Reducing Implicit Class Imbalance in Unlabeled Datasets Using Text-Specified Sensitive Attributes

Kosei Suayama* and Kazuaki Nakamura*
* Tokyo University of Science
* E-mail: 4625521@ed.tus.ac.jp, nakamura.kazuaki@rs.tus.ac.jp

***Abstract*—Modern AI models sometimes output biased results on sensitive attributes (SA) such as "race" and "age", which can cause a social issue. A reason for these biased outputs is a class imbalance problem in the training dataset. Therefore, reducing the class imbalance of a given dataset is important. Although the class imbalance could be implicitly contained in unlabeled datasets that have no class labels, no existing studies focus on the problem of implicit class imbalance in unlabeled datasets, except for our previous work. Even our method has a serious drawback: it cannot explicitly specify which attribute should be focused on as SA. In this paper, we propose a method for reducing the implicit class imbalance in unlabeled datasets that allows us to specify SA by text. The proposed method first extracts an SA-related visual feature from each image in a target imbalanced dataset using a vision-language model called CLIP. The extracted features are next utilized to cluster the target dataset. Then, we train a conditional diffusion model, regarding the cluster IDs as pseudo-class labels, and generate the same number of images from each pseudo-class to construct a balanced dataset. The results of our experiments focusing on two types of SA, "race" and "age", demonstrated that the proposed method is capable of effectively reducing the implicit class imbalance in a given unlabeled face dataset.**

I. Introduction

Recently, machine learning-based AI models have been rapidly advancing. On the other hand, such AI models sometimes output biased results on sensitive attributes (SA) such as "race", "age", and "gender", which are socially unacceptable. For instance, output images of image-generation AI such as DALL-E2 and Stable Diffusion v1.4 correlate with the demographics of US labor, where minority groups tend to be under-represented [1]. In addition, some commercial gender classification models perform better on images of males compared to those of females and also show higher accuracy on images with lighter skin than those with darker skin [2]. Such biased outputs of AI models are mainly caused by a class imbalance problem in the training dataset. For instance, the training dataset of the above gender classification models may contain many face images with lighter skin and only a few face images with darker skin. Hence, methods for reducing the class imbalance in training datasets are needed to avoid the unacceptable behaviors of AI models.

So far, a lot of methods for reducing the class imbalance in labeled datasets have been proposed. Typical examples include over-sampling [3], [4], the strategy of generating new samples of rare classes and adding them to the original labeled dataset, and under-sampling, the strategy of removing some samples of common classes from the original dataset [5].

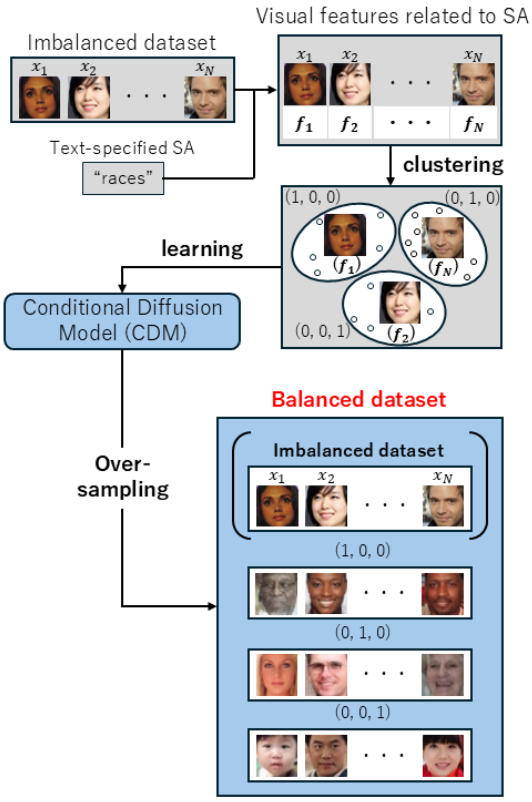

Fig. 1. Overview of the proposed method for reducing the implicit class imbalance in unlabeled datasets, where we can explicitly specify SA by text.

Loss functions for training image recognition models or object detection models with imbalanced datasets have also been actively studied [6], [7], [8].

However, to the best of our knowledge, no existing methods focus on unlabeled datasets even though they could be implicitly imbalanced, similar to labeled datasets. For example, even an unlabeled face dataset could have many face images of White people and only a few face images of Black and Asian people when observed with human eyes. If such an unlabeled face dataset containing an implicit class imbalance is used to train face detection models or face generation models, the trained models will provide biased results. This issue cannot be solved by the above existing methods because they focus only on labeled datasets.

Based on the discussion up to here, we focus on a method for reducing the implicit class imbalance in unlabeled datasets

979-8-3315-7206-8/25/$31.00

to balance them. We have previously proposed such a method employing an over-sampling strategy, experimenting with unlabeled face datasets [9]. However, this method has two drawbacks. First, additional samples generated by our previous method do not have sufficient image quality. Second, we cannot specify which attribute is critically sensitive for a target dataset. Due to this issue, our previous method tends to consider various attributes such as illumination conditions, face directions, facial expressions, and so on, even when "race" is only a sensitive attribute (SA). To tackle this problem, this paper proposes a novel method for reducing the implicit class imbalance that can explicitly specify SA by text. Using the text-specified SA, the proposed method first extracts an SA-related visual feature from each image in the target dataset. A set of extracted features is then clustered to give a pseudo-label to each image. Then, we apply an over-sampling strategy based on the pseudo-labels, as shown in Fig. 1. The contributions of this paper are summarized as follows.

- This paper realizes a mechanism for giving SA-related pseudo-labels to each image using a text-specified SA and a vision-language model, i.e., CLIP [10].
- This paper improves the quality of over-sampled images generated by the proposed balancing method by employing a Conditional Diffusion Model (CDM) [11] trained with pseudo-labels.
- This paper demonstrates the effectiveness of the proposed method by conducting two experiments that focus on different types of SA, i.e., "race" and "age", respectively.

II. Related Works

A. Approach for Class Imbalance in Labeled Datasets

There are two existing approaches for tackling the class imbalance problem in labeled datasets: one is to develop a loss function suitable for imbalanced datasets, and the other is to reduce the class imbalance.

1) To develop a loss function: Lin et al. proposed a loss function called Focal Loss, which is suitable for training object detection models [6] with an imbalanced dataset. Since the number of background pixels is much larger than that of foreground pixels in object detection, Focal Loss adds a modularization term into Cross-Entropy Loss that balances the weights of the foreground and background pixels. This is specialized for object detection models and cannot be applied directly to other tasks. Therefore, Cui et al. proposed Class-Balanced Loss suitable for classification tasks [7]. They introduce the concept of "effective number" to estimate the number of samples for each class and use the reciprocal of the effective number as a weight term for the corresponding class. This makes the loss of rare classes greater. Many other loss functions, such as Affinity Loss [8], have also been proposed to address the imbalance problem. However, these loss functions require class labels and cannot be applied to unlabeled datasets.

2) To reduce the class imbalance: Reducing the class imbalance in labeled datasets is achieved by under-sampling or over-sampling, as mentioned in Section I. For the former, Wei-Chao et al. proposed an under-sampling technique using clustering [12]. This method first clusters samples of non-rare classes into K clusters, where K is the number of samples in the rarest class. Then, only the center or medoid of each cluster is maintained, and the remaining data are removed. Hou et al. proposed a density-based under-sampling technique [13]. This method first estimates the density distribution of samples of non-rare classes using kernel density estimation and then finds the local maxima of the estimated distribution. Finally, only the data located on the local maxima or their neighboring area are maintained. Many other under-sampling methods have been proposed, whose details are reported in Devi's survey [5]. For over-sampling, SMOTE (Synthetic Minority Oversampling Technique) [3] is a representative method, which reduces the class imbalance by connecting each sample of a rare class and its neighbor with a line and generating new samples along the line. Many other over-sampling methods like ADASYN (Adaptive Synthetic Sampling Approach) [4] have also been proposed. However, all of the above methods require class labels and cannot be applied to unlabeled datasets.

Joo et al. proposed another balancing method, particularly focusing on face image datasets [14]. They train an encoder that can separately extract SA and non-SA attributes from a face image and a decoder that reconstructs a face image from a pair of SA and non-SA attributes. Using them, only the SA of each image is manipulated from a non-rare value (e.g., White race) to a rare value (e.g., Black race). This method also requires labels of SA to train the encoder and decoder. Therefore, we cannot apply it to unlabeled datasets.

B. Methods for Balancing Unlabeled Datasets

Unlike existing methods, we focused on unlabeled datasets and proposed a balancing method leveraging the theoretical characteristics of GANs [15] in our previous work [9]. We found that the generation probability of GAN's images is correlated with their appearance variation caused by a perturbation in the input vector; images with lower generation probability (i.e., rare images) tend to have larger appearance variation. Based on this characteristic, our previous method calculates the "rarity" of each generated image using its appearance variation. Then, only the images with higher rarity are added to a target unlabeled dataset to balance it. However, this method has two drawbacks, as mentioned in Section I. First, since the method calculates the appearance variation directly in the pixel space, not only SA (e.g., "race" and "age" for face images) but also various non-SA attributes (e.g., face directions, illumination conditions, etc.) are taken account into the calculation of the "rarity". Second, high-rarity images tend to have low quality since GAN's training dataset does not include low-quality images. We tackle these problems in this work.

III. Proposed Method

A. Overview

In the proposed method, we explicitly specify which attribute is SA by text to solve the problem described above. For instance, if we want to balance a target dataset in terms

of race, we input "races" into our proposed system as a text for specifying SA.

We employ over-sampling as a basic strategy for reducing the implicit class imbalance in the target dataset. The over-sampling strategy can be applied not only to labeled datasets but also to unlabeled datasets if we can assign some pseudo-label to each image in the datasets. Note that, in our situation, we have to construct a set of pseudo-labels according to the text-specified SA. The overall procedure of the proposed method to satisfy this requirement is as follows.

1) Extract a visual feature closely that is related to the text-specified SA from each image in the imbalanced target dataset.
2) Cluster the target dataset based on the extracted visual features, regarding resultant cluster IDs as pseudo-labels.
3) Train a CDM using the target dataset with the pseudo-labels, which is used to generate new images to construct a balanced dataset (i.e., to perform over-sampling).

We describe the above steps 1) to 3) in detail in Subsections III-B to III-D, respectively.

B. Extracting Visual Features Closely Related to Specified SA

1) Extracting visual features using CLIP: Since SA is specified by text, we need a mechanism that bridges the two different modalities, text and images, to extract an SA-related visual feature from each image. To this end, we utilize a vision-language model, CLIP [10], which consists of a text encoder (TE) that embeds texts into a feature space and a vision encoder (VE) that embeds images into the same feature space. In CLIP, since the two modalities share a common feature space, a text and an image containing similar semantic content are embedded into locations close to each other. Based on this property, the proposed method extracts visual features that are closely related to SA as follows. First, k other texts related to the specified SA are prepared. The way of achieving this is described later. Hereafter, we refer to these k texts as *related texts*. The *related texts* are expected to contain some attribute values of SA. For instance, when the specified SA is "races", one possible *related text* is "a photo of a yellow race, which means an Asian race", where "yellow" and "Asian" are the attribute values. Next, we input each *related text* w_i $(i = 1, \ldots, k)$ to TE and obtain a corresponding embedding vector $\mathrm{TE}(w_i)$. At the same time, we also input each image x_j $(j = 1, \ldots, N)$ in the target dataset to VE and obtain an embedding vector $\mathrm{VE}(x_j)$. Then, we calculate the cosine similarity between $\mathrm{TE}(w_i)$ and $\mathrm{VE}(x_j)$ as

$$f_{ij} = \frac{\mathrm{TE}(w_i)^\top \mathrm{VE}(x_j)}{||\mathrm{TE}(w_i)|| \cdot ||\mathrm{VE}(x_j)||} . \tag{1}$$

Finally, we use $\boldsymbol{f}_j = (f_{1j} \quad \cdots \quad f_{kj})^\top$ as an SA-related visual feature for image x_j.

2) Preparing related texts: As mentioned above, *related texts* are expected to contain some attribute values. For example, "White", "Asian", and so on should be contained in the *related texts* if SA is "races", while "young", "fifties", and so on should be contained if SA is "age". To satisfy this requirement, we additionally specify a few (one or two) attribute values in conjunction with the SA. The proposed way of obtaining the *related texts* is below.

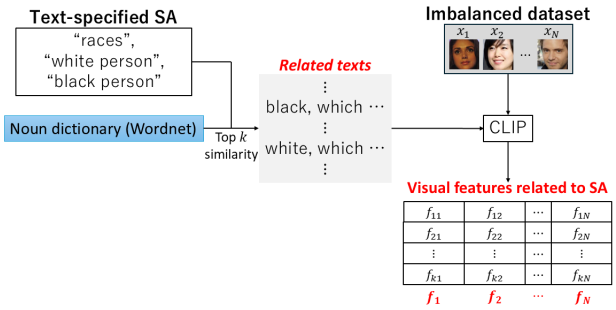

Fig. 2. Overall procedure for SA-related feature extraction.

In advance, we prepare a word dictionary that lists English nouns and their corresponding descriptions, where the nouns are used as candidates for the *related texts*. Specifically, WordNet [16] is used for this purpose. Let W_h be h-th noun in WordNet and W_h^+ be the combination of W_h and its description. For example, in WordNet, the noun "yellow race" is described as "an Asian race". In this case, W_h = "yellow race" and W_h^+ = "yellow race, which means an Asian race". Using these descriptions, we first input W_h^+ to the CLIP text encoder and obtain the corresponding embedding vector $\mathrm{TE}(W_h^+)$ for all $h \in \{1, 2, ...\}$. Next, we combine the specified SA and its additional attribute values to form a text sentence W_{SA}. For instance, if SA is "races" and its additional attribute values are "White person" and "Black person", W_{SA} = "races such as White person, Black person". Note that the SA itself can be used as W_{SA} if there are no additional attribute values. Then, we input W_{SA} to the CLIP text encoder and obtain an embedding vector $\mathrm{TE}(W_{\mathrm{SA}})$. After that, we calculate the cosine similarity between $\mathrm{TE}(W_{\mathrm{SA}})$ and $\mathrm{TE}(W_h^+)$ as

$$S_h = \frac{\mathrm{TE}(W_{\mathrm{SA}})^\top \mathrm{TE}(W_h^+)}{||\mathrm{TE}(W_{\mathrm{SA}})|| \cdot ||\mathrm{TE}(W_h^+)||} \tag{2}$$

for all h. Finally, W_h^+ corresponding to the top k highest similarity is selected as the *related texts*.

The overall procedure for obtaining SA-related visual features is depicted in Fig. 2.

C. Obtaining Pseudo-Labels by Clustering Target Dataset

To assign a pseudo-label to each image in the target dataset, we cluster it utilizing the visual features extracted by the above method. Each cluster obtained with this process is expected to correspond to a certain SA attribute value. For instance, if SA is "races", the clusters corresponding to "White", "Black", "Asian", and so on are expected to be obtained since race-related visual features are used in the clustering process. Thus, based on the clustering result, we assign each image the ID of the cluster to which it belongs as a pseudo-label. Hereafter, we let m denote the assigned pseudo-label, which is given in the form of a one-hot vector.

D. Reducing Implicit Class Imbalance by Over-sampling

As described in Section II-B, our GAN-based previous method [9] is disadvantageous in the quality of over-sampled images. Therefore, the proposed method employs a CDM [11] to train a conditional image generation model for over-sampling, using the pseudo-labels m as a conditional input. CDM is an extension of Diffusion Model [17] that can deal with a conditional input and is capable of stably generating higher quality images than Conditional GAN (CGAN) [18]. This property is beneficial for solving the above issue. After training a CDM, we use it to generate the same number of images from each cluster (pseudo-label), where we regard a set of generated images as a final output, i.e., a balanced dataset.

IV. Experiments

A. Experimental Setup

We experimentally demonstrated the effectiveness of the proposed method using UTKFace [19]. UTKFace is a face image dataset with race, age, and gender labels and is imbalanced in terms of race and age. Thus, we separately conducted two experiments where "race" and "age" are specified as SA, respectively, and attempted to reduce the implicit class imbalance from each aspect. In the first experiment focusing on race as SA, we used all images in UTKFace except those labeled as "Others"; in more detail, we created a subset of UTKFace consisting of 10078 images of "White" people, 4527 images of "Black" people, 3434 images of "Asian" people, and 3976 images of "Indian" people (22015 images in total). We refer to this image set as a "racially imbalanced dataset" in the remainder. On the other hand, in the second experiment focusing on age as SA, we rearranged the age labels of UTKFace, which are originally represented in integers, into the following five classes: under ten (–10), twenties (20–30), forties (40–50), sixties (60–70), and over eighty (80–). Then, we created another subset consisting of 3062 images of people under 10 years old, 7344 images of people in their twenties, 2245 images of people in their forties, 1318 images of people in their sixties, and 523 images of people over 80 years old (14492 images in total). We refer to this image set as an "age-imbalanced dataset". These two datasets were to be balanced by the proposed method in each experiment, respectively. Note that we used the race and age labels of UTKFace only for creating these imbalanced datasets; we ignored the label information when reducing the class imbalance.

For the hyperparameter setting in the proposed method, we set the number of *related texts* as 20 ($k = 20$). We employed X-means [20] as the algorithm for clustering the SA-related visual features. X-means can automatically determine the optimal number of clusters, but when running it, we limited the maximum number of clusters to 20. For evaluation criteria, we employed entropy as an index of the class imbalance, which is calculated as

$$H = -\sum_{c \in C} P_c \ln P_c \ . \tag{3}$$

Fig. 3. Examples of generated images of each cluster listed in Table I.

In the above formula, c is a specific attribute value of SA (e.g., "White", "Black", etc. in the first experiment), C is a set of all attribute values, and P_c is the ratio of images corresponding to the attribute value c. The effectiveness of the proposed method can be demonstrated if a balanced dataset produced by the proposed method shows higher entropy than the original imbalanced dataset for a specified SA.

B. Performance Evaluation on Racially Imbalanced Dataset

In the first experiment, we evaluated the race distribution and the entropy of the following datasets to compare them.

(A) Racially imbalanced dataset created in the last section.
(B) Resultant dataset when we balanced the above (A) by our previous method [9].
(C) Resultant dataset when we balanced the above (A) by only specifying "races" as SA in the proposed method.
(D) Resultant dataset when we balanced the above (A) by specifying an additional attribute value "white person" in addition to "races" as SA.
(E) Resultant dataset when we balanced the above (A) by specifying two additional attribute values, "white person" and "black person", in addition to "races" as SA.
(F) Resultant dataset when we balanced the above (A) by directly setting four ideal related texts, "White person", "Black person", "Asian person", and "Indian person", in the proposed method.

Since the images over-sampled by the proposed method and our previous one are unlabeled, we estimated their race class using CLIP as an open-set classifier to calculate the race distribution and the entropy of each dataset. Note that the race estimation accuracy of CLIP was 90.3% when we tested it with (A). To create the datasets (B) to (F), we over-sampled 2000 images in each case. More specifically, we created (B) by generating 20000 images using a GAN and selecting only the images with the top 2000 highest rarity. When creating (C) to (F), we generated $2000/M$ images from each cluster (or pseudo-label), where M is the number of clusters obtained by the process of Section III-C.

TABLE I
CLUSTERING RESULT IN THE FIRST EXPERIMENT (DATASET (E)).

Cluster ID	White	Black	Asian	Indian	Majority class
0	105	4	648	140	Asian (72.2%)
1	884	10	18	21	White (94.7%)
2	58	13	393	594	Indian (56.1%)
3	133	25	776	328	Asian (61.5%)
4	1041	25	58	86	White (86.0%)
5	963	2	72	37	White (89.7%)
6	1443	8	25	47	White (94.7%)
7	743	23	10	37	White (91.4%)
8	118	44	266	730	Indian (63.0%)
9	1507	7	20	72	White (93.8%)
10	36	10	170	354	Indian (62.1%)
11	422	86	54	342	White (46.7%)
12	71	1076	25	64	Black (87.1%)
13	374	47	240	157	White (45.7%)
14	109	1165	42	130	Black (80.6%)
15	66	430	37	56	Black (73.0%)
16	31	720	7	38	Black (90.5%)
17	29	746	9	55	Black (88.9%)
18	1853	37	72	108	White (89.5%)
19	92	49	492	580	Indian (47.8%)

Table I shows the clustering result of the proposed method for dataset (E). Besides, Fig. 3 shows examples of the generated images of each cluster. We can qualitatively and quantitatively see from these results that each cluster tends to correspond to a single race class. In addition, images generated by the proposed method have enough quality for various races. Table II shows the race distribution and the entropy of datasets (A) to (F). As shown in this table, the implicit class imbalance is successfully reduced in (C) to (E) compared with (A). These results indicate the effectiveness of the proposed method. However, in the case of (C), the obtained *related texts* did not necessarily contain words expressing race, and therefore, we could not observe a clear correlation between cluster IDs and race classes. Comparing (D) and (E), the implicit class imbalance is more reduced in (D) than in (E). This is because we limit the maximum number of clusters to 20. In fact, when we removed this limitation and conducted the same experiment, the result for (E) outperformed that for (D). On the other hand, the performance of the proposed method (i.e., entropy of (C), (D) and (E)) is almost the same as that of our previous method (i.e., entropy of (B)). This is due to an undesired property of the *related texts*; some of them contained words expressing not only race but also gender, such as "white **man**", particularly in the cases of (D) and (E). Due to this property, non-SA attributes were taken into account when extracting SA-related visual features. In (F), there is almost no imbalance thanks to the ideal *related texts*. This indicates that the performance of the proposed method could be improved by a better acquisition method of *related texts*.

C. Performance Evaluation on Age-Imbalanced Dataset

In the second experiment, we compared the age group distribution and the entropy of the following datasets.

(A) Age-imbalanced dataset created in Section IV-A.
(B) Resultant dataset when we balanced the above (A) by our previous method [9].

TABLE II
RACE DISTRIBUTION AND ENTROPY OF EACH DATASET.

	White	Black	Asian	Indian	Entropy
(A)	45.80	20.56	15.60	18.10	1.282
(B)	37.85	23.30	21.80	17.05	1.341
(C)	39.35	23.50	16.95	20.20	1.331
(D)	37.20	25.85	17.85	19.10	1.341
(E)	39.15	24.50	19.70	16.65	1.330
(F)	27.95	22.15	25.80	24.10	**1.383**

TABLE III
CLUSTERING RESULT IN THE SECOND EXPERIMENT (DATASET (E)).

Cluster ID	Age group					Majority class
	–10	20–30	40–50	60–70	80–	
0	0	356	172	118	9	20–30 (54.4%)
1	2	58	75	267	117	60–70 (51.4%)
2	0	9	15	107	164	80– (55.6%)
3	1	1	0	19	148	80– (87.6%)
4	31	1070	168	13	2	20–30 (83.3%)
5	15	861	122	17	5	20–30 (84.4%)
6	9	696	68	5	0	20–30 (89.5%)
7	2	201	299	273	40	40–50 (36.7%)
8	618	37	3	0	0	–10 (93.9%)
9	1148	46	4	0	0	–10 (95.8%)
10	438	283	4	0	0	–10 (60.4%)
11	590	46	1	0	0	–10 (92.6%)
12	16	808	117	3	0	20–30 (85.6%)
13	3	306	482	273	27	40–50 (44.2%)
14	38	843	48	5	0	20–30 (90.3%)
15	113	606	24	0	0	20–30 (81.6%)
16	1	286	399	145	9	40–50 (72.1%)
17	36	498	20	2	0	20–30 (89.6%)
18	0	158	35	8	0	20–30 (78.6%)
19	1	175	189	63	2	40–50 (44.0%)

(C) Resultant dataset when we balanced the above (A) by only specifying "age groups" as SA in the proposed method.
(D) Resultant dataset when we balanced the above (A) by specifying an additional attribute value "youth" in addition to "age groups" as SA.
(E) Resultant dataset when we balanced the above (A) by specifying two additional attribute values, "youth" and "old age", in addition to "age groups" as SA.

Similar to the first experiment, generated images do not have labels for evaluating the age group distribution and entropy. Therefore, we trained an age group classifier based on dataset (A) and used it to estimate the age group of each generated image. Note that the estimation accuracy of the trained classifier was 82.5% when we tested it on another subset of UTKFace that is different from (A). To create datasets (B) to (F), we generated 2000 images in each case by the same procedure as the previous section.

Table III shows the clustering result of the proposed method for dataset (E), and Fig. 4 shows examples of the generated images of each cluster. In these results, we can see the same tendency as Table I and Fig. 3; most clusters correspond to a single age group, and the generated images have good quality for various age groups. Table IV shows the age group distribution and the entropy of datasets (A) to (E). As shown in this table, the entropy of (C), (D), and (E) is much higher than

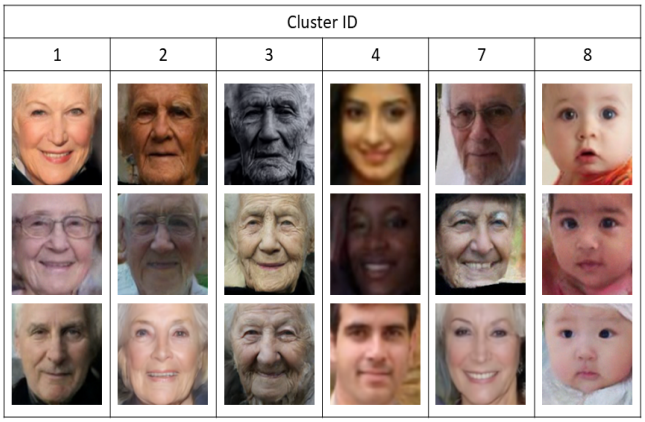

Fig. 4. Examples of generated images of each cluster listed in Table III.

TABLE IV
AGE GROUP DISTRIBUTION AND ENTROPY OF EACH DATASET.

	Age group					Entropy
	–10	20–30	40–50	60–70	80–	
(A)	21.13	50.68	15.49	9.09	3.61	1.300
(B)	12.95	55.95	4.05	8.55	18.50	1.242
(C)	30.05	31.15	4.85	14.60	19.35	**1.470**
(D)	19.15	42.25	5.30	15.15	18.15	1.432
(E)	18.60	39.20	4.90	20.50	16.80	1.452

that of (A). The percentage of people in their twenties, which is high in the original dataset (A), becomes significantly lower in (C) to (E). Besides, the percentage of people in their sixties and that of over eighty people, which are very low in (A), becomes higher in (C) to (E). These results demonstrate the effectiveness of the proposed method. In contrast, the entropy of (B) is worse than the original dataset (A). This is due to the inability to specify SA, which is solved in the proposed method. However, even in (C), (D), and (E), the percentage of some age groups is far from the ideal value, i.e., 20%. To solve this problem, we need to improve the acquisition method of *related texts*.

V. CONCLUSION

In this paper, we proposed a method for reducing the implicit class imbalance in unlabeled datasets by explicitly specifying SA by text. To achieve this, we first extract SA-related visual features using CLIP and then utilize them to cluster the target imbalanced dataset. Then, we train a CDM, regarding the cluster IDs as pseudo-labels, and finally generate the same number of images from each cluster to obtain a balanced dataset. We evaluated the proposed method on a face dataset, where race and age are specified as SA. As a result, the proposed method successfully reduced the implicit class imbalance. In addition, the quality of the generated images was much improved compared to our previous method [9]. However, the imbalance reduction capability of the proposed method is still limited. To improve this, we will update the method of obtaining *related texts* of SA in future work. We will also test the impact of the proposed method on the performance of downstream tasks, i.e., the avoidability of AI's biased outputs. This study is partially supported by JST CREST Grant (JPMJCR20D3).

REFERENCES

[1] A. S. Luccioni, C. Akiki, M. Mitchell, and Y. Jernite, "Stable bias: Evaluating societal representations in diffusion models," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, pp. 56 338–56 351, 2023.

[2] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. the 1st Conf. on Fairness, Accountability and Transparency*, vol. 81, 2018, pp. 77–91.

[3] D. Elreedy and A. F. Atiya, "A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance," *Information Sciences*, vol. 505, pp. 32–64, 2019.

[4] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. 2008 IEEE Int'l Joint Conf. on Neural Networks*, 2008, pp. 1322–1328.

[5] D. Devi, S. K. Biswas, and B. Purkayastha, "A review on solution to class imbalance problem: Undersampling approaches," in *Proc. of 2020 Int'l Conf. on Computational Performance Evaluation (ComPE)*, 2020, pp. 626–631.

[6] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. of 2017 IEEE International Conf. on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[7] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9268–9277.

[8] M. Hayat, S. Khan, S. W. Zamir, J. Shen, and L. Shao, "Gaussian affinity for max-margin class imbalanced learning," in *Proc. 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV)*, 2019, pp. 6468–6478.

[9] K. Suyama and K. Nakamura, "Detoxification of unlabeled dataset: reducing implicit class imbalance using pseudo-jacobian of gan's generator," in *Proc. 31st Int'l Conf. on Multimedia Modeling (MMM)*, 2025, pp. 277–290.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. 38th Int' l Conf. on Machine Learning (ICML)*, vol. 139, 2021, pp. 8748–8763.

[11] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *Proc. NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021.

[12] L. Wei-Chao, T. Chih-Fong, H. Ya-Han, and J. Jing-Shang, "Clustering-based undersampling in class-imbalanced data," *Neurocomputing*, vol. 409–410, pp. 17–26, 2017.

[13] Y. Hou, B. Li, L. Li, and J. Liu, "A density-based under-sampling algorithm for imbalance classification," *Journal of Physics: Conference Series*, vol. 1302, no. 2, pp. 1–10, 2019.

[14] J. Joo and K. Kärkkäinen, "Gender slopes: counterfactual fairness for computer vision models by attribute manipulation," in *Proc. Int' l Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia*, 2020, pp. 1–5.

[15] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. 27th Int'l Conf. on Neural Information Processing Systems (NeurIPS)*, 2014, pp. 2672–2680.

[16] G. A. Miller, "Wordnet: a lexical database for english." *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 6840–6851, 2020.

[18] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[19] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5810–5818.

[20] D. Pelleg and A. Moore, "X-means: extending k-means with efficient estimation of the number of clusters," in *Proc. 17th Int'l Conf. on Machine Learning (ICML)*, 2000, pp. 727–734.