, and also better generative properties [19]. Hence, NFs hold many promises towards learning more powerful nonlinear data augmentations that ultimately lead to better discriminative models, including transformers, which are known to be highly effective but data/label hungry.

In this paper, we devise a novel conditional NF model for nonlinear image augmentation that learns a bidirectional mapping between ambient and latent spaces. Unlike previous methods that rely on unimodal, mingled latent spaces, and unconditional NFs for augmentation [12, 20, 21], our proposed NF makes it possible to model the distribution of image classes as disentangled multi-modal gaussians where each one is assigned to a single class. Consequently, our

NF design achieves both conditional image generation (i.e., augmentation) and classification thereby acting as a pseudo-oracle capable of labeling the augmented images. It's worth noticing that our NF model also renders the interpolation of images lying on highly nonlinear manifolds in the ambient space possible thanks to the NF which makes image interpolations more tractable (linear) in the latent space. Considering this issue, we investigate different augmentation methods based on single and pairwise image interpolations as well as more *robust* augmentation methods operating in the span of the eigenvectors of training data in the latent space. Extensive experiments, conducted on the EuroSAT benchmark, show the impact of our augmentations on vision transformers trained from scratch, with different (labeled) data regimes.

2. PROPOSED METHOD

2.1. A Glimpse on Normalizing Flows

Let $\mathbf{X}$ be a random variable standing for all possible images taken from an existing but *unknown* probability distribution $P_{\mathbf{X}}$ in an ambient space $\mathcal{X} \subseteq \mathbb{R}^d$. Considering $\mathbf{Z}$ as a latent representation associated to $\mathbf{X}$ drawn from a *known* probability distribution $P_{\mathbf{Z}}$ in a latent space $\mathcal{Z} \subseteq \mathbb{R}^d$; normalizing flows aim at learning a diffeomorphism f from $\mathcal{X}$ to $\mathcal{Z}$ (together with its inverse g). Given $\mathbf{x} \in \mathcal{X}$, one may write

$$P_{\mathbf{X}}(\mathbf{x}) = P_{\mathbf{Z}}(f(\mathbf{x})) \left| \det \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \right| = P_{\mathbf{Z}}(f(\mathbf{x})) \left| \det \mathbf{J}_{f(\mathbf{x})} \right|, \quad (1)$$

where $\mathbf{J}_{f(\mathbf{x})} \in \mathbb{R}^{d \times d}$ is the Jacobian of f w.r.t. $\mathbf{x}$ and $|\det(.)|$ stands for determinant magnitude. In practice, f is a neural network composed of several smaller invertible flows chosen to make $\mathbf{J}_{f(\mathbf{x})}$ computationally efficient. As defined in [15, 17], each flow is usually made of an Actnorm layer, an Invertible 1×1 convolution, and a Coupling Layer stacked together. Let $\mathbf{x}_{1:d}$ be a d-dimensional vector, a Coupling Layer maps $\mathbf{x}_{1:d}$ to two subvectors $\tilde{\mathbf{x}}_{1:d/2}$ and $\tilde{\mathbf{x}}_{d/2+1:d}$ being $\tilde{\mathbf{x}}_{1:d/2} = \mathbf{x}_{1:d/2}$ and $\tilde{\mathbf{x}}_{d/2+1:d} = \mathbf{x}_{d/2+1:d} \odot \exp(s(\mathbf{x}_{1:d})) + b(\mathbf{x}_{1:d})$, $s(.)$, $b(.)$ are two neural networks, $\odot$ the Hadamard product and $\exp(.)$ is applied entrywise. Invertible 1×1 convolutions are generalized permutation layers that enhance expressivity by allowing permutations between image channels to be learned [17]. An Actnorm layer is an invertible equivalent of Batch Normalization that increases stability and performance. NFs are usually trained to minimize the negative log-likelihood of Eq. 1. From transport theory point of view [22], NFs pushforward a complex ambient distribution into a simpler latent one as the mono-modal normal. In what follows, we first describe our extension of standard NFs that makes them conditional by allowing their latent distributions to be multi-modal and class-dependent.

2.2. Conditioning

Let $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$ denote a collection n of labeled images with $\mathbf{x}_i$ belonging to an ambient space $\mathcal{X}$ and $\mathbf{y}_i$ its underlying class-label taken from a discrete set $\mathcal{Y} = \{1, \ldots, K\}$. Given a pair $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, one may write the conditional form of Eq. 1 as

$$P_{\mathbf{X}}(\mathbf{x}|\mathbf{y}) = P_{\mathbf{Z}}(f(\mathbf{x})|\mathbf{y}) \left| \det \mathbf{J}_{f(\mathbf{x})} \right|, \quad (2)$$

here $P_{\mathbf{Z}}(.|\mathbf{y})$ is set a priori to a given distribution, *viz.*, gaussian, denoted for a given class $\mathbf{y}$ as $\mathcal{N}_{\mathbf{y}}$. Our goal here is to train the parameters of the NF (denoted as Θ) together with the hyperparameters of the underlying gaussians (referred to as $\Psi = \{(\mu_{\mathbf{y}}, \Sigma_{\mathbf{y}})\}_{\mathbf{y} \in \mathcal{Y}}$) while guaranteeing better generation performances of the resulting NF. To further ensure gaussians are not overlapping through classes, a Kullback-Leibler Divergence (KLD) is applied amongst all the gaussian pairs

$$\mathcal{L}_{KLD}(\Psi) = - \sum_{\mathbf{y}, \mathbf{y}' \in \mathcal{Y} : \mathbf{y} \neq \mathbf{y}'} \text{KLD}(\mathcal{N}_{\mathbf{y}} \,||\, \mathcal{N}_{\mathbf{y}'}). \quad (3)$$

With the above term, we define our global loss as

$$\mathcal{L}(\Theta, \Psi) = \mathcal{L}_{NF}(\Theta, \Psi) + \lambda \mathcal{L}_{KLD}(\Psi). \quad (4)$$

This formulation has the advantage of producing a NF that can also obtain good classification performance, by assigning labels using $\mathbf{argmax}_{\mathbf{y}} P_{\mathbf{Z}}(\mathbf{z}|\mathbf{y})$. In order to optimize the loss in Eq. 4, we use an EM-like procedure. Two steps are alternatively applied: in the E-step, we fix the hyperparameters Ψ, and we train our NF while in the M-step we fix the NF parameters Θ and we optimize only Ψ using gradient descent. In order to make the training of the covariances $\{\Sigma_{\mathbf{y}}\}_{\mathbf{y}}$ tractable while guaranteeing the positive definiteness of these matrices, we consider anisotropic diagonal $\{\Sigma_{\mathbf{y}}\}_{\mathbf{y}}$ using a reparametrization function $\psi(.) = a(1+\exp\{-\beta(.)\})^{-1} + c$, forcing them to remain strictly positive after gradient descent. Indeed, the positive variables a, b and c respectively control the amplitude (scale), the slope (smoothness) as well as the shift of the reparametrization ψ. Furthermore, $\frac{a+c}{c}$ controls the conditioning of the learned diagonal covariance matrices, and to some extent, the shape of the learned multi-modal gaussians in the latent space. The aforementioned E and M steps are run using two disjoint subsets $\mathcal{D}_{\text{nf}}$ and $\mathcal{D}_{\text{g}}$ (taken from $\mathcal{D}$) in order to mitigate the co-adaptation between Θ and Ψ. In practice, $|\mathcal{D}_{\text{g}}| = 0.1 \times |\mathcal{D}|$ and $|\mathcal{D}_{\text{nf}}| = 0.9 \times |\mathcal{D}|$.

2.3. Augmentation

In what follows, we generate labeled images by disrupting the original ones in the latent space, using unary and pairwise as well as principal modes perturbations and interpolations.

Unary Image Augmentation. The latent representation of a given image $\mathbf{x} \in \mathcal{D}$ is disrupted by adding a normal noise $\epsilon \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ as

$$\hat{\mathbf{x}} = g\big(f(\mathbf{x}) + \alpha \, (\epsilon \odot \text{diag } \boldsymbol{\Sigma}_{\mathbf{y}}^{\frac{1}{2}})\big), \quad (5)$$

being $g = f^{-1}$ the NF generation function, $\alpha > 0$, $\mathbf{y}$ the label of $\mathbf{x}$, and $\mathbf{\Sigma}_{\mathbf{y}}^{\frac{1}{2}}$ the square root of the covariance matrix of class $\mathbf{y}$. With this perturbation, and thanks to the learned multi-modal latent distributions, the generated image inherits the same label as the original image.

Pairwise Image Augmentation. Given any arbitrary $\mathbf{x}_1$ and $\mathbf{x}_2$ in $\mathcal{D}$, a new $\hat{\mathbf{x}}$ is generated by combining the latent representations of these images, and inverting the result as

$$\begin{array}{rcl} \hat{\mathbf{z}} & = & t \cdot f(\mathbf{x}_1) + (1-t) \cdot f(\mathbf{x}_2), \quad \text{with } t \in [0,1] \\ \hat{\mathbf{x}} & = & g(\hat{\mathbf{z}}). \end{array} \tag{6}$$

We also consider a *rescaled* variant of $\hat{\mathbf{x}}$ (related to [23]) as

$$\hat{\mathbf{x}} = g\left(\left[t \cdot \|\mathbf{z}_1\| + (1-t) \cdot \|\mathbf{z}_2\| \right] \cdot \frac{\hat{\mathbf{z}}}{\|\hat{\mathbf{z}}\|} \right), \tag{7}$$

being $\mathbf{z}_1 = f(\mathbf{x}_1)$, $\mathbf{z}_2 = f(\mathbf{x}_2)$ the NF latent representations of images $\mathbf{x}_1$, $\mathbf{x}_2$, and $\|.\|$ the L2 norm with t a coefficient that controls the weight given to each image. Note that (i) if $\mathbf{x}_1$ and $\mathbf{x}_2$ have the same label $\mathbf{y}$, then the generated $\hat{\mathbf{x}}$ inherits $\mathbf{y}$, (ii) otherwise, an extra step is considered (see later in this section and also Eq. 11) in order to decide about its label and whether it is reliable. The settings (i) + (ii) are respectively dubbed as *intra* and *inter* (class) augmentation.

Principal Modes Augmentation. Given a class $\mathbf{y}$ and the underlying *dense* covariance matrix $\widehat{\mathbf{\Sigma}}_{\mathbf{y}}$ estimated on the latent representations of training images belonging to class $\mathbf{y}$, we consider the singular value decomposition (SVD) of $\widehat{\mathbf{\Sigma}}_{\mathbf{y}}$

$$\widehat{\mathbf{\Sigma}}_{\mathbf{y}} = \mathbf{V}_{\mathbf{y}} \cdot \Lambda_{\mathbf{y}} \cdot \mathbf{V}_{\mathbf{y}}^{\top}, \tag{8}$$

with $\mathbf{V}_{\mathbf{y}}$ and $\Lambda_{\mathbf{y}}$ being respectively the eigenvectors and eigenvalues of $\widehat{\mathbf{\Sigma}}_{\mathbf{y}}$. Augmentation is achieved by disrupting data in the span of the eigenvectors $\mathbf{V}_{\mathbf{y}}$ as

$$\begin{array}{rcl} \hat{\mathbf{z}} & = & \pi_{\mathbf{y}}^{-1}\big(\pi_{\mathbf{y}}(\mathbf{z}) + \alpha \, (\epsilon \odot \text{diag } \Lambda_{\mathbf{y}})\big) \\ \hat{\mathbf{x}} & = & g(\hat{\mathbf{z}}), \end{array} \tag{9}$$

here $\pi_{\mathbf{y}}$ defines a projection from the latent to the eigenspace (corresponding to the aforementioned SVD decomposition) and $\pi_{\mathbf{y}}^{-1}$ its backprojection in the latent space, i.e.,

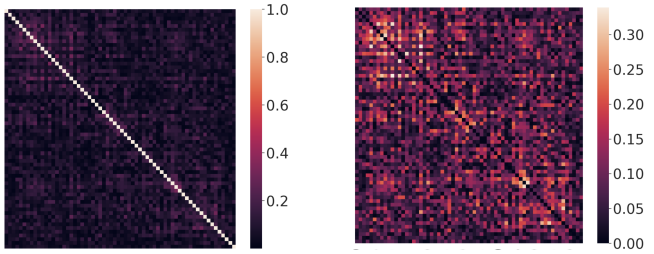
$$\begin{array}{rcl} \pi_{\mathbf{y}}(\mathbf{z}) & = & (\mathbf{z} - \mu_{\mathbf{y}}) \cdot \mathbf{V}_{\mathbf{y}} \\ \pi_{\mathbf{y}}^{-1}(\hat{\mathbf{z}}) & = & \hat{\mathbf{z}} \cdot \mathbf{V}_{\mathbf{y}}^{\top} + \mu_{\mathbf{y}}, \end{array} \tag{10}$$

with $\mu_{\mathbf{y}}$ being the mean of training images in the latent space (belonging to class $\mathbf{y}$). Similarly to unary augmentation, the generated image inherits the same label as the original image. As shown in experiments, this principal modes perturbation scheme is more relevant (compared to the two others) as it considers the eigenvectors that define the principal variation modes of data in the latent space. Besides, it takes into account the dependencies between all latent dimensions (see Fig. 1) prior to disentangle those dimensions using SVD. This eigenspace can also be interpreted as an extra latent space — built on top of the NF latent space — that provides uncorrelated and disentangled latent representations.

Label reliability. Given a generated image $\hat{\mathbf{x}}$, and $\mathbf{y}$ its possible label, this pair is kept if

$$P_{\mathbf{Z}}(\hat{\mathbf{z}}|\mathbf{y}) \geq \tau_{\mathbf{y}}, \tag{11}$$

being $\tau_{\mathbf{y}}$ a cut-off threshold that keeps 95% of training images (belonging to class $\mathbf{y}$) whose likelihoods $\{P_{\mathbf{Z}}(.|\mathbf{y})\}$ are the largest. Note that for pairwise augmentation experiments, we consider small t-values (in Eq. 6) so only the generated images whose labels correspond to $\mathbf{y}_1$ are kept in practice.



(a) (b)

Fig. 1: Figure 1a shows the correlation matrix obtained from $\widehat{\mathbf{\Sigma}}_{\mathbf{y}}$. Figure 1b shows this correlation (after subtracting the identity matrix and omitting the signs of these correlations for better visibility). We observe high correlations suggesting that achieving principal modes augmentation in the span of the eigenvectors of the dense reestimated covariance matrices (instead of the diagonal ones) is more relevant and faithful w.r.t. training data.

3. EXPERIMENTS

3.1. Dataset and Settings

Experiments have been conducted on the EuroSAT dataset [4, 24] which includes 27k satellite images of 64×64 pixels belonging to 10 different classes; 80% of these data (21.6k images) are used for training and the remaining 20% (5.4k images) for testing. This test set has the same cardinality as [3,4,25] and differs from [2] which includes 2.7k images. The discriminative model used in our experiments is the vision transformer in [26] which is suitable for mid-scale datasets; it uses a Swin architecture [27] of 7M parameters trained with the Adam optimizer [28]. Besides standard data augmentations, we also use cutmix [7], mixup [8], auto-augment [29] and repeat augment [30] as well as random erasing [31] prior to train the transformer. In all experiments, the Swin model is trained for 100 epochs with a batch size of 256 (see [26] for more details). The NF backbone is taken from Matrix Exponential Flow [32] which is trained for 350 epochs using the Adamax optimizer [28] with a learning rate of 0.01, decreased by half at the 200, 250 and 300 epochs. The NF model is also

NF Augmentation Used	10% Data	25% Data	50% Data	100% Data	Latent Code Mixing
Baseline (no NF augmentation)	96.47	98.33	98.83	99.15	✗
Unary ($\alpha = 0.24$)	96.59	**98.44**	99.82	99.12	✗
Pairwise+Intra ($t = 0.02$)	96.43	98.41	**98.94**	99.18	✓
Pairwise+Intra+rescaled ($t = 0.02$)	96.39	98.28	98.88	99.14	✓
Pairwise+Inter ($t = 0.04$)	96.57	98.21	98.79	99.12	✓
Pairwise+Inter+rescaled ($t = 0.06$)	96.55	98.36	98.84	99.14	✓
Principal Modes ($\alpha = 0.035$)	**96.72**	98.35	98.86	**99.21**	✗

Table 1: Comparison and ablation study of the proposed augmentation methods for different training data regimes. Each reported accuracy corresponds to an average over 5 different runs. Again, intra (resp. inter) corresponds to pairwise interpolations involving training images with the same (resp. different) labels. Latent code mixing stands for generated images obtained by combining 2 distinct original images. Green and red cells respectively indicate whether the obtained accuracy is higher or lower than the baseline.

trained with standard augmentations including horizontal and vertical flip, padding with 12 degree rotation, random crop and color jitter that modifies the brightness+contrast with an intensity of 25, and the saturation with an intensity of 12. All these models are trained using Pytorch 2 on Nvidia V100 and A100 GPUs.

3.2. Performances and Comparison

Table 1 shows the performance of the proposed augmentation methods, at different data regimes, obtained by training the NF and the transformers on 10%, 25%, 50% and 100% of the training sets which are afterwards augmented by a factor of 20. From these results, principal modes augmentation consistently outperforms the baseline while pairwise interpolations, involving images with the same labels, are generally worse than the baseline at lower data regimes and better at higher regimes; a different behavior is observed when images, involved in augmentation, have different labels particularly at low data regimes (see also Table 2).

		NF-based Augmentation Configurations					
Data %	Basic Baseline	Unary	Pairwise +intra	Pairwise +intra +rescaled	Pairwise +inter	Pairwise +inter +rescaled	Principal Modes
10%	90.81	92.56	90.69	90.66	91.04	91.65	**92.64**
5%	87.29	89.59	87.21	87.05	87.27	89.26	**90.27**
1%	73.2	75.86	73.36	73.19	73.87	75.13	**80.26**

Table 2: Accuracies obtained at 1%, 5% and 10% data regimes. "Basic" baseline uses flipping and cropping transformations only.

Method	Accuracy	Generative Model	Discriminative Model	Pretrained	IS	Params (M)
Ours	99.21	NF	Transformer	✗	64^2	7
[2]	99.2	GAN	Wide Resnet50	✓	64^2	67
[3]	99.22	✗	Transformer	✓	384^2	307
[25]	99.2	✗	Transformer	✓	384^2	307
[4]	99.2	✗	Resnet50	✓	224^2	24

Table 3: Comparison of our method against related SOTA. IS stands for Image Size (in pixels). We observe that our method, at least, matches the accuracy of larger pretrained models that operate on larger images.

Fig. 2: Sample of generated images using principal modes augmentation (upper) versus original training images (lower). From left-to-right and top-to-bottom, classes correspond to: annual crop, highway, sea/lake, pasture, river, residential, permanent crop, forest, herbaceous vegetation and industrial. We observe that images have different appearances while preserving the structures of the original classes.

Table 3 shows a comparison of our best augmentation method (namely principal modes) against related methods using GANs [2] and various discriminative classifiers on EuroSAT [3,4,25]. The accuracies obtained match those of larger, fine-tuned discriminative models, trained on larger-size images. Finally, Fig. 2 shows a sample of augmented images together with the underlying original training images.

4. CONCLUSION

In this paper, we introduce a novel image augmentation method based on normalizing flows. The strength of our method resides in its ability to learn a bidirectional mapping that takes data from a highly nonlinear ambient space to a well disentangled latent space where class distributions become more tractable. This mapping also allows achieving nonlinear data augmentation in the ambient space through the latent space and the NF mapping. The impact of this augmentation process is demonstrated through extensive experiments in remote sensing image classification using transformers on the EuroSAT benchmark. These experiments show a clear positive impact of our augmentation on the trained transformers. As a future work, we are currently investigating the extension of this method to other transformer models and other applications.

Acknowledgment. This work has been supported with computer and storage resources by GENCI at IDRIS thanks to the grant 2023-AD011013954 on the supercomputer Jean Zay with the V100 partition.

5. REFERENCES

[1] Quentin Oliveau and Hichem Sahbi, "Learning attribute representations for remote sensing ship category classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 6, pp. 2830–2840, 2017.

[2] Oluwadara Adedeji, Peter Owoade, Opeyemi Ajayi, and Olayiwola Arowolo, "Image augmentation for satellite images," *arXiv preprint arXiv:2207.14580*, 2022.

[3] Andrea Gesmundo, "A continual development methodology for large-scale multitask dynamic ml systems," *arXiv preprint arXiv:2209.07326*, 2022.

[4] Maxim Neumann, Andre Susano Pinto, Xiaohua Zhai, and Neil Houlsby, "In-domain representation learning for remote sensing," *arXiv preprint arXiv:1911.06721*, 2019.

[5] Hichem Sahbi and Sebastien Deschamps, "Adversarial label-efficient satellite image change detection," in *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2023, pp. 5794–5797.

[6] Sebastien Deschamps and Hichem Sahbi, "Reinforcement-based display selection for frugal learning," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 1186–1193.

[7] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *In IEEE/CVF ICCV*, 2019, pp. 6023–6032.

[8] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[9] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[10] Christopher Beckham, Sina Honari, Vikas Verma, Alex M Lamb, Farnoosh Ghadiri, R Devon Hjelm, Yoshua Bengio, and Chris Pal, "On adversarial mixup resynthesis," *Advances in neural information processing systems*, vol. 32, 2019.

[11] David Berthelot, Colin Raffel, Aurko Roy, and Ian Goodfellow, "Understanding and improving interpolation in autoencoders via an adversarial regularizer," *arXiv preprint arXiv:1807.07543*, 2018.

[12] Genki Osada, Budrul Ahsan, Revoti Prasad Bora, and Takashi Nishide, "Regularization with latent space virtual adversarial training," in *ECCV 2020*. Springer, 2020, pp. 565–581.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[14] Jonathan Howe, Kyle Pula, and Aaron A Reite, "Conditional generative adversarial networks for data augmentation and adaptation in remotely sensed imagery," in *Applications of Machine Learning*. SPIE, 2019, vol. 11139, pp. 119–131.

[15] Laurent Dinh, David Krueger, and Yoshua Bengio, "Nice: Non-linear independent components estimation," *arXiv preprint arXiv:1410.8516*, 2014.

[16] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio, "Density estimation using real nvp," *arXiv preprint arXiv:1605.08803*, 2016.

[17] Durk P Kingma and Prafulla Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *Advances in neural information processing systems*, vol. 31, 2018.

[18] Hanlin Wu, Ning Ni, Shan Wang, and Libao Zhang, "Blind super-resolution for remote sensing images via conditional stochastic normalizing flows," *arXiv preprint arXiv:2210.07751*, 2022.

[19] Ivan Kobyzev, Simon J. D. Prince, and Marcus A. Brubaker, "Normalizing Flows: An Introduction and Review of Current Methods," vol. 43, no. 11, pp. 3964–3979.

[20] Oguz Kaan Yüksel, Sebastian U Stich, Martin Jaggi, and Tatjana Chavdarova, "Semantic perturbations with normalizing flows for improved generalization," in *In IEEE/CVF ICCV*, 2021, pp. 6619–6629.

[21] GENKI OSADA, Budrul Ahsan, and Takashi Nishide, "Mixed samples data augmentation with replacing latent vector components in normalizing flow," in *First Workshop on Interpolation Regularizers and Beyond at NeurIPS 2022*, 2022.

[22] Cédric Villani and American Mathematical Society, *Topics in Optimal Transportation*, Graduate studies in mathematics. American Mathematical Society, 2003.

[23] Chin-Wei Huang, Laurent Dinh, and Aaron Courville, "Augmented normalizing flows: Bridging the gap between generative flows and latent variable models," *arXiv preprint arXiv:2002.07101*, 2020.

[24] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth, "Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2018, pp. 204–207.

[25] Andrea Gesmundo and Jeff Dean, "An evolutionary approach to dynamic introduction of tasks in large-scale multitask learning systems," *arXiv preprint arXiv:2205.12755*, 2022.

[26] Hanan Gani, Muzammal Naseer, and Mohammad Yaqub, "How to train vision transformer on small-scale datasets?," in *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. 2022, BMVA Press.

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *In IEEE/CVF ICCV*, 2021, pp. 10012–10022.

[28] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[29] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le, "Autoaugment: Learning augmentation policies from data," *arXiv preprint arXiv:1805.09501*, 2018.

[30] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *In IEEE/CVF CVPR workshops*, 2020, pp. 702–703.

[31] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang, "Random erasing data augmentation," in *Proceedings of the AAAI conference on artificial intelligence*, 2020, vol. 34, pp. 13001–13008.

[32] Changyi Xiao and Ligang Liu, "Generative flows with matrix exponential," in *ICML*. PMLR, 2020, pp. 10452–10461.