



Swapping Semantic Contents for Mixing Images

Rémy Sun, Clément Masson, Gilles Hénaff, Nicolas Thome, Matthieu Cord

► To cite this version:

Rémy Sun, Clément Masson, Gilles Hénaff, Nicolas Thome, Matthieu Cord. Swapping Semantic Contents for Mixing Images. 2022 26th International Conference on Pattern Recognition (ICPR), Aug 2022, Montreal, Canada. pp.1280-1286, 10.1109/ICPR56361.2022.9956602 . hal-03951744

HAL Id: hal-03951744

<https://hal.science/hal-03951744>

Submitted on 23 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Swapping Semantic Contents for Mixing Images

Rémy Sun*, Clément Masson†, Gilles Hénaff†, Nicolas Thome‡ and Matthieu Cord*

*MLIA, ISIR, Sorbonne Université, Paris, France

†Thales Land and Air Systems, Elancourt, France

‡VERTIGO, CEDRIC, Conservatoire National des Arts et Métiers, Paris, France

Abstract—Deep architecture have proven capable of solving many tasks provided a sufficient amount of labeled data. In fact, the amount of available labeled data has become the principal bottleneck in low label settings such as Semi-Supervised Learning. Mixing Data Augmentations do not typically yield new labeled samples, as indiscriminately mixing contents creates between-class samples. In this work, we introduce the SciMix framework that can learn to replace the global semantic content from one sample. By teaching a StyleGan generator to embed a semantic style code into image backgrounds, we obtain new mixing scheme for data augmentation. We then demonstrate that SciMix yields novel mixed samples that inherit many characteristics from their non-semantic parents. Afterwards, we verify those samples can be used to improve the performance semi-supervised frameworks like Mean Teacher or Fixmatch, and even fully supervised learning on a small labeled dataset.

I. INTRODUCTION

Deep architectures have proven capable of reliably solving a variety of tasks such as classification [1], [2], object detection [3] or machine translation [4]. This is however contingent on there being a large amount of labeled data to train models on. This is seldom the case in practical applications where labellisation tends to be costly.

Data Augmentation [5], [6] - the creation of artificial samples from existing ones - has long been used to help models train on small datasets. Of particular interest in low label settings like Semi-Supervised Learning [7], [8] (SSL), Mixing Samples Data Augmentations [9], [10] (MSDA) can be used to combine the few samples that are either labeled or reliably pseudo-labeled with the large pool of unlabeled data. Unfortunately, Mixing Data Augmentations mix contents indiscriminately and as such create between-class hybrids for classification. While such hybrids have proven very useful for model regularization [11], [12], this process strongly perturbs the semantic information from reliably labeled samples.

We argue that, with the right adjustments, mixing data augmentations can still be used to teach semantic invariance to neural networks. Indeed, if we can mix the semantic content of one sample with the non-semantic content of another, then the generated samples will still be actual in-class samples. For instance, Fig. 1 shows that mixing two street numbers with MixUp or CutMix typically leads to no real number appearing on the image whereas carefully selecting the contents to be mixed leads to a mixed sample that remains realistic.

In this paper, we introduce SciMix a new framework that learns to separate semantic from non semantic content and generate hybrids that preserve most of the specified



Fig. 1. Standard mixing augmentations mix contents indiscriminately whereas our method mixes the semantic number from one sample with the background information from another.

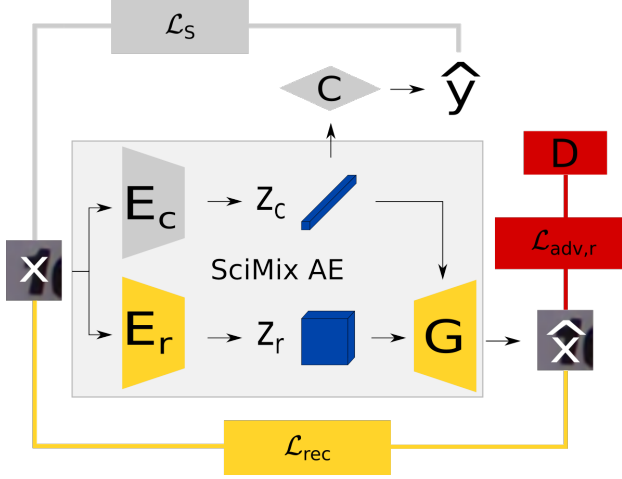
non semantic content while still properly representing the required semantic content. Moreover, we demonstrate hybrids generated from this framework improve model performance through extensive low label experiments, primarily on the Semi-Supervised Learning problem (CIFAR10 and SVHN). We therefore propose three main contributions: **1)** A new mixing paradigm designed to create artificial in-distribution samples that embed the non-semantic content of one sample into the non-semantic context of another. This new approach generates a new type of data augmentation for deep learning. **2)** A new auto-encoding architecture and associated learning scheme that trains a generator to mix semantic and non-semantic contents. In particular, we purposefully train a model to separate semantic and non-semantic contents into two representations, and train a style-inspired generator to embed the semantic content (“style” code) into the non-semantic background (traditional input). **3)** A new learning process to leverage our new mixing data augmentation. We show mixed samples can be used to optimize an additional supervised objective that significantly improves classifier performance.

II. SCIMIX FRAMEWORK

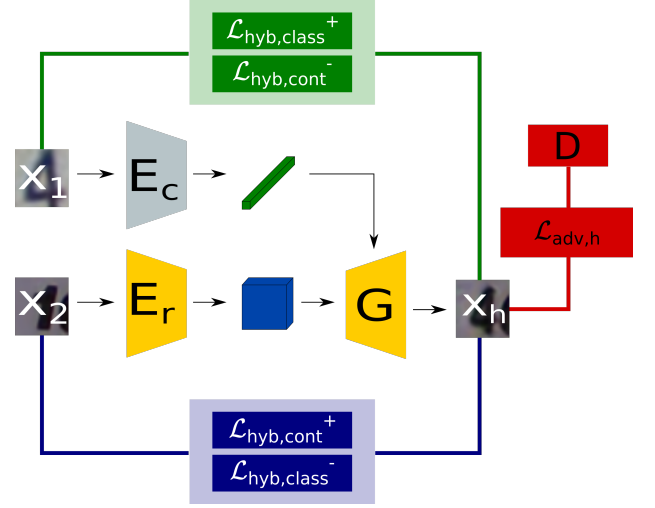
We propose in this paper a new mixing data augmentation that mixes the semantic content of a sample with the non-semantic content of another. As such, we first detail our scheme to train a model capable of mixing samples as per our exact specifications (Sec. II-A). We then explain our data augmentation strategy for visual classification leveraging our generated semantic hybrid samples (Sec. II-B). Finally, we discuss SciMix in the broader context of content mixing techniques (Sec. II-C).

A. Learning to generate hybrids

a) Auto-encoding architecture: Our framework is based on a novel auto-encoder architecture presented in Fig. 2a that treats semantic information as a global characteristic of



(a) Semantic auto-encoder of SciMix's generator. x is encoded into z_c (made semantic by \mathcal{L}_S) and z_r , which are decoded into \hat{x} optimized by $\mathcal{L}_{AE} = \mathcal{L}_{rec} + \mathcal{L}_{adv,r}$.



(b) Hybridization in SciMix's generator. Semantic components are extracted from x_1 and x_2 to obtain a hybrid optimized through \mathcal{L}_{hyb} (detailed in 5 sub-losses here).

Fig. 2. Overview of the SciMix generator architecture. SciMix trains an auto-encoder with two latent spaces, one of which is semantic

an encoded image. An input x is projected into a semantic latent space z_c by an encoder E_c as well as a complementary non-semantic latent space z_r by an encoder E_r . While our framework should primarily be understood as an auto-encoder framework, the distinct nature of the latent spaces z_c and z_r requires more careful consideration. We choose in this paper to focus on the definition and exploitation of the semantic features z_c , and simply treat z_r as information irrelevant to z_c . In other words, we design the framework so that the semantic information z_c controls what the generator G reconstructs. As the notion of semantic information is fundamentally tied to that of the tasks under consideration, we define z_c with respects to a classifier. More precisely, we treat the semantic latent space z_c as the feature space of a classifier. To this end, we add a linear neural layer C on top of z_c (see Fig. 2a) that outputs a class prediction \hat{y} . Note that $E_c \circ C$ therefore constitutes a standard CNN classifier [1], [13]. We ensure the classifier $E_c \circ C$ correctly learns semantic information through classification loss term \mathcal{L}_S on its output ¹.

Contrarily to [16], we complete the encoding with a generator G that computes a reconstruction \hat{x} using the non-semantic features z_r as direct inputs and the semantic features z_c as style codes. This makes the non-semantic content z_r easier to transfer, which is fortunate considering we can ensure semantic transfer more easily through a classifier $E_c \circ C$ (see Sec. III-E for the converse approach). As we treat the model as an autoencoder, we use a reconstruction loss \mathcal{L}_{rec} to tie the reconstruction \hat{x} to the input x . We further refine this reconstruction by using an adversarial critic to smooth out details [16]. We add a discriminator network D (see Fig. 2a)

to predict whether the considered image is a real sample or a reconstruction. Conversely, E_r , E_c and G are trained to fool D into seeing reconstructions \hat{x} as real images. The resulting loss $\mathcal{L}_{adv,r}$ serves to improve reconstructions learned through the reconstruction loss.

$$\mathcal{L}_{AE} = \mathcal{L}_{rec} + \mathcal{L}_{adv,r} = \sum_{x \in \mathcal{D}} \|x - G(E_c(x), E_r(x))\|_2 + \sum_{x \in \mathcal{D}} -\log(D(G(E_c(x), E_r(x))))). \quad (1)$$

Finally, our learning scheme is based on the minimization of the loss \mathcal{L}_{gen} composed of \mathcal{L}_S and \mathcal{L}_{AE} , plus an additional loss \mathcal{L}_{hyb} term:

$$\mathcal{L}_{gen} = \mathcal{L}_{AE} + \mathcal{L}_S + \mathcal{L}_{hyb}, \quad (2)$$

which is described in the following hybridizing scheme.

b) Hybridization losses: Simply training the auto-encoder architecture, even with z_c as the feature space of a classifier, is not enough to ensure that hybrids correctly inherit characteristics from their parents. To force the model to properly inject semantic content into the general background of known samples, we design explicit hybridization losses (studied more closely in Sec. III-E)

$$\mathcal{L}_{hyb} = \mathcal{L}_{hyb,class}^+ + \mathcal{L}_{hyb,cont}^+ + \mathcal{L}_{hyb,class}^- + \mathcal{L}_{hyb,cont}^- + \mathcal{L}_{adv,h}. \quad (3)$$

$\mathcal{L}_{hyb,class}^+$ explicitly trains our model to rely on z_c to generate the main semantic object in the generated reconstruction/hybrid. Indeed, we rely on the classifier $E_c \circ C$'s ability to identify and classify the main object in inputs (see Fig. 2b).

Put plainly, we generate hybrids $x_h = G(E_c(x_1), E_r(x_2))$ from pairs of samples in a batch and obtain logits predictions

¹ \mathcal{L}_S can correspond to any classifier training framework (e.g. supervised training, FixMatch [14]). In semi-supervised experiments, we use Mean Teacher [15] as a simple classifier guide in order to leverage SSL datasets.

$C(E_c(x_h))$ for those hybrids. $\mathcal{L}_{hyb,class}^+$ optimizes the model so that this prediction on the logits match the prediction on the semantic parent x_1 of x_h . Importantly, we only optimize the hybridization process that generates x_h : we do not optimize the classifier’s prediction on x_h or x_1 . The idea is that the autoencoder learns to place x_h in the right class manifold, while said class manifold does not move to accommodate x_h :

$$\mathcal{L}_{hyb,class}^+ = \sum_{x \in \mathcal{D}} \|C(E_c(x_1)) - C(E_c(G(E_c(x_1), E_r(x_2))))\|_2. \quad (4)$$

Similarly, $\mathcal{L}_{hyb,cont}^+$ optimizes the model so that a generated hybrids x_h ’s non-semantic representation $z_{r,h}$ matches its non semantic parent x_2 ’s non semantic component $z_{r,2}$. As in the semantic case, we only optimize the generative process that leads to the generation of x_h but do not optimize E_r to project x_h close to its non-semantic parent:

$$\mathcal{L}_{hyb,cont}^+ = \sum_{x \in \mathcal{D}} \|E_r(x_2) - E_r(G(E_c(x_1), E_r(x_2)))\|_2. \quad (5)$$

We also train hybrids to differ from their parents through the negative semantic hybridization loss $\mathcal{L}_{hyb,class}^- = \sum_{x \in \mathcal{D}} \|C(E_c(x_2)) - C(E_c(G(E_c(x_1), E_r(x_2))))\|_2$ and the negative non-semantic hybridization loss $\mathcal{L}_{hyb,cont}^- = \sum_{x \in \mathcal{D}} \|E_r(x_1) - E_r(G(E_c(x_1), E_c(x_2)))\|_2$. In practice, this means maximizing the distance between hybrids and semantic (resp. non-semantic) parent in non-semantic (resp. semantic) space.

To ensure the quality of generated hybrids, we train the discriminator D to also recognize hybrids as synthetic images. With this discriminator we can simply add an adversarial loss term $\mathcal{L}_{adv,h}$ to ensure hybrids look realistic (as far as D is concerned).

B. Training a classifier by leveraging our Data Augmentation

We now have a novel mixing data augmentation that can embed the semantic content of one sample to the non-semantic context of other samples, given a trained generator. This provides a useful and new way to improve any standard training method “X” by adding a single additional loss term $\mathcal{L}_{contradict}$: $\mathcal{L}_{SciMix} = \mathcal{L}_X + \mathcal{L}_{contradict}$.

a) Generating hybrids given a trained autoencoder:

Generating hybrids given a trained model is straightforward (Fig. 2b shows how a hybrid is mixed). Specifically, given samples $x^{(1)}$ (with known label $y^{(1)}$) and $x^{(2)}$, we extract the relevant features $z_c^{(1)} = E_c(x^{(1)})$, $z_r^{(1)} = E_r(x^{(1)})$, $z_c^{(2)} = E_c(x^{(2)})$ and $z_r^{(2)} = E_r(x^{(2)})$. $x_h = G(z_c^{(1)}, z_r^{(2)})$ is now a sample with class $y^{(1)}$. As a conservative measure, we only keep the generated hybrid if $C(E_c(x_h)) = y^{(1)}$ to avoid disturbing decision boundaries too much. Note that with this, we generate a strong augmentation of x_1 and teach the classifier to group x_1 with its strongly augmented version in a similar line to work in contrastive representation learning [17].

b) *Training a new classifier f* : We now propose a way to leverage our novel hybrids to improve the training of standard models such as Mean Teacher [15] or FixMatch [14]. To this end, we compute hybrids that mix the semantic content of each sample in the batch with the non-semantic content of other samples in the batch. We leverage those hybrids by optimizing an additional loss:

$$\mathcal{L}_{contradict} = \sum_{x_c, x_r \in \mathcal{B}, perm(\mathcal{B})} [l_{MSE}(f(x_h), \alpha * f(x_c) + (1 - \alpha)f(x_r))]. \quad (6)$$

This new loss takes advantage of our mixing paradigm by mostly imputing the semantic parent’s label to our hybrids, with only a slight dependence on the non-semantic parent to acknowledge the imperfection of the mixing process. Contrarily to standard mixing augmentations, the ratio $\alpha > 0.5$ is a fixed hyperparameter (in the spirit of label smoothing [18]).

C. Mixing contents in the literature

Mixing of one type of content with another is more readily found in unsupervised image-to-image translation: models are trained to translate the content of one image to the “domain” of another, though these terms are rarely well defined. Interestingly, bi-modal auto-encoding architectures appear fairly early on in this literature [19], [20]. Incidentally, more recent works in few-shot translation [21] and unsupervised translation [22] have even started associating the domain (or class) information to a style code fed as input to a StyleGan inspired decoder. In a more supervised fashion, such methods have been used to combine textures (domain information) with structural information (image information) [16]. This line of work however is specifically tailored to image generation and fails to leverage information to from fully fledged classifier to learn complex semantic variations.

Style transfer actually tackles the issue of mixing different types of contents in a similar fashion to our framework. The main difference between such frameworks and our problem lies in the definition of the contents to mix. In style transfer, the distinction is made between a style code and a structure code while we seek to mix semantic and non-semantic content. As such, our work reprises feature map modulation mechanisms that have been proven to work in style transfer [23], [24], [25] but makes use of additional losses to ensure we mix semantic and non-semantic contents.

It is worth noting that our goal of generating images in a way such that semantic content can be modified independently of non semantic content echoes that of disentangled generation [26], [27], [28], [29]. Importantly however, we aim to modify only a single coarse attribute rather than separate a multitude of fine grained characteristics.

III. EXPERIMENTS

We demonstrate here how our SciMix data augmentation can be leveraged to improve training in low-label settings. To this end, we conduct extensive experiments on the semi-supervised problem on the CIFAR10 [30] and SVHN [31]

TABLE I
MIXING SAMPLES WITH SciMix AS A DATA AUGMENTATION IMPROVES THE PERFORMANCE OF MEAN TEACHER AND FIXMATCH. SIGNIFICANT ACCURACY (%) GAINS ARE OBSERVED ON CIFAR10 (WITH 100, 250 AND 500 LABELS) AND SVHN (WITH 60 AND 100 LABELS).

Method	CIFAR10			SVHN	
	100	250	500	60	100
Mean Teacher, [15]	40.5 \pm 6.4	63.1 \pm 0.9	72 \pm 3	48.7 \pm 23.0	82.3 \pm 5.5
SciMix w/ Mean Teacher	46.4 \pm 1.2	68.0 \pm 1	77.2 \pm 0.5	83.4 \pm .5	87.3 \pm 2.8

(a) Mean Teacher

Method	CIFAR10	SVHN
	100	60
FixMatch, [8]	88.6 \pm 0.7	96.4 \pm 0.3
SciMix w/ FixMatch	90.7 \pm 0.2	96.5 \pm 0.9

(b) FixMatch

datasets. Additionally, we propose in Sec. III-G a study of SciMix’s performance in a fully supervised setting with few labels on a variation of the CUB-200 [32] dataset.

In the semi-supervised case, we show SciMix improves two backbone methods: Mean Teacher [15] and FixMatch [8] (refer to Sec. II-B for how we apply our framework to these methods). We chose Mean Teacher as a reference consistency-based baseline. Beyond its widespread use in SSL, consistency induces a stabilization we feel would help extract invariant semantic features. FixMatch is a state-of-the-art SSL method based on strong augmentation, and often serves as a reference or backbone in the literature [33], [34].

We operate on a standard WideResNet-28-2 [13] for our classifiers (both f and $E_c \circ C$). E_r follows the same architecture as E_c . The skeleton of G follows a StyleGanv2 [35] architecture. Hyperparameters and optimizers were generally taken to follow settings reported in the base methods’ original papers [15], [8].

We report the *mean* \pm *std* classification accuracy over 3 seeded runs for varying numbers of labeled samples in a dataset (the rest are treated as unlabeled). The SciMix generators used to train a model with N labeled samples are also trained with only N labeled samples. One generator is trained per setting, and classifiers trained with the generator’s mixed samples are trained on the same split of labeled/unlabeled data to avoid information leakage. More details for all experiments are given in Appendix.

A. Performance gains

Tab. I shows that adding our optimization on hybrids with $\mathcal{L}_{contradict}$ - as described in Sec. II-B - does indeed lead to improved performance on CIFAR10 and SVHN. Indeed, training with SciMix hybrids leads to significant accuracy gains with a Mean Teacher classifier with a wide range of labeled samples. We also observe improvements over FixMatch on very low labeled settings (FixMatch performance quickly saturates on higher labeled settings). Interestingly, two concurrent behaviors can be observed on Mean Teacher: SciMix hybrids become more useful when less labeled samples are available but the quality of generated hybrids becomes unreliable with few labeled samples. Indeed, at 100 labeled samples on CIFAR10, the Mean Teacher classifier used to train the generator is too weak to provide very useful hybrids. With a strong SciMix hybridizer (trained with all samples), SciMix

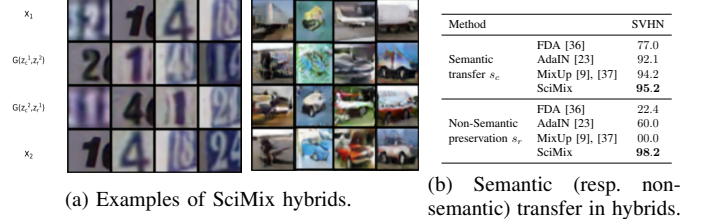


Fig. 3. SciMix hybrids properly mix semantic and non-semantic contents.

data augmentation would bring a Mean Teacher classifier to an accuracy of 60.4 ± 1.7 with 100 labels on CIFAR10.

B. Quality of hybridization

We now verify the auto-encoder described in Sec. II-A indeed yields interesting hybrids for data augmentation. This can be observed qualitatively by considering the generated hybrids (see Fig. 3a) and quantitatively in Tab. 3b through an observation of the generated hybrids in relation to their parents.

Fig. 3a shows hybrids generated by the method. As can be observed, SciMix creates hybrids that match the semantic content of the semantic parent while closely matching the non-semantic parents in every other regard. Qualitative samples suggest the framework properly identifies semantic content in the parent samples and successfully embeds the semantic content of the semantic parent into the non-semantic background of the non-semantic parent.

a) Preservation of semantic and non-semantic characteristics: We quantify how well the generated hybrids x_h inherit properties from the semantic and non-semantic parents x_c and x_r through the metrics s_c and s_r . The semantic transfer rate s_c is the accuracy of an “oracle” classifier (trained on the entire dataset, as a proxy for human evaluation of hybrid labels) over a dataset built from hybrids that are assumed to have inherited the label of their semantic parent. Conversely, the non-semantic preservation rate s_r is the proportion of hybrids x_h that are closer to the non-semantic parent x_r in pixel space (ie, $\|x_h - x_c\| > \|x_h - x_r\|$).

Tab. 3b shows SciMix compares favorably to texture altering hybrids (FDA, AdaIN) and standard mixing augmentations (MixUp with the label of the dominant samples as suggested in [37]) on a strong generator (SVHN 250 labels). While most existing hybridizations do tend to preserve the semantic content, SciMix shines in that it transfers semantic content

while keeping hybrids very close to their non-semantic parent. Indeed, other hybrids remain much closer to their semantic parent (always the case - by design - for MixUp).

C. Comparison to Mixing Data Augmentation

TABLE II
COMPARISON OF SciMix WITH OTHER DATA AUGMENTATIONS IN LOW LABEL SETTINGS.

Method	CIFAR10	SVHN
	250	60
Mean Teacher, [15]	63.1 ± 0.9	48.7 ± 23.0
Mean Teacher + MixUp [9]	64.8 ± 3.5	61.8 ± 1.0
Mean Teacher + CutMix [10]	55.0 ± 7.5	17.3 ± 3.7
SciMix w/ Mean Teacher (ours)	68 ± 1.0	83.4 ± 0.5

We now show that in very low label settings, the “artificial” labeled samples SciMix can outperform the regularization offered by more traditional mixing Data Augmentation. Tab. II shows that SciMix mostly outperforms MixUp and CutMix for CIFAR10 with 250 labeled samples (the generator is too weak with 100 labels) and SVHN with 60 labeled samples (hardest setting). While MixUp does perform similarly to SciMix on CIFAR10, this is likely due to the low performance of the classifier $E_c \circ C$ trained with only 250 labels.

D. Ablation: Regularization vs. Data Augmentation

TABLE III
COMPARISON ON THE PERFORMANCES OF THE CLASSIFIER TRAINED WITH OUR GENERATOR (SEC. II-A) AND MODELS TRAINED FROM SCRATCH WITHOUT DATA AUGMENTATION (SEC. II-B).

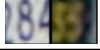

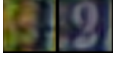
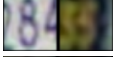
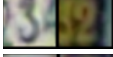
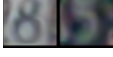
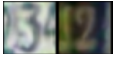
Method	SVHN
	60
Mean Teacher, [15]	48.7 ± 23.0
Generator classifier (Sec. II-A)	72.0 ± 2.7
SciMix w/ Gen. Init.	72.6 ± 2.3
SciMix w/ Mean Teacher (Sec. II-B)	$83.4 \pm .5$

While we evaluate our framework as a general data augmentation framework by using mixed samples from a pre-trained generator to train a model from scratch, it is interesting to note that our mixing auto-encoder also trains and uses a classifier model. We study in Tab. III the relative performance of the generator classifier (tied to the auto-encoder and regularized by \mathcal{L}_{hyb} but without mixed data augmentation) and of a classifier trained from scratch with our mixed data augmentation (but without explicit loss regularization). While the generator classifier does outperform the backbone classifier in all cases (showcasing the usefulness of the regularization scheme), training from scratch with data augmentation leads to better results on every setting. Interestingly, keeping the weights of the auto-encoder classifier as initialization to train a model with SciMix mixing does not necessarily outperform a

random initialization. This can be explained by the fact that the separation characterized by the data augmentation is already learned by the auto-encoder classifier and therefore training with our hybrids fails to teach the model anything interesting.

E. Model analysis: Importance of the learning scheme

TABLE IV
COMPARISON OF VARIOUS ARCHITECTURAL VARIANTS (SEC. II-A) ALONG WITH SAMPLES OF GENERATED HYBRIDS FOR EACH VARIANT ON SVHN 60 LABELS.

Hybrids parent pairs x_r/x_c					
Method	Accuracy	s_c	s_r	sample hybrids	
Structural z_c	39.5 ± 9.3	16.0	56.0		
No \mathcal{L}_{hyb}	48.5 ± 14.2	11.7	99.8		
Basic \mathcal{L}_{hyb}	66.3 ± 1.0	66.2	99.4		
Non Frozen criterion \mathcal{L}_{hyb}	81.8 ± 3.0	75.1	73.8		
Full Setup	83.4 ± 0.5	76.7	98.8		

To better explore how our framework facilitates the incrustation of semantic content in the general context of existing samples, we first propose a rapid ablation study on the quality of the samples generated by variants of our auto-encoder on our hardest SVHN setting (60 labels). We evaluate this through the classification accuracy of models trained with the generator’s mixed samples, the semantic transfer s_c , the non-semantic transfer s_r , and a visual evaluation of two hybrids (e.g. the leftmost hybrid mixes a blue 8 x_c with a yellow 3 x_r).

We consider 4 variations on SciMix’s generator to demonstrate the merits of our chosen method: z_r as a style code. We flip the roles of z_c and z_r , to demonstrate z_c is better used as a style code in SciMix. **No \mathcal{L}_{hyb}** . We train without an explicit optimization loss, to show \mathcal{L}_S and \mathcal{L}_{rec} are not sufficient to create good hybrids in SciMix. **Basic \mathcal{L}_{hyb}** . We demonstrate the orthogonalization losses $\mathcal{L}_{hyb,class}^-$ and $\mathcal{L}_{hyb,cont}^-$ contribute to the generator by considering a variant that does not optimize them. **No frozen criterion \mathcal{L}_{hyb}** . We do not force the generator to only optimize the generation of the hybrids when optimizing \mathcal{L}_{hyb} (the projection of the hybrids in the latent spaces is also modified).

Tab. IV shows that without explicit hybridization optimization, the model fails to properly transfer semantic characteristics (low s_c score). Since the model does properly transfer semantic content with only $\mathcal{L}_{hybrids}^+$, the addition of the orthogonalization constraints $\mathcal{L}_{hybrids}^-$ is not necessary to obtain useful hybrids. However, this orthogonalization increases the diversity in the generated hybrids and therefore leads to a better augmentation procedure. Not freezing the projection heads when training for hybridization on the other hand leads

TABLE V
COMPARISON OF VARIOUS LOSSES TO LEVERAGE HYBRIDS IN \mathcal{L}_{SciMix}
(SEC. II-B) ON SVHN 60 LABELS.

Method	Accuracy
Baseline	48.7 ± 23.0
$\mathcal{L}_{labeled}$	29.9 ± 17.6
$\mathcal{L}_{pseudo-label}$	62.8 ± 5.4
$\mathcal{L}_{consistency}$	77.8 ± 4.8
$\mathcal{L}_{contradict}$	83.4 ± 0.4

to a general deterioration of the training process and can therefore be felt in both transfer rates. Predictably, using z_r as a global style code leads a very poor correspondence of hybrids to their non-semantic parents as things like backgrounds can become very complicated to reproduce with a modulation based generator.

F. Model analysis: Leveraging the hybrids as Data Augmentation

We considered 4 alternative methods to exploit the hybrids generated by our method: **Labeled** Only hybrids with a labeled semantic parent are considered (supervised training with a hard label) **Pseudo-label** Hybrids are treated as labeled samples (hard labels), with the labels inherited from the semantic parent’s pseudo-labels. **Consistency** Hybrids are made to follow their semantic parent’s prediction. **Contradict** Hybrids are made to match both their semantic parent’s consistency target and their non-semantic parent targets.

Tab. V shows that the best results are obtained with $\mathcal{L}_{contradict}$, but $\mathcal{L}_{pseudo-label}$ and \mathcal{L}_{cons} also outperform the baseline method on SVHN 60 labels (hardest setting). The fact $\mathcal{L}_{contradict}$ outperforms other methods is interesting in that the loss does not actually treat the hybrids as pure labeled samples, but assumes some semantic content/noise is retained from the non-semantic samples. $\mathcal{L}_{labeled}$ ’s failure suggests that even with proper mixing, it not possible to improve models by only generalizing a few labeled samples.

G. Pushing SciMix on CUB-200

We now push SciMix further by studying versions of the more complex **Caltech-UCSD Birds 200** (CUB-200) dataset [32] (6033 pictures of 200 bird species). Given CUB-200 inherently presents few labels, we directly study how fully supervised training benefits from SciMix on low labels settings (\mathcal{L}_S is a standard cross-entropy loss). Furthermore, we take advantage of CUB-200’s higher native resolution to go beyond the limitations of 32×32 images in CIFAR 10 and SVHN: we use SciMix on commonly studied input sizes 64×64 (e.g. Tiny ImageNet [38]) and 96×96 (e.g. STL-10 [39]).

a) *Quality of generated hybrids:* As can be observed on Fig. 4a, the SciMix autoencoder learns to generate interesting hybrids for different resolutions of the CUB-200 dataset. While the lower resolution 64×64 hybrids inherit more semantic characteristics of the relevant parent, 96×96 hybrids still retain semantic patterns tied to the semantic parent’s class.

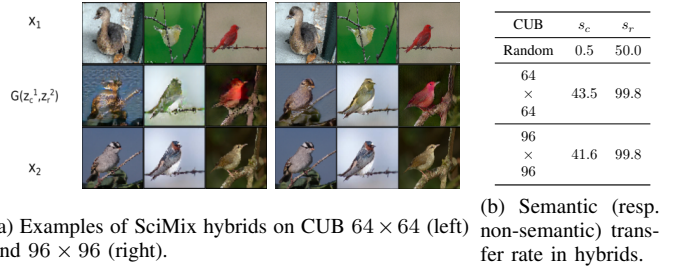


Fig. 4. SciMix hybrids mix semantic and non-semantic contents on CUB-200.

TABLE VI
IMPACT OF SciMix DATA AUGMENTATION FOR THE CUB-200 DATASET
AT RESOLUTIONS 64×64 AND 96×96 .

Method	CUB-200	
	64×64	96×96
Supervised	58.9 ± 1.0	65.2 ± 0.8
SciMix w/ Supervised	60.2 ± 0.6	65.6 ± 0.9

Interestingly, SciMix has no difficulty producing hybrids close to their non-semantic parent as can be shown in Tab. 4b by reprising the analysis of the non-semantic transfer rate s_r from Sec. III-B. Analyzing the semantic transfer rate s_c proves more difficult as our best “oracle” classifiers remain unreliable (around 60% accuracy). Nevertheless, the semantic transfer rates s_c in Tab. 4b indicate hybrids generated by SciMix are properly classified by the oracle classifier as having inherited their semantic parent’s class about 40% of the time (orders of magnitude more than attributable to random chance).

b) *Performance gains:* Tab. VI shows a fully supervised version of SciMix data augmentation improves supervised models on both 64×64 and 96×96 versions of CUB-200. This demonstrates that while SciMix augmentation strongly benefits from a large amount of unlabeled data, it can still generate hybrids diverse enough to benefit training with only a small set of data to generate hybrids from.

IV. CONCLUSION

In conclusion, we propose in this paper a new approach to mixing data augmentation that mixes the semantic content of one sample and the non-semantic content of the other. Making use of advances in style transfer and modular image generation, we train an auto-encoder that learns to extract and combine semantic and non-semantic content from multiple images. Through extensive experiments, we show the intricate hybridization loss we propose leads to the generation of interesting mixed samples. We furthermore demonstrate it is better to use semantic information as an indirect “style code” input to our StyleGan decoder instead of non-semantic information. Afterwards, we prove such data augmentation significantly improves the training of supervised and semi-supervised models when few labels are available.

REFERENCES

- [1] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision*, 2016.
- [2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012.
- [3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015.
- [4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.
- [5] Z. He, L. Xie, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Data augmentation revisited: Rethinking the distribution gap between clean and augmented data," *Arxiv preprint*, 2019.
- [6] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," *Computer Vision and Pattern Recognition*, 2019.
- [7] O. Chapelle, B. Scholkopf, and A. Zien, *Semi-Supervised Learning*. The MIT Press, 2006.
- [8] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *Advances in Neural Information Processing Systems*, 2019.
- [9] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.
- [10] Y. Yang and S. Soatto, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *International Conference on Computer Vision*, 2019.
- [11] L. Carratino, M. Cissé, R. Jenatton, and J.-P. Vert, "On mixup regularization," in *ArXiv preprint*, 2020.
- [12] S. Thulasidasan, G. Chennupati, J. A. Bilmes, T. Bhattacharya, and S. Michalak, "On mixup training: Improved calibration and predictive uncertainty for deep neural networks," in *Advances in Neural Information Processing Systems*, 2019.
- [13] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *British Machine Vision Conference*, 2016.
- [14] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in Neural Information Processing Systems*, 2020.
- [15] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems*, 2017.
- [16] T. Park, J.-Y. Zhu, O. Wang, J. Lu, E. Shechtman, A. Efros, and R. Zhang, "Swapping autoencoder for deep image manipulation," in *Advances in Neural Information Processing Systems*, 2020.
- [17] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Computer Vision and Pattern Recognition*, 2020.
- [18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Computer Vision and Pattern Recognition*, 2016.
- [19] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems*, 2017.
- [20] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *European Conference on Computer Vision*, September 2018.
- [21] M.-Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, and J. Kautz, "Few-shot unsupervised image-to-image translation," in *International Conference on Computer Vision*, 2019.
- [22] K. Baek, Y. Choi, Y. Uh, J. Yoo, and H. Shim, "Rethinking the truly unsupervised image-to-image translation," in *International Conference on Computer Vision*, 2021.
- [23] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *International Conference on Computer Vision*, 2017.
- [24] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Computer Vision and Pattern Recognition*, 2019.
- [25] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," in *Advances in Neural Information Processing Systems*, 2021.
- [26] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "Beta-vae: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2017.
- [27] T. Robert, N. Thome, and M. Cord, "Dualdis: Dual-branch disentangling with adversarial learning," *Arxiv preprint*, 2019.
- [28] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Advances in Neural Information Processing Systems*, 2016.
- [29] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Computer Vision and Pattern Recognition*, 2017.
- [30] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *Tech. Rep.*, 2009.
- [31] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *Advances in Neural Information Processing Systems*, 2011.
- [32] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-UCSD Birds 200," *Tech. Rep.*, 2010.
- [33] J. Li, C. Xiong, and S. C. Hoi, "Comatch: Semi-supervised learning with contrastive graph regularization," in *International Conference on Computer Vision*, 2021.
- [34] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," in *Advances in Neural Information Processing Systems*, 2021.
- [35] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Computer Vision and Pattern Recognition*, June 2020.
- [36] Y. Yang and S. Soatto, "Fda: Fourier domain adaptation for semantic segmentation," in *Computer Vision and Pattern Recognition*, 2020.
- [37] H. Inoue, "Data augmentation by pairing samples for images classification," *Arxiv preprint*, 2018.
- [38] P. Chrabaszcz, I. Loshchilov, and F. Hutter, "A downsampled variant of imagenet as an alternative to the cifar datasets," *Arxiv preprint*, 2017.
- [39] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *International Conference on Artificial Intelligence and Statistics*, 2011.