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.
Abstract—Deep architectures 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 labelisation 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 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. SCI Mix 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
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 through a classifier $C$. 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. 2b) 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 $L_S$ on its output.

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. IIIB-C for the converse approach). As we treat the model as an autoencoder, we use a reconstruction loss $L_{rec}$ 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. 2b) 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 $L_S$ on its output.

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. IIIB-C for the converse approach). As we treat the model as an autoencoder, we use a reconstruction loss $L_{rec}$ 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. 2b) 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 $L_S$ on its output.

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. IIIB-C for the converse approach). As we treat the model as an autoencoder, we use a reconstruction loss $L_{rec}$ 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. 2b) 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 $L_S$ on its output.

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. IIIB-C for the converse approach). As we treat the model as an autoencoder, we use a reconstruction loss $L_{rec}$ 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. 2b) 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 $L_S$ on its output.

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. IIIB-C for the converse approach). As we treat the model as an autoencoder, we use a reconstruction loss $L_{rec}$ 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. 2b) 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 $L_S$ on its output.

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. IIIB-C for the converse approach). As we treat the model as an autoencoder, we use a reconstruction loss $L_{rec}$ 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. 2b) 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 $L_S$ on its output. 
the relevant features $C(E_c(x_h))$ for those hybrids. $L^+_{hyb, class}$ 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 $x_h$:

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

Similarly, $L^+_{hyb, cont}$ 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:

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

We also train hybrids to differ from their parents through the negative semantic hybridization loss $L^+_{hyb, class} = \sum_{x \in D} ||C(E_c(x_2)) - C(E_c(G(E_c(x_1), E_r(x_2))))||_2$ and the negative non-semantic hybridization loss $L^+_{hyb, cont} = \sum_{x \in D} ||E_r(x_1) - E_r(G(E_c(x_1), E_r(x_2)))||_2$. 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 $L_{adv,h}$ 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 new additional loss term $L_{contradict}$: $L_{SciMix} = L_X + L_{contradict}$. 

#### 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 $z^{(1)}_c = E_c(x^{(1)})$, $z^{(2)}_c = E_c(x^{(2)})$ and $z^{(2)}_r = E_r(x^{(2)})$. $x_h = G(z^{(1)}_c, z^{(2)}_c)$ 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:

$$L_{contradict} = \sum_{x \in \mathbb{B} \text{perm}(\mathbb{B})} [\text{MSE}(f(x_h), \alpha \ast f(x_c) + (1 - \alpha)f(x_c))].$$  

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]
classifiers (both for backbone in the literature [33], [34]. Semantic features. FixMatch is a state-of-the-art SSL method based on consistency methods). We chose Mean Teacher as a reference consistency method to avoid information leakage. More details for all experiments are given in Appendix.

We operate on a standard WideResNet-28-2 [13] for our classifiers (both $f$ and $E_c \circ \phi$), $E_r$ follows the same architecture as $E_c$. The skeleton of $G$ follows a StyleGan2 architecture. Hyperparameters and optimizers were generally taken to follow settings reported in the base methods’ original papers [15], [8].

We report the mean ± 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. 1 shows that adding our optimization on hybrids with $L_{contradict}$ - 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 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 $x_s$ 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 (i.e., $||x_h - x_r|| > ||x_h - x_s||$).

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.

### Table I

<table>
<thead>
<tr>
<th>Method</th>
<th>CIFAR10</th>
<th>SVHN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>250</td>
</tr>
<tr>
<td>Mean Teacher [15]</td>
<td>40.5 ± 6.4</td>
<td>63.1 ± 0.9</td>
</tr>
<tr>
<td>SciMix w/ Mean Teacher</td>
<td>46.4 ± 1.2</td>
<td>68.0 ± 1</td>
</tr>
<tr>
<td>FixMatch [6]</td>
<td>88.6 ± 0.7</td>
<td>96.4 ± 0.3</td>
</tr>
<tr>
<td>SciMix w/ FixMatch</td>
<td>90.7 ± 2</td>
<td>96.5 ± 0.9</td>
</tr>
</tbody>
</table>

(a) Mean Teacher

(b) FixMatch

### Fig. 3

SciMix hybrids properly mix semantic and non-semantic contents.
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

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

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 $L_{hyb}$. We train without an explicit optimization loss, to show $L_S$ and $L_{rec}$ are not sufficient to create good hybrids in SciMix. Basic $L_{hyb}$. We demonstrate the orthogonalization losses $L_{hyb, class}$ and $L_{hyb, cont}$ contribute to the generator by considering a variant that does not optimize them. No frozen criterion $L_{hyb}$. We do not force the generator to only optimize the generation of the hybrids when optimizing $L_{hyb}$ (the projection of the hybrids in the latent spaces is also modified).

To better explore how our framework facilitates the incorporation 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 $x_r$).

E. Model analysis: Importance of the learning scheme

We do

<table>
<thead>
<tr>
<th>Methods</th>
<th>SVHN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Teacher, [15]</td>
<td>48.7 ± 23.0</td>
</tr>
<tr>
<td>Generator classifier (Sec. II-A)</td>
<td>72.0 ± 2.7</td>
</tr>
<tr>
<td>SciMix w/ Gen. Init.</td>
<td>72.6 ± 2.3</td>
</tr>
<tr>
<td>SciMix w/ Mean Teacher (Sec. II-B)</td>
<td>83.4 ± 5.5</td>
</tr>
</tbody>
</table>

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.

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 $L_{hyb}$ 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.

### Table II

**Comparison of SciMix with other Data Augmentations in low label settings.**

<table>
<thead>
<tr>
<th>Method</th>
<th>CIFAR10</th>
<th>SVHN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Teacher, [15]</td>
<td>63.1 ± 0.9</td>
<td>48.7 ± 23.0</td>
</tr>
<tr>
<td>Mean Teacher + MixUp [9]</td>
<td>64.8 ± 3.5</td>
<td>61.8 ± 1.0</td>
</tr>
<tr>
<td>Mean Teacher + CutMix [10]</td>
<td>55.0 ± 7.5</td>
<td>17.3 ± 3.7</td>
</tr>
<tr>
<td>SciMix w/ Mean Teacher (ours)</td>
<td>68 ± 1.0</td>
<td>83.4 ± 0.5</td>
</tr>
</tbody>
</table>

### 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).**

<table>
<thead>
<tr>
<th>Method</th>
<th>SVHN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Teacher, [15]</td>
<td>48.7 ± 23.0</td>
</tr>
<tr>
<td>Generator classifier (Sec. II-A)</td>
<td>72.0 ± 2.7</td>
</tr>
<tr>
<td>SciMix w/ Gen. Init.</td>
<td>72.6 ± 2.3</td>
</tr>
<tr>
<td>SciMix w/ Mean Teacher (Sec. II-B)</td>
<td>83.4 ± 5.5</td>
</tr>
</tbody>
</table>
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 $\mathcal{L}_\text{contradict}$, but $\mathcal{L}_\text{pseudo-label}$ and $\mathcal{L}_\text{cons}$ also outperform the baseline method on SVHN 60 labels (hardest setting). The fact $\mathcal{L}_\text{contradict}$ 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. $\mathcal{L}_\text{labeled}$’s failure suggests that even with proper mixing, it is 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.

Interestingly, SciMix has no difficulty producing hybrids close to their non-semantic parent as can be shown in Tab. 4b by re-running 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.