Unbiased Manifold Augmentation for Coarse Class Subdivision

Baoming Yan, Ke Gao, Bo Gao, Lin Wang, Jiang Yang, and Xiaobo Li

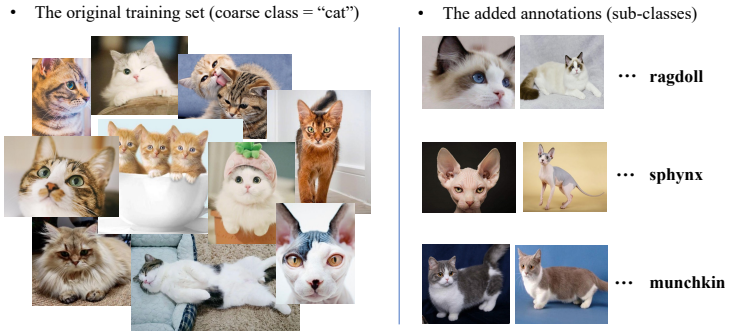
Alibaba Group
{andy.ybm, gaoke.gao, leo.gb, youlin.wl, yangjiang.yj, xiaobo.lixb}@alibaba-inc.com

Abstract. Coarse Class Subdivision (CCS) is important for many practical applications, where the training set originally annotated for a coarse class (e.g. bird) needs to further support its sub-classes recognition (e.g. swan, crow) with only very few fine-grained labeled samples. From the perspective of causal representation learning, these sub-classes inherit the same determinative factors of the coarse class, and their difference lies only in values. Therefore, to support the challenging CCS task with minimum fine-grained labeling cost, an ideal data augmentation method should generate abundant variants by manipulating these sub-class samples at the granularity of generating factors. For this goal, traditional data augmentation methods are far from sufficient. They often perform in highly-coupled image or feature space, thus can only simulate global geometric or photometric transformations. Leveraging the recent progress of factor-disentangled generators, Unbiased Manifold Augmentation (UMA) is proposed for CCS. With a controllable StyleGAN pretrained for a coarse class, an approximate unbiased augmentation is conducted on the factor-disentangled manifolds for each sub-class, revealing the unbiased mutual information between the target sub-class and its determinative factors. Extensive experiments have shown that in the case of small data learning (less than 1% fine-grained samples of commonly used), our UMA can achieve 10.37% average improvement compared with existing data augmentation methods. On challenging tasks with severe bias, the accuracy is improved by up to 16.79%. We release our code at https://github.com/leo-gb/UMA.

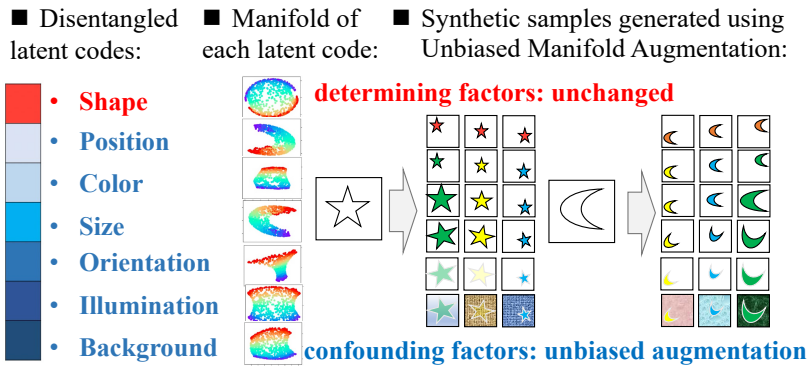
Keywords: Coarse Class Subdivision,Causal Representation Learning, Factor-disentangled Generator,Unbiased Manifold Augmentation

1 Introduction

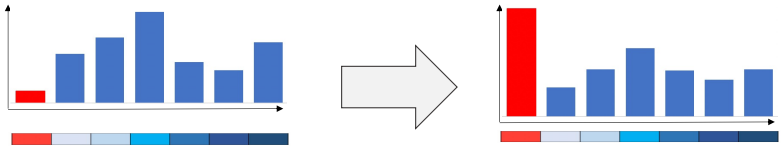
Different from the conventional classification tasks where the original class and the new target are of similar level of semantic granularity, this paper focuses on Coarse Class Subdivision (CCS) which is a very practical problem. Given an existing training set for a coarse class, the target of CCS is to further recognize its sub-classes with minimum fine-grained labeling cost, as shown in Figure 1 (a). From the perspective of causal representation learning [17,20,26], the generating



(a) The illustration of Coarse Class Subdivision task



(b) Sample synthesis using our UMA



(c) The change of factor importance for a target sub-class before/after using UMA

Fig. 1. The proposed Unbiased Manifold Augmentation (UMA) for coarse class subdivision task. (a) An example of task overview: given a training set for coarse class, and only very few samples for each sub-class. (b) Abundant and unbiased samples generation using our UMA. It should be noted that the semantic of each latent-code manifold is implicit. (c) The problem caused by limited and biased fine-grained data, and the effect of our UMA.

factors of images are composed of determinative factors and confounding factors. In the case of CCS, the sub-classes inherit the same determinative factors of the coarse class, and their difference lies only in values. This task requires identifying the determinative factors of these sub-classes and distinguishing subtle differences between them. Fine-grained classification [10, 16, 17] is very challenging in itself, limited samples with agnostic bias make the task more difficult. For instance, given a training set with the coarse label "bird", a sub-class model fine-tuned with a few white-swan images often fails while discriminating between black swan and crow, revealing "color" as an unreliable determinative factor for the target sub-class.

Traditional data augmentation methods have been proved to be effective for common small data learning problems [27, 28], but show limited effect on this challenging task. Ideal augmentation strategies for CCS should make full use of the fine-grained samples, and manipulate them at the granularity of generating factors. Even if their distribution on the generating factors is biased, we should generate abundant variants based on them, revealing the unbiased mutual information between the target sub-classes and their generating factors. In this way, the classifiers can be guided to focus on the determining factors which have the best generalization performance, as shown in Figure 1 (b)(c). However, most existing data augmentation methods perform in highly-coupled image or feature space, thus can only simulate global geometric or photometric transformations, which is far from sufficient for this challenging task.

Fortunately, the rapid development of factors-disentangled and controllable generative models illuminates an entirely new avenue for overcoming this problem [18,23]. To support various attribute-level manipulation of a given image, an idea controllable generative network is forced to learn disentangled manifolds for all generating factors of the target class, thus can model any variations of these factors with well-structured latent representations. Consequently, different from the traditional data augmentation methods conducted in highly-coupled image space or feature space, progressive manipulation in the disentangled manifolds can lead to an approximate unbiased distribution of all generating factors for the target category.

However, the factor-decoupling and controllable-manipulation effects of existing generators are far from ideal, especially the correspondence between editable latent-codes and the generating factors are implicit. Despite there are many methods aiming to find out these relations [8,15], most of them are too costly to be a practical option. In this paper, we propose a novel method called Unbiased Manifold Augmentation method (UMA), which is a simple and effective solution for the above difficult problems.

It should be noted that although the editable generator is essential for our UMA, there are many off-the-shelf generic generators can be utilized directly. For example, a generic face generator trained on the well-known datasets FFHQ [3] or CelebA [1]can be used in our UMA for any facial attribute recognition task. And any general bird generator trained on LSUN [25] or [2] can be used to improve the recognition of "swan" or "gull".

Our main contributions are as follows:

- A novel and systematic data augmentation mechanism called Unbiased Manifold Augmentation (UMA) is proposed for coarse class subdivision problem. Given an existing training set for a coarse class, our UMA can support subclasses recognition with minimum fine-grained labeling cost.
- The UMA is conducted on latent-code manifolds of a controllable generator pretrained for a coarser category, instead of the traditional highly-coupled image or feature space. Using a simple and effective progressive synthesis strategy, an approximate unbiased augmentation at the granularity of generating factors is achieved, even with limited labeled samples and agnostic bias. By revealing the unbiased mutual information between the target class and all of its impact factors, the classifier can be guided to focus on the right determining factors of the target sub-classes. In conjunction with it, a phase of progressive robust learning is further integrated, to keep a good balance of the diversity and reliability of these synthetic samples.
- Extensive experiments have shown that in the case of small data learning (less than 1% fine-grained samples of commonly used), our UMA can achieve 10.37% average improvement compared with existing data augmentation methods. On challenging tasks with severe bias, the accuracy is improved by up to 16.79%.

2 Related work

Although the coarse class subdivision task has rarely been formally defined, it is a very practical problem and the related research is ubiquitous in deep learning. An exhaustive list of these work is out of the scope of this paper. Here we highlight data-augmentation strategies, especially based on the exploration of the recent progress of controllable generators with disentangled manifolds.

Data Augmentation. As one of the most hopeful means to alleviate small data learning problem in CCS task, the data augmentation methods have been actively studied [5,6,9,13,24,27,28,30]. Considering the different spaces in which the operations are performed, most of the widely used methods can be roughly divided into three categories. 1) Image Augmentation: given two training samples, Mixup [28] interpolates both the image and labels, while CutMix [27] partially mixes the patches and labels. The approaches conducted in the original image space [5,27,28] can improve the DNNs' robustness over some common noise, but the effects are often limited due to these simple variants of global geometric or photometric. 2) Feature Augmentation: Some methods such as [10,24] regularizes the DNNs by random interpolation of feature maps. It should be noted that the manifolds mentioned in Manifold Mixup [24] are actually the traditional feature maps in classification networks, not the manifolds of latent codes derived from controllable generative models, as adopted in this paper. 3) Style Augmentation: Since [7,9]decomposed the feature maps of the DNNs into separated representations of image content and style, many data augmentation methods using style or content manipulation has been proposed [8–10,14,29]. However, the diversity

and reasonableness of the above synthetic data are difficult to guarantee, because both the feature space and style space are still highly coupled in semantic. Different from them, Our UMA is conducted in a well-structured latent space of disentangled generating factors, leveraging the recent progress of controllable generative models.

Controllable Generative Models. In recent years, controllable generators have witnessed great progress. It has been demonstrated that StyleGAN and StyleGAN2 [4] can offer strong editing capabilities with a disentangled latent space. Motivated by it, various methods have used StyleGAN to perform some specific manipulation for any given image [18,19,23]. They first encode an image into the latent space W of a pre-trained StyleGAN, then edit the latent code in a semantically meaningful way to obtain a new code, according to different image editing requirements. A desired image is then generated with this new code using the StyleGAN. An extended latent space W+ as the concatenation of 18 different W vectors is often used, one for each input layer of StyleGAN [18,23]. The latent space is considered as an ideal model of natural images' inherent distribution [4,18,19,23]. Sample synthesis in these disentangled semantic manifolds can lead to an approximate unbiased distribution of all generating factors of the target category, thus will guide the classifier to correctly focus on the determining factors which have the best generalization performance.

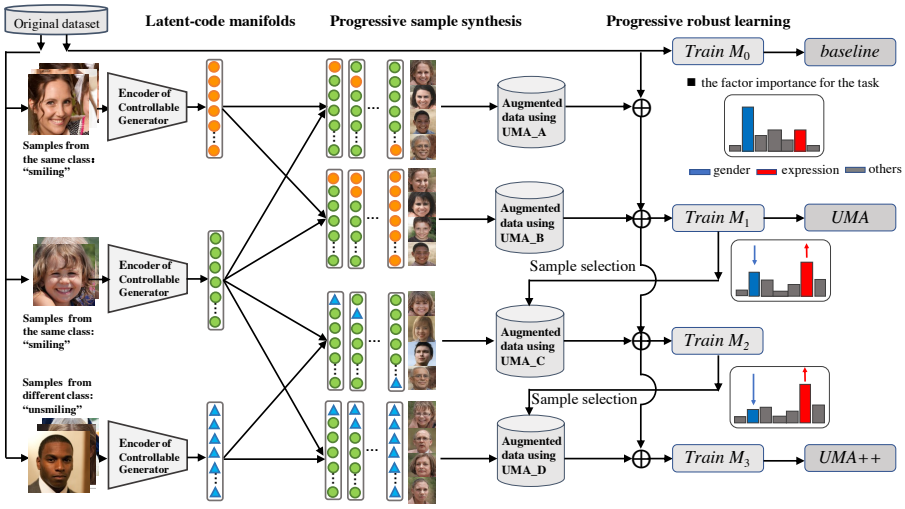


Fig. 2. The architecture of the proposed Unbiased Manifold Augmentation (UMA) for coarse class subdivision task. Take the coarse class "face" for example, here the target sub-classes are "smiling" and "unsmiling". Randomly selected fine-grained samples often have inevitable bias on the confounding factor "gender", instead of the determining factor "expression" for the target task. Our UMA can generate unbiased samples for each sub-class and lead the classifier to focus on determining factors with better generalization.

3 The UMA method

Although the advanced controllable generative models can support image manipulations at the granularity of generating factors, as for coarse class subdivision task, there are still three questions need to be answered:

1. How to synthetize abundant variants with only very few fine-grained samples?

2. How to synthetize unbiased variants without knowing what factors are the determinative ones?

3. How to guarantee that these synthetic variants are fine-grained-label preserving?

The key idea of our approach is based on an important observation: From the perspective of causal representation learning, the generating factors of images are composed of determinative factors and confounding factors. From the perspective of information theory, the determinative factors often have consistently high mutual information with the target class under most scenarios, whereas the confounding factors vary wildly, thus determinative factors have better generalization ability. In the case of coarse class subdivision, the sub-classes inherit the same determinative factors of the coarse class, and their difference lies only in values. Consequently, given two samples within the same sub-class, if we exchange their correspondence factors, even without knowing their semantic or whether they are determinative factors of the target class, the generated samples will still fall into the same sub-class. This observation is consistent with the experimental results.

Furthermore, the progress of controllable generators illuminates a new avenue for overcoming these problems, because the semantic-disentangled manifolds have excellent properties of local linearity to support abundant and unbiased augmentation even with limited and biased samples for each sub-class.

Based on the above observations, the Unbiased Manifold Augmentation (UMA) is proposed. It consists of a simple and effective progressive synthesis strategy, and a phase of progressive robust learning.

3.1 Progressive sample synthesis

As shown in Figure 2, the proposed UMA consists of two complementary phases to ensure the diversity and credibility of the generated samples respectively. We now first turn to describe the details of the progressive sample synthesis strategy. In the phase of progressive sample synthesis, given a target category and only a few training samples, a series of simple and effective synthesis strategies is conducted, in the latent-code manifolds of a pretrained editable generator.

It should be noted that the controllable generator needed by UMA is any StyleGAN or similar architecture whose latent codes containing the generating factors of the target class, so the editable generator can be trained at a much coarser granularity. E.g., a generic face generator can support any facial attribute recognition task with UMA, while the recognition of “swan” or “gull” can be improved with any generic bird generator. Since the controllable generative models

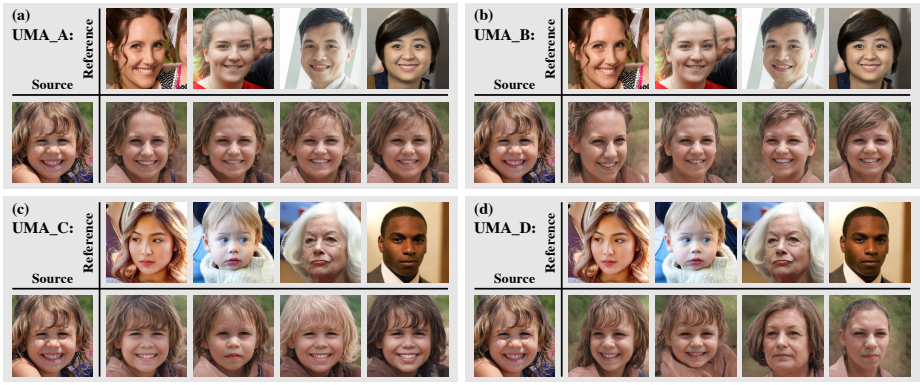


Fig. 3. Examples of the progressive diversity of synthetic samples for sub-class "smiling face", corresponding to UMA_A to UMA_D. The full version of UMA and UMA++ are different integration of them.

have seen rapid improvement recently, there are many off-the-shelf models pre-trained on well-known large datasets for coarse categories. They can be adopted as the cornerstone as our UMA, to support the unbiased manifold augmentation of any subclass of them, which meets the needs of many practical applications.

The advanced controllable generators often consist of two components, named the encoder-decoder architecture. First, a mapping network converts a given image into the latent code vectors $(W^0 \cdots W^{17})$ of a pre-trained StyleGAN, corresponding to its 18 layers of progressive generator with different resolutions. Second, after some manipulations, the modified latent vectors are then fed into the synthesis network to generate a new image. It has been demonstrated that different layers correspond to different semantic levels of image attributes from coarse to fine, taking facial attributes manipulation for example, from global pose to local details of the hair.

However, most existing manipulation methods are not suitable for unbiased augmentation due to the lack of safety and diversity guarantee. The generated images must still fall into the same class, and variants of the generating factors should be covered as many as possible, even with limited and biased training samples of this category.

For the purpose of coarse class classification, the progressive sample synthesis is proposed, based on the full exploration of the controllable generators' underlying properties.

Based on the observation introduced before, a series of progressive sample synthesis strategy denoted as UMA_A to UMA_D is proposed. Given a pair of samples x_i and x_j, the corresponding operations are as following:

$$W_i^k = e(x_i),\ W_j^k = e(x_j),\ k \in [0, 17] \tag{1}$$

$$\left\{W_i^k \bigodot W_j^k\right\}_t = \left\{W_j^k\ |_{t=k}, W_i^k\ |_{t\neq k}\right\}_t,\ t \in [0, 17] \tag{2}$$

$$\{\widehat{x}_{ij\_single}\}_t = \left\{g\left(W_i^k \bigodot W_j^k\right)\right\}_t \tag{3}$$

$$\left\{W_i^k \bigoplus W_j^k\right\}_t = \left\{W_j^k \mid_{t\leq k}, W_i^k \mid_{t>k}\right\}_t,\ t \in [0, 17] \tag{4}$$

$$\{\widehat{x}_{ij\_multiple}\}_t = \left\{g\left(W_i^k \bigoplus W_j^k\right)\right\}_t \tag{5}$$

Here e and g denote the encoder and decoder of the used controllable generator, while k means 18 latent-code layers. For UMA_A, the pair of seed images is randomly sampled from the target category, and new codes are generated by single layer switching, leading to 18 new images in all. UMA_B uses progressive switching layer by layer. Figure. 3 shows that the simple operation is very useful to guarantee the safety of the synthetic samples, as it only swaps the real attribute values within the same class, thus the determining factors and their manifolds are kept unchanged. Meanwhile, the progressive switching also brings in many reasonable variants of other factors. To further improve diversity, UMA_C and UMA_D use one sample in the target class and another from the coarser category, conducting single or multiple layers swapping respectively. We can see from Figure. 3 that the diversity and reliability of the synthetic samples gradually change from UMA_A to UMA_D.

Algorithm 1 : UMA (Unbiased Manifold Augmentation)

Input: a source dataset $S(s+, s-)$ for the target category c;
a controllable generator $G(e, g)$ for a coarser category C
Output: learned classification model M for task c

Initialize: $S_{training} \leftarrow S$
for $n = 1, ..., N$ **do**
  sample a pair of seed images (x_i, x_j) from $s+$:
  $S_{training} \leftarrow \{(\widehat{x}_{ij\_single}, y+)\}_t$, using $Eq.(1)(2)(3)$
  $S_{training} \leftarrow \{(\widehat{x}_{ij\_multiple}, y+)\}_t$, using $Eq.(1)(4)(5)$
  sample a pair of seed images (x_i, x_j) from $s-$:
  $S_{training} \leftarrow \{(\widehat{x}_{ij\_single}, y-)\}_t$, using $Eq.(1)(2)(3)$
  $S_{training} \leftarrow \{(\widehat{x}_{ij\_multiple}, y-)\}_t$, using $Eq.(1)(4)(5)$
end for
train M_{UMA} using $S_{training}$
for $n = 1, ..., N$ **do**
  sample seed pair (x_i, x_j), x_i from $s+$, x_j from C:
  $S_{training} \leftarrow \{(\widehat{x}_{ij\_single}, \widehat{y})\}_t$, using $Eq.(1)(2)(3)$
  $S_{training} \leftarrow \{(\widehat{x}_{ij\_multiple}, y)\}_t$, using $Eq.(1)(4)(5)$
  $\widehat{y}$ (or weight) is determined using robust learning
end for
train M_{UMA++} using $S_{training}$
return M_{UMA} or M_{UMA++}

3.2 Progressive robust learning

Theoretically, using the proposed progressive sample synthesis strategy mentioned above, infinite new samples can be generated. But the trustworthiness of their pseudo labels should also be taken into account.

We can see from Figure 3 that based on the underlying properties of the controllable generators, using seed images sampled from the target category, UMA_A and UMA_B can generate label-preserving samples. To further improve diversity, UMA_C and UMA_D are also introduced, but they often result in unreliable labels. For example, the layer swapping between images of a smiling person and a person who doesn't smile, may lead to a new one without obvious expression. Consequently, the phase of progressive robust learning is further integrated, to keep a good balance of the diversity and reliability for the synthetic samples.

As noisy labels may severely degrade the generalization, robust learning with noisy samples has gained significant attention in the machine learning community. Please refer to the comprehensive survey [22] for more information.

As introduced before, the diversity and reliability of the synthetic samples gradually change with UMA_A to D. In conjunction with it, a progressive robust learning strategy is proposed to cope with different scenarios. For datasets with random distribution, a classifier for the target category is trained on augmented dataset with label-preserving synthetic samples using UMA_A and B, which is called UMA in this paper. For a dataset with severe bias, only using UMA may still generate biased samples. To cope with it, UMA++ should be used. It consists of the complete series of progressive synthesis strategies and the progressive robust learning strategy. Here a simple method wildly used in semi-supervised learning [12, 21, 22] is adopted. A classifier is trained using UMA first. Then the classifier itself can be used to filter out unreliable samples obtained with UMA_C and D, and fine-tune in an iterative manner. The appropriate combination of the progressive sample synthesis and selection makes UMA a flexible mechanism. Thus, a good balance of the diversity and reliability of the synthetic samples can be achieved.

4 Experiments

To verify the effectiveness of the proposed UMA and UMA++ methods, extensive experiments are conducted on three publicly available datasets, CelebA [1], Stanford-Cars [11] and LSUN-Horses [25]. The performance of classification accuracy is compared with ten widely used data augmentation methods [5, 6, 9, 13, 24, 27, 28, 30], on various settings including random distribution and severe bias.

4.1 Datasets and settings

Different from traditional data augmentation methods which simulate global transformations, UMA can support the unbiased augmentation at the granularity of generating factors, thus we perform experiments on a diverse set of

challenging domains to illustrate the generalization of our approach. For facial domain, we perform various facial attributes recognition tasks from CelebA dataset [1]. Recognition tasks on Stanford Cars dataset [11] and LSUN horse dataset [25] are also conducted.

For sample synthesis, we start with the pretrained StyleGAN [4] generator for each coarse domain, e.g., a generic face generator pretrained on FFHQ dataset [3]. Then the latent-code encoder is further obtained by e4e [23] framework through image manipulation (or StyleGAN inversion) task. Hence, the latent code of each image could be extracted, which contains 18 layers with 512 dimensions. For the UMA_A and UMA_B synthesis mode, the source image and reference image are from the same class. In mode A, the latent code of the two images is extracted, and single layer is switched, then fed into the StyleGAN [4] generator to synthesis a new image. For B mode, the latent code of source image is replaced by the reference image from the first layer to the 18th layer. For the UMA_C and UMA_D synthesis mode, the source image and reference image are from the different class. In mode C, single layer of the latent code is switched, while the latent code is gradually replaced in mode D.

For these recognition tasks, we train the model from scratch, and no extra data or pretrained models are used. ResNet-18 is used as the default CNN backbone for feature extraction, and all images are resized to 224×224 size. The training batch size is 32 for all the tasks, and Adam optimizer is used during training with initial learning rate 0.004. We perform warm-up schedule in the first 5 epochs, then the learning rate is decayed by 0.1 every 20 epochs and total 50 epochs are trained. The training set is 256, and the test set is 1024. To train all the parameters in our model, we compute the cross-entropy between the prediction and target as the loss function. In the robust learning process, the probability distribution over the classes of the synthesized image is predicted. If the probability of the most likely class is higher than a predetermined threshold of 0.95, it would be added to the training dataset. The label of the sample would be assigned to its most likely class.

4.2 Evaluation of datasets with agnostic bias

For datasets with agnostic bias, we validate the effective ness of UMA by measuring classification accuracies on 11 challenging recognition tasks, across facial, car and horse domains [1, 11, 25], and compare the performance with 10 widely used data augmentation methods [5, 6, 9, 13, 24, 27, 28, 30]. All the results in Table 1 and Figure 4 show that the UMA is very effective, and consistently outperforms other augmentation methods significantly.

As shown in Table 1, we perform five single attributes recognition tasks (black hair, eyeglasses, heavy makeup, smiling and bald) and two tasks for combined attributes recognition (CA #1 is heavy makeup and smiling, CA #2 is male with black hair) in facial domain. The subclass recognition tasks in car or horse domain are also included. We can see that the performances of other methods are not stable, revealing that they are not suitable for all scenarios. By contrast, UMA performs best in all tasks. Especially, in the eyeglass-wearing recognition

Table 1. Accuracy comparison on 11 coarse class subdivision tasks with 10 widely used data augmentation methods. All methods use an identical backbone architecture and the same default parameters. The number of training samples is only 32. More details of the task settings please refer to section 4.

Method	Publication	Black Hair	Eye-glasses	Makeup	Smilling	Bald	CA#1
MixUp [28]	ICLR'18	74.28	77.39	67.58	55.76	76.56	64.45
CutMix [27]	ICCV'19	70.68	74.45	62.89	51.76	76.17	65.82
Auto Augment [5]	CVPR'19	71.55	77.76	59.67	54.98	74.22	63.77
Manifold Mixup [24]	ICML'19	75.76	75.74	67.29	58.40	76.95	66.70
Random Erasing [30]	AAAI'20	74.24	75.55	60.94	50.78	73.05	62.79
Random Augment [6]	NIPS'20	71.52	72.06	59.28	54.69	73.83	59.18
MoEx [13]	CVPR'21	74.09	78.68	69.73	53.22	81.64	66.80
StyleMix [9]	CVPR'21	68.71	78.86	63.67	61.23	71.88	51.76
StyleCutMix [9]	CVPR'21	72.34	69.30	64.84	60.94	75.39	66.11
StyleCutMix(auto-γ) [9]	CVPR'21	68.67	71.32	63.09	58.98	74.21	67.09
UMA	-	**79.20** (3.44↑)	**90.99** (12.13↑)	**74.41** (4.68↑)	**67.87** (6.64↑)	**84.77** (3.13↑)	**76.95** (9.86↑)
UMA++	-	**79.89** (4.13↑)	**93.01** (14.15↑)	**77.05** (7.32↑)	**69.53** (8.30↑)	**87.50** (5.86↑)	**78.03** (10.94↑)

Method	Publication	CA#2	Sedan	SUV	Brown Horse	White Horse	Average
MixUp [28]	ICLR'18	74.51	62.89	62.30	79.69	84.38	70.89
CutMix [27]	ICCV'19	71.00	65.04	62.50	79.69	79.69	69.06
Auto Augment [5]	CVPR'19	72.75	68.75	66.99	85.94	78.91	70.48
Manifold Mixup [24]	ICML'19	74.32	63.96	61.91	76.56	80.47	70.73
Random Erasing [30]	AAAI'20	72.07	61.52	59.08	78.91	78.12	67.91
Random Augment [6]	NIPS'20	71.78	66.99	68.55	81.25	81.25	69.13
MoEx [13]	CVPR'21	71.97	60.16	57.62	85.16	86.72	71.44
StyleMix [9]	CVPR'21	73.34	63.96	62.50	78.12	81.25	68.66
StyleCutMix [9]	CVPR'21	72.75	64.36	59.57	75.00	78.91	69.05
StyleCutMix(auto-γ) [9]	CVPR'21	72.27	64.84	62.65	76.56	85.16	69.53
UMA	-	**76.86** (2.35↑)	**72.56** (3.81↑)	**71.29** (2.74↑)	**89.84** (3.90↑)	**88.28** (1.56↑)	**79.37** (7.93↑)
UMA++	-	**81.74** (7.23↑)	**74.61** (5.86↑)	**74.90** (6.35↑)	**91.41** (5.47↑)	**92.19** (5.47↑)	**81.81** (10.37↑)

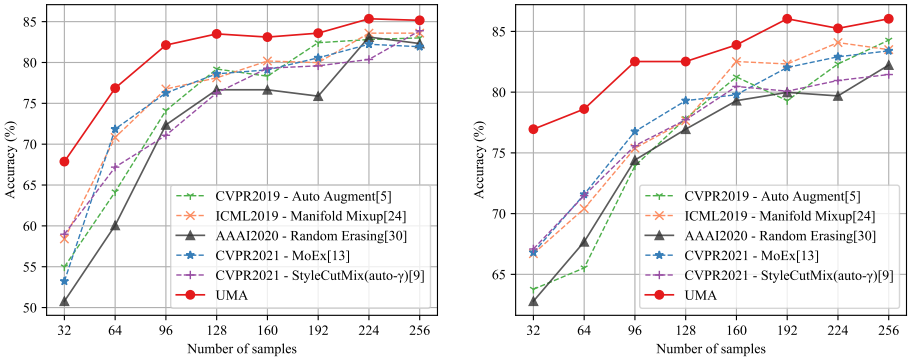
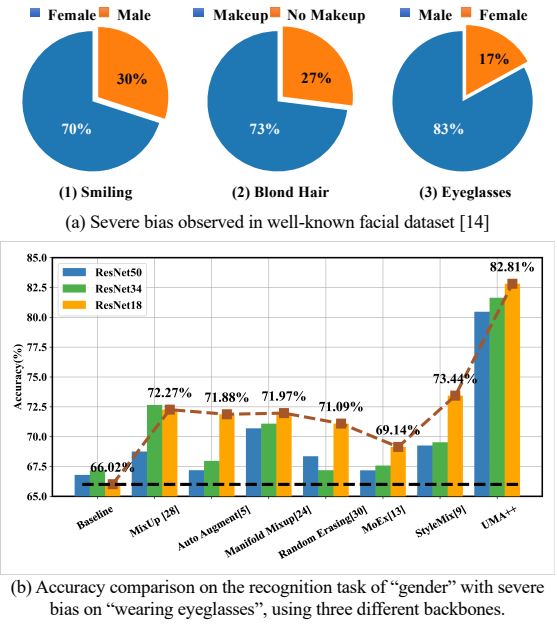


Fig. 4. Accuracy comparison on various sub-class recognition tasks, with different number of training samples. To simulate the challenging small-data-learning scenario of coarse class subdivision task, we only provide less than 1% fine-grained samples of commonly used.

task, our UMA and UMA++ are superior to the best one with a large margin of 12.13% and 14.15%. On average, UMA and UMA++ gain 7.93% and 10.37% improvement than the best method.

Figure 4 shows the details of accuracy varying with different number of training samples. In general, the accuracy substantially improves with the increase of training data, whereas the marginal gain decreases gradually. This means that the less data, the more important data augmentation is. UMA has a greater advantage with limited training data. With the whole training data, 2% improvement is obtained. While using only 12.5% training data, UMA can achieve up to 7% improvement in smiling recognition and 9.8% improvement in combined attribute recognition task of heavy makeup with smiling. This is a compelling proof of the diversity and reliability of the progressive synthesis and selection method in UMA.

4.3 Evaluation of datasets with severe bias



(a) Severe bias observed in well-known facial dataset [14]

(b) Accuracy comparison on the recognition task of "gender" with severe bias on "wearing eyeglasses", using three different backbones.

Fig. 5. The observation of severe bias in wildly-used datasets, and the accuracy comparison between our UMA and the state-of-the-art data augmentation methods.

For a dataset with severe bias, only UMA is not enough, and UMA++ is very suitable for this problem. Severe bias is very common in many well-known datasets. As shown in Figure. 5(a), 70% of the annotation samples in [1] with

smiling are female, while 73% annotated blond hair are heavily makeup, and 83% person with eyeglasses are male. The highly-coupled attributes can become the confounding factors during recognition. E.g., "smiling" would hinder the discrimination of "gender" attribute.

To study the effectiveness of UMA++ with severe bias, we conduct gender recognition on a biased dataset. In the training dataset, 95% of the male samples are wearing eyeglasses, and 5% of the female samples are wearing eyeglasses, while in the test dataset, 50% of the samples are wearing eyeglasses for all genders.

We start with a baseline classification model with only random horizontal flip augmentation, and achieves 66.02% accuracy. Because of the severe bias in the training dataset, the model heavily relies on the factors of eyeglasses, while the intrinsic factors of gender are neglected. To relieve this problem, samples of male without eyeglasses and female with eyeglasses are desired. Hence, we generate samples using UMA_C and UMA_D, which selects reference images outside the target class. It can introduce valuable variants, e.g., male samples without eyeglasses and female samples with eyeglasses. In UMA++, samples synthesized by UMA_C are assigned with the same label with the source image, and added to the original dataset to train a base classifier. Then samples are further synthesized by UMA_D and selected with robust learning such as [12], achieving 16.79% improvement compared with baseline, as shown in Figure. 5(b). Compared with the state-of-the-art methods, the UMA++ still outperforms them by 9.37%.

We also conduct experiments with different backbones, Figure. 5(b) shows that UMA++ outperforms others consistently with various backbones. With limited and biased datasets, ResNet-18 achieves the best performance, which implies that deeper backbone could learn to fit the confounding factors much easier and reduce the performance.

Table 2. Performance comparison with several state-of-the-art FGVC methods, on a dataset with severe bias. Baseline stands for resnet-18 backbone with one fully connected layers. UMA++ and various FGVC methods are added to evaluate the performance.

Method	Publication	Baseline	Baseline + UMA++	Baseline + FGVC	Baseline +FGVC +UMA++
SPS [16]	ICCV'21	66.02%	82.81%	72.27%	83.59%(**11.31**↑)
ProtoTree [10]	CVPR'21			66.80%	84.38%(**17.58**↑)
CAL [17]	CVPR'21			73.05%	84.77%(**11.72**↑)

For sub-class recognition tasks, many Fine-Grained Visual Classification(FGVC) methods have been proposed [10, 16, 17], focusing on the design of networks or attention mechanism, complementary with data augmentation discussed in this

paper. Table 2 shows that compared with the state-of-the-art FGVC strategies, our UMA can still perform best with dataset bias. Besides, our UMA is complementary with these FGVC strategies, and further improve the performance by over 11.31%.

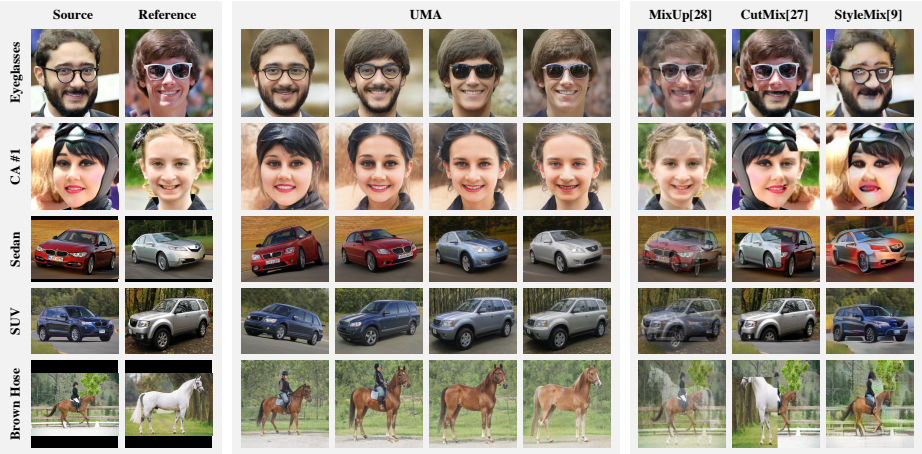


Fig. 6. Visualization of the proposed UMA and other wildly used data augmentation methods. The reliability and diversity of the synthetic samples using the UMA are better than other augmentation methods conducted in image space or feature space

5 Conclusion

To support the challenging CCS task with minimum fine-grained labeling cost, the Unbiased Manifold Augmentation (UMA) is proposed. Leveraging the recent progress of controllable generators, unbiased and reliable sample synthesis is conducted in the disentangled latent-code manifolds, at the granularity of generating factors or called attributes, different from traditional augmentation in highly-coupled image or feature space. UMA can reveal the approximate unbiased mutual information between the target class and all of its impact factors, thus guides the classifier to focus on causal factors.The proposed framework is independent of the adopted editable generator or the specific robust learning method. Experiments have shown that with a generator for a coarser category, our UMA can greatly improve the generalization ability of the DNNs for subclasses recognition, on datasets with agnostic even severe bias.

References

1. Celeba, http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
2. Cub-200-2011, http://www.vision.caltech.edu/visipedia/CUB-200-2011.html
3. ffhq-dataset, https://github.com/NVlabs/ffhq-dataset
4. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8188–8197 (2020)
5. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 113–123 (2019)
6. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 702–703 (2020)
7. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016)
8. Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S.: Ganspace: Discovering interpretable gan controls. Advances in Neural Information Processing Systems **33**, 9841–9850 (2020)
9. Hong, M., Choi, J., Kim, G.: Stylemix: Separating content and style for enhanced data augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14862–14870 (2021)
10. Huang, S., Wang, X., Tao, D.: Stochastic partial swap: Enhanced model generalization and interpretability for fine-grained recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 620–629 (2021)
11. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 554–561 (2013)
12. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3, p. 896 (2013)
13. Li, B., Wu, F., Lim, S.N., Belongie, S., Weinberger, K.Q.: On feature normalization and data augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12383–12392 (2021)
14. Li, Y., Yu, Q., Tan, M., Mei, J., Tang, P., Shen, W., Yuille, A., Xie, C.: Shape-texture debiased neural network training. arXiv preprint arXiv:2010.05981 (2020)
15. Lin, J., Zhang, R., Ganz, F., Han, S., Zhu, J.Y.: Anycost gans for interactive image synthesis and editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14986–14996 (2021)
16. Nauta, M., van Bree, R., Seifert, C.: Neural prototype trees for interpretable fine-grained image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14933–14943 (2021)
17. Rao, Y., Chen, G., Lu, J., Zhou, J.: Counterfactual attention learning for fine-grained visual categorization and re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1025–1034 (2021)
18. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2287–2296 (2021)

19. Roich, D., Mokady, R., Bermano, A.H., Cohen-Or, D.: Pivotal tuning for latent-based editing of real images. arXiv preprint arXiv:2106.05744 (2021)
20. Schölkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A., Bengio, Y.: Toward causal representation learning. Proceedings of the IEEE **109**(5), 612–634 (2021)
21. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in Neural Information Processing Systems **33**, 596–608 (2020)
22. Song, H., Kim, M., Park, D., Shin, Y., Lee, J.G.: Learning from noisy labels with deep neural networks: A survey. arXiv preprint arXiv:2007.08199 (2020)
23. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for stylegan image manipulation. ACM Transactions on Graphics (TOG) **40**(4), 1–14 (2021)
24. Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., Bengio, Y.: Manifold mixup: Better representations by interpolating hidden states. In: International Conference on Machine Learning. pp. 6438–6447. PMLR (2019)
25. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015)
26. Yue, Z., Wang, T., Sun, Q., Hua, X.S., Zhang, H.: Counterfactual zero-shot and open-set visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15404–15414 (2021)
27. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6023–6032 (2019)
28. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
29. Zheng, X., Chalasani, T., Ghosal, K., Lutz, S., Smolic, A.: Stada: Style transfer as data augmentation. arXiv preprint arXiv:1909.01056 (2019)
30. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 13001–13008 (2020)