k-SALSA: *k*-anonymous synthetic averaging of retinal images via local style alignment

Minkyu Jeon[1,2], Hyeonjin Park[5,*], Hyunwoo J. Kim[2,†], Michael Morley[3,4,†], and Hyunghoon Cho[1,†]

[1] Broad Institute of MIT and Harvard, Cambridge, MA, USA
{mjeon, hhcho}@broadinstitute.org
[2] Korea University, Seoul, Republic of Korea
hyunwoojkim@korea.ac.kr
[3] Harvard Medical School, Boston, MA, USA
[4] Ophthalmic Consultants of Boston, Boston, MA, USA
mgmorley@eyeboston.com
[5] NAVER CLOVA, Seoul, Republic of Korea
hyeonjin.park.ml@navercorp.com

Abstract. The application of modern machine learning to retinal image analyses offers valuable insights into a broad range of human health conditions beyond ophthalmic diseases. Additionally, data sharing is key to fully realizing the potential of machine learning models by providing a rich and diverse collection of training data. However, the personally-identifying nature of retinal images, encompassing the unique vascular structure of each individual, often prevents this data from being shared openly. While prior works have explored image de-identification strategies based on synthetic averaging of images in other domains (e.g. facial images), existing techniques face difficulty in preserving both privacy and clinical utility in retinal images, as we demonstrate in our work. We therefore introduce k-SALSA, a generative adversarial network (GAN)-based framework for synthesizing retinal fundus images that summarize a given private dataset while satisfying the privacy notion of k-anonymity. k-SALSA brings together state-of-the-art techniques for training and inverting GANs to achieve practical performance on retinal images. Furthermore, k-SALSA leverages a new technique, called local style alignment, to generate a synthetic average that maximizes the retention of fine-grain visual patterns in the source images, thus improving the clinical utility of the generated images. On two benchmark datasets of diabetic retinopathy (EyePACS and APTOS), we demonstrate our improvement upon existing methods with respect to image fidelity, classification performance, and mitigation of membership inference attacks. Our work represents a step toward broader sharing of retinal images for scientific collaboration. Code is available at https://github.com/hcholab/k-salsa.

Keywords: Medical image privacy, k-anonymity, generative adversarial networks, fundus imaging, synthetic data generation, style transfer

[*] This work was performed while the author was at Korea University

[†] Corresponding authors

1 Introduction

Retinal imaging is a fast, non-invasive, and cost-effective platform to study a range of systemic diseases, e.g. cardiovascular and neurological disorders [46]. Recent advances in machine learning (ML) are accelerating this transformation, equipping researchers with tools to recognize clinically relevant biomarkers across diverse imaging modalities, such as fundus imaging and optical coherence tomography (OCT). Studies have demonstrated the effectiveness of deep learning models in predicting clinical traits such as cardiovascular risk factors as well as other health-related information such as age, sex, and smoking status [26,36,49].

However, privacy concerns prevent the sharing of retinal images, presenting a hurdle for ML in ophthalmology [43,44]. Despite not being legally recognized as a biometric identifier in certain cases (e.g. HIPAA [45]), retinal images are widely regarded as sensitive because they include individual-specific patterns like blood vessel structure [28]. Reflecting these concerns, medical institutions have begun to refrain from using retinal images in grand rounds, lectures, and publications, leading to difficulties in research and education. We aim to tackle these challenges by transforming retinal images to protect privacy while preserving clinical utility.

To this end, a seminal work by Newton et al. [33] on face de-identification introduced a class of techniques known as the "k-Same" algorithms, wherein mutually disjoint clusters of k images in the dataset are individually replaced with a single representative synthetic image that summarizes the visual characteristics of the images in each cluster. This naturally leads to a synthetic dataset satisfying the classical privacy notion of k-anonymity [41], which requires that each data instance in the released dataset cannot be distinguished among at least k underlying individuals. k-anonymity has been widely considered in the medical literature as a meaningful privacy notion and has been used as a core principle in real-world systems and polices [10,12,19]. Furthermore, recent successes of generative adversarial networks (GANs) [13,15] in synthesizing realistic images in diverse domains suggest a promising approach for generating high-quality representatives of individual clusters by taking an average of images in the latent embedding space of a GAN, also known as the k-Same-Net algorithm [30].

Despite the promise of these approaches, several key challenges remain in applying these methods to retinal images. First, while several works have explored the use of GANs to generate synthetic retinal images [4,5,6,7,34], none to our knowledge have addressed the problem of effectively summarizing these images in the latent space, making the feasibility of this approach for retinal images an unknown. Next, the difficulty of capturing fine-grain visual patterns of retinal images (e.g. hemorrhages or lipid deposits) poses an additional challenge in preserving the clinical utility of these images. Finally, because k-anonymity does not directly imply privacy (individual images could potentially be inferred from the average), a direct evaluation of privacy offered by k-anonymity in the context of retinal images is needed before these tools may be used in practice [35].

To address these challenges, we developed k-SALSA, an end-to-end pipeline for synthesizing a k-anonymous dataset given a private dataset of retinal images. We modernize the approach of k-Same-Net to use state-of-the-art techniques for training and inverting GANs. This allows us to map the source images to an embedding space, "average" them, and generate a representative synthetic image. We improve upon the existing methodology of taking the Euclidean average of embedding vectors by introducing a new technique called *local style alignment*, which aims to maximize the retention of local texture information from the source images. This ensures that the output keeps clinically relevant features.

We evaluate our pipeline on two benchmark datasets (APTOS and EyePACS) and demonstrate the enhanced visual fidelity of the synthetic images generated by our approach with respect to the Fréchet inception distance [18], a standard quality metric for synthetic images. We also show that the synthetic dataset generated by our approach enables accurate training of downstream classifiers for predicting varying degrees of diabetic retinopathy. Lastly, we evaluate the privacy of our approach with respect to membership inference attacks (MIA), where an adversary tries to predict whether a given retinal image was part of the cluster represented by a specific synthetic image. Our results show that synthetic images generated by k-SALSA provide strong mitigation of MIA, while prior k-Same approaches using pixel-wise or eigenvector-based averaging fail to do so, despite the fact that all of these approaches ostensibly satisfy k-anonymity.

Summary of our contributions. (1) We demonstrate the feasibility of GAN-based k-anonymization of retinal images. (2) We present a modernized k-Same algorithm using state-of-the-art GAN techniques, which are crucial for practical performance. (3) We introduce a novel technique—local style alignment—for generating a synthetic average with enhanced fidelity and downstream utility. (4) We perform comprehensive experiments on two datasets, evaluating the fidelity, utility, and privacy of our method compared to existing techniques.

2 Related Work

Traditional approaches for removing identifying features from private images (e.g., faces and medical images) involve direct manipulation of pixels, including masking, blurring, and pixelation [38,3,31,39]. However, these heuristics have been found to provide insufficient privacy protection [1,37]. In response, a *learning*-based approach to de-identification has been increasingly studied [14]. This is enabled by recent advances in GANs, which intuitively provide a more powerful approach to manipulate images according to their intrinsic manifold [15].

The existing literature on GAN-based transformation of images for privacy protection largely focuses on face de-identification. A common approach for formalizing privacy protection in this problem is to combine multiple images to obtain k-anonymity through the k-Same framework [21,30,33]. A key difficulty in these works has been generating high-quality images that capture useful information in the original image. To this end, recent works have focused on developing techniques to disentangle and preserve non-identity attributes of the image, such as pose and facial expression [20,21,29,50]. However, these methods are not directly applicable to our setting given the unclear distinction between identity vs. non-identity features in retinal images beyond the blood vessel structure.

GAN-based approaches to generate images with differential privacy [8] have also been proposed [27,52], but current techniques lead to significant degradation of image quality (see supplement for an example application to retinal images).

Several recent works have successfully explored the use of GANs for generating realistic retinal images. Niu et al. [34] proposed a method to generate an image consistent with the given pathological descriptors. Both Zhou et al. [54] and Chen et al. [6] developed GAN models to synthesize retinal images conditioned on a semantic segmentation to improve disease classification performance. Yu et al. [53] introduced multi-channel GANs that improve the quality of generated retinal images by separately considering different elements of the image, including the blood vessels and the optic disc.

3 Method

3.1 Overview of k-SALSA

We consider a dataset D of retinal fundus images that the user wishes to release in a privatized form. We assume that the user has access to an auxiliary dataset D_0 that can be used to pre-train the GAN components of k-SALSA. Note that D_0 could simply be a publicly available dataset like the ones we used in our work or a subset of D. This choice does not affect the k-anonymity property of k-SALSA. Given these datasets, k-SALSA proceeds in four steps:

1. **Pre-training.** We first train a GAN on D_0. Let $G(\cdot)$ be the trained generator, which maps a latent code $w \in \mathcal{W}$ to a synthetic image $G(w)$. This step intuitively constructs the latent embedding space $\mathcal{W}$, which we will later rely on for the averaging operation. Next, k-SALSA trains a GAN inversion model $E(\cdot)$ (can be viewed as an encoder), which maps an image $x \in \mathcal{X}$ to a particular latent code $E(x) \in \mathcal{W}$. Note that this inversion process is lossy in that the synthetic image $G(E(x))$ will only be approximately similar to the original image x, constrained by the limitations of encoder E and generator G. Together, these two functions approximate a bijection between the space of retinal images $\mathcal{X}$ and the latent embedding space $\mathcal{W}$.
2. **Clustering.** Next, k-SALSA performs *same-size clustering* of the target input dataset D based on the inverted codes $E(x)$ for each $x \in D$, partitioning D into groups of exactly k similar images. Here k is the user parameter determining the k-anonymity of the final output. If the size of D is not divisible by k, one can disregard a small number of samples to resolve the issue, given that k is typically small (e.g. 10).
3. **Averaging.** For each cluster, k-SALSA summarizes the k source images as a single representative image via local style alignment—our new approach. This leads to k-anonymity, since each average image represents k individuals as a whole and does not distinguish among them.
4. **Release.** Finally, k-SALSA constructs a k-anonymous synthetic dataset $\tilde{D}$ to release by associating each average image (one per cluster) with the aggregated labels of the k images in the corresponding cluster (if labels were

provided in the input dataset). This synthetic dataset can then be used for downstream analysis, such as training a classifier to predict the labels.

A graphical illustration of our workflow is provided in Fig. 1.

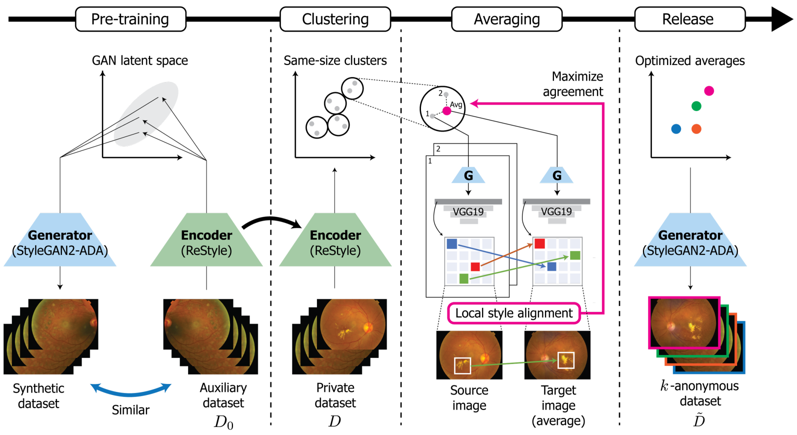

Fig. 1: **Workflow of k-SALSA.** GAN generator and inversion encoder are first trained to be used in subsequent steps. Same-size clustering groups images into groups of k, then the representative of each cluster is optimized via our local style alignment approach to preserve salient visual patterns. Images are synthesized from the optimized averages and released. Avg: average, G: generator

3.2 Local Style Alignment: Our New Approach to Image Averaging

While prior k-Same approaches developed for facial images have considered the Euclidean average of latent codes within each cluster as the representative, we found that this straightforward approach leads to significant loss of detail in synthetic retinal images, where clinically relevant patterns such as hemorrhages and exudates are often omitted (see Fig. 2).

We make the following two observations toward addressing this key limitation. First, unlike facial images where the salient structural features (e.g. eyes and nose) generally appear in consistent regions within the image, which facilitates the disentanglement of latent features, important patterns in retinal images can appear in different areas and thus are more easily diluted when averaging the features in the latent space. Second, in contrast to the importance of shape information in facial images, the patterns of interest in retinal images that are not directly linked to personal identity tend to be associated with *texture-level* information (e.g. colored dots of varying granularity and frequency). In fact, the geometric structure of the blood vessels is a prominent identifying feature of concern that we are interested in obfuscating in the image via averaging.

Algorithm 1 k-SALSA

```
Input: Private dataset $X = (x_1, \dots, x_n)$, auxiliary dataset $X_0$ for GAN model training, integer $k > 1$ (assume $n = mk$ for integer $m$ without loss of generality), number of iterations $T$, loss ratio parameter $\lambda$
Output: Synthetic dataset $\tilde{X}$ of size $m$ with $k$-anonymity
Train a GAN generator $G$ and a GAN inversion encoder $E$ on $X_0$
Obtain latent code $w_i = E(x_i)$ for each $i \in [n]$ and let $W = \{w_i\}_{i=1}^{n}$
$(C_1, \dots, C_m) = \mathsf{SameSizeClustering}(W, k)$ ▷ $C_j \subset W$, $|C_j| = k$, $|C_j \cap C_{j' \neq j}| = 0$, $\forall j$
Initialize $\tilde{X} = \emptyset$
for each cluster $j \in [m]$ do
    Let $C_j = (w'_1, \dots, w'_k)$, and $x'_i$ the original image of $w'_i$ for each $i$
    Compute $w_0 = \frac{1}{k} \sum_{i=1}^{k} w'_i$ and generate $x_0 = G(w_0)$
    Initialize $w_{\text{avg}}^{(0)} = w_0$
    for each iteration $t \in [T]$ do
        Generate $x_{\text{avg}}^{(t-1)} = G(w_{\text{avg}}^{(t-1)})$
        Compute content loss $\mathcal{L}_{\text{content}}(x_0, x_{\text{avg}}^{(t-1)})$ using Eq. 6
        Compute local style alignment loss $\mathcal{L}_{\text{style}}((x'_1, \dots, x'_k), x_{\text{avg}}^{(t-1)})$ using Eq. 7
        Compute total loss $\mathcal{L}_{\text{total}} = \lambda \mathcal{L}_{\text{content}} + (1 - \lambda) \mathcal{L}_{\text{style}}$
        Update $w_{\text{avg}}^{(t)}$ using $w_{\text{avg}}^{(t-1)}$ and the gradient $\nabla_{w_{\text{avg}}^{(t-1)}} \mathcal{L}_{\text{total}}$
    end for
    Add $G(w_{\text{avg}}^{(T)})$ to $\tilde{X}$
end for
return $\tilde{X}$
```

Our local style alignment technique takes advantage of these observations to obtain higher quality representative images of each cluster. We first draw the connection between texture patterns of interest and the "style" of the image from the style transfer literature [11]. Since we are interested in local texture patterns in the image, we capture the *local style features* by constructing the feature covariance matrix in a sliding window of image patches, rather than over the whole image. We then consider the correspondence of the local style features between the source images in the cluster and the target representative image, allowing for source texture patterns to appear in different locations in the target and simultaneously enforcing that these patterns are recapitulated somewhere in the target image. Optimizing the latent code for the representative image with respect to the augmented loss function considering both the local style features and general content similarity, we obtain an enhanced representative image for each cluster. We provide the details of each step below and in Algorithm 1.

Construction of local style features. Following the approach of style transfer, we view style and texture information as being captured by the cross-channel feature correlations in the intermediate layers of a pre-trained convolutional neural network (CNN), such as VGG19 [40].

Formally, let $F_{\text{source}}^{(i)}$ and F_{target} be n-by-n-by-c tensors for the i-th source image and the target representative, respectively, representing the activation output of an n-by-n intermediate CNN layer across c channels. Note that we use the second layer of VGG19 in all our experiments. We spatially partition the activation output into a grid of p submatrices, $\{F^{(i)}_{\text{source},j}\}_{j=1}^{p}$ and $\{F_{\text{target},j}\}_{j=1}^{p}$, each corresponding to a local patch in the image. For each patch j, we define the local style features $S^{(i)}_{\text{source},j}$ and $S_{\text{target},j}$ as c-by-c matrices, where

$$(S^{(i)}_{\text{source},j})_{u,v} := \langle \mathsf{Vec}((F^{(i)}_{\text{source},j})_{:,:,u}), \mathsf{Vec}((F^{(i)}_{\text{source},j})_{:,:,v})\rangle \tag{1}$$

$$(S_{\text{target},j})_{u,v} := \langle \mathsf{Vec}((F_{\text{target},j})_{:,:,u}), \mathsf{Vec}((F_{\text{target},j})_{:,:,v})\rangle \tag{2}$$

and $\mathsf{Vec}(\cdot)$ denotes vectorization, $\langle\cdot,\cdot\rangle$ denotes dot product, and $M_{:,:,u}$ for a tensor M denotes a slice corresponding to channel u. The sets $\{S^{(i)}_{\text{source},j}\}_{j=1}^{p}$ for each source image i in the cluster and $\{S_{\text{target},j}\}_{j=1}^{p}$ for the target fully characterize the local texture information we aim to capture in our model.

Alignment of local style features. To introduce flexibility in determining where a visual pattern from the source image appears in the target image, we quantify the agreement in local style features via a correspondence. Inspired by the recent work of Wang et al. on dense contrastive learning [47], we compute the cosine similarity of style features between every pair of patches between the source and the target and take the optimal match for each patch in the source image to be included in the loss function. This induces positional flexibility while penalizing complete omission of texture patterns from the source. Note that the prior work [47] did not consider style information in their approach.

We define correspondence index $a(i,j)$ for patch j in the source image i as:

$$a(i,j) := \arg\max_{j' \in \{1,\ldots,p\}} \mathsf{CosineSimilarity}(\mathsf{Vec}(S^{(i)}_{\text{source},j}), \mathsf{Vec}(S_{\text{target},j'})). \tag{3}$$

It is worth noting that, while a naïve implementation of all pairwise comparison of patches leads to significant runtime overhead, our implementation efficiently utilizes matrix operations to maintain computational efficiency.

Optimization of representative images. To synthesize an informative representative image for each cluster, leveraging the correspondence of local style features, we frame the process as an optimization problem as follows.

Let $E(\cdot)$ and $G(\cdot)$ be the encoder of the pre-trained GAN inversion model and the pre-trained GAN generator, respectively. We directly optimize the target latent code w_{avg} whose corresponding image $G(w_{\text{avg}})$ is the desired representative of the cluster. We initialize w_{avg} to the baseline Euclidean average w_0 of the source image embeddings given by

$$w_0 := \frac{1}{k}\sum_{i=1}^{k} E(x^{(i)}), \tag{4}$$

where $x^{(i)}$ denotes the i-th image in the cluster. We then iteratively optimize the solution using gradient descent with the loss function

$$\mathcal{L}_{\text{total}} = \lambda\mathcal{L}_{\text{content}} + (1-\lambda)\mathcal{L}_{\text{style}}, \tag{5}$$

where λ determines the ratio between the two terms given by

$$\mathcal{L}_{\text{content}} = 1 - \langle F(G(w_0)), F(G(w_{\text{avg}}))\rangle, \tag{6}$$

$$\mathcal{L}_{\text{style}} = \sum_{i=1}^{k}\sum_{j=1}^{p} \| S^{(i)}_{\text{source},j} - S_{\text{target},a(i,j)}(w_{\text{avg}}) \|^2_{\mathcal{F}}. \tag{7}$$

Note that $\|\cdot\|_{\mathcal{F}}$ denotes the Frobenius norm, and $F(\cdot)$ represents a pre-trained encoder network, which we use to induce high-level similarity between the optimized representative and the Euclidean average to avoid degenerate cases and to prioritize refining of the baseline solution. In our experiments, we set F to the pre-trained MoCo network [16], which was trained on the ImageNet dataset [9] via unsupervised contrastive learning. MoCo is recognized for its effectiveness in capturing semantic information of natural images beyond the available labels in the original dataset. We also note that, although style transfer approaches typically take many iterations to converge, our initialization scheme using the Euclidean average greatly simplifies this process, requiring fewer iterations to obtain the final solutions in our experiments.

3.3 GAN-based Image Generation and Encoding

We train a GAN model, StyleGAN2-ADA [22], on retinal images to learn to generate realistic fundus images from a latent vector space. The StyleGAN family of methods [22,23,24] generate high-resolution images using a progressive architecture, where increasingly fine-grain details are added to the image as we get deeper into the network. We consider the latent space associated with this network to be extended multi-scale $\mathcal{W}$, which modulates the activation of units in all layers of the generator hierarchy. StyleGAN2-ADA is one of the latest in this class, which stabilizes training on limited data using the *adaptive discriminator augmentation* (ADA) mechanism. Likely because the size of the public retinal image datasets is small compared to other types of image datasets, we observed that the use of ADA leads to a considerable improvement in image quality. We also note that the existing work on GANs for retinal images (e.g. [6,34,53,54]) leverages additional labeled information such as vessel segmentation, and thus are not directly applicable to our setting where we use the raw fundus images.

To manipulate and summarize real retinal images in the latent space equipped with a generator to synthesize new images, we need an encoder to map a given image to the GAN latent space, a task known as GAN inversion [51]. In our framework, we use this encoder to invert every image in the input dataset, then use the latent codes both to define the clusters of size k to be averaged and to find the Euclidean centroid for the cluster to use as an initialization point, as described in the previous sections. To this end, we use ReStyle [2], a recently developed approach to GAN inversion which achieved a significant scalability improvement over the previous methods by adopting an iterative refinement approach. The ReStyle model takes the target image and the current synthetic image (the result of inversion) as input, and learns to generate an update to the latent code that improves consistency between the two images. For k-SALSA, adopting this approach was key to building a practical pipeline—it reduced the inversion time by an order of magnitude (from 80 to 3 seconds per image).

While the generator and the encoder individually draws from prior work, we note that the combination of these state-of-the-art techniques have not been previously studied in the context of privatizing retinal images and were in fact key enabling factors of the practical performance of k-SALSA in our experiments.

3.4 Same-Size Clustering

To partition the input dataset into groups of exactly k images to average, we employ a greedy nearest neighbor clustering to the inverted latent codes of the input images. At each iteration, a point with the maximum average distance to the rest of the dataset is chosen, with the goal of prioritizing outliers. Then $k-1$ nearest neighbors of the chosen point are identified to form a new cluster of size k. The points in the new cluster are removed from the dataset and the above process is repeated. In our experiments, we downsample the dataset to a multiple of k to avoid a leftover cluster smaller than k. Our experiments show the effectiveness of this efficient clustering approach in downstream tasks.

4 Experiments

4.1 Benchmark Datasets and Evaluation Setting

Our experiments are conducted on public fundus image datasets APTOS[6] and EyePACS[7], widely used for diabetic retinopathy (DR) classification. The images in both datasets are labeled by ophthalmologists with five grades of DR based on severity: 0 (normal), 1 (mild DR), 2 (moderate DR), 3 (severe DR), and 4 (proliferative DR). EyePACS images were acquired from different imaging devices, leading to variations in image resolution, aspect ratio, intensity, and quality. Hence, EyePACS represents a more challenging evaluation setting.

For both datasets, we train the GAN generator and the GAN inversion model on the training set (see supplement for details). We then apply k-SALSA to the training set with the pre-trained GAN models to generate a k-anonymous dataset of synthetic images with aggregated labels. To evaluate the downstream utility of the synthetic dataset, we trained DR classifiers based on the synthetic images and evaluated the classifiers on the test set using real images and labels.

4.2 Baseline Approaches

We compare k-SALSA with the following baseline methods. To demonstrate the advantage of our novel local style alignment-based averaging scheme, we consider the same method as k-SALSA, except using the Euclidean average (centroid)

[6] https://www.kaggle.com/c/aptos2019-blindness-detection

[7] https://www.kaggle.com/c/diabetic-retinopathy-detection

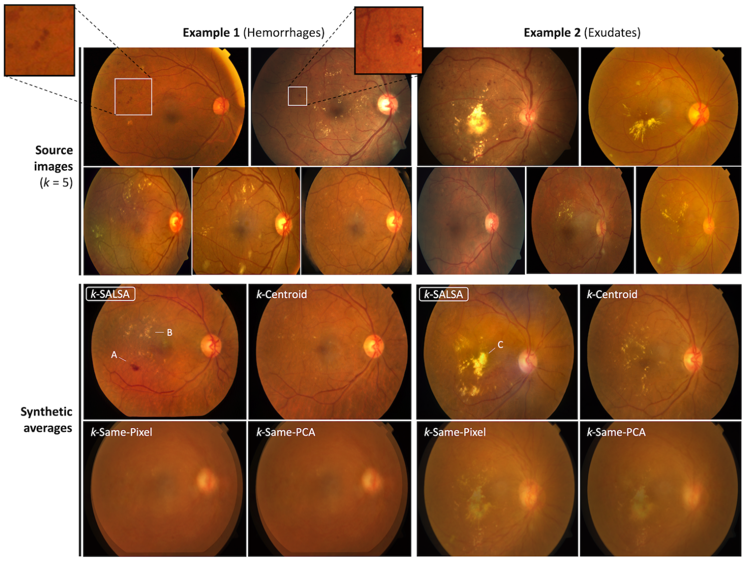

Fig. 2: **Examples of synthetic average of retinal images.** Two example clusters ($k = 5$) of real retinal images (*top*) along with synthetic averages generated by different methods (*bottom*). k-SALSA better captures clinically relevant features such as hemorrhages (A) and exudates (B and C)

of each cluster in our GAN latent space to generate the representative image (**k-Centroid**). To illustrate the importance of GANs in synthetic averaging of retinal images, we also evaluate less sophisticated schemes based on pixel-wise averaging (**k-Same-Pixel**) and averaging in the low-dimensional latent space obtained by principal components analysis (**k-Same-PCA**). Note that k-Same-PCA is equivalent to the method proposed in the original work on k-Same algorithms [33], and k-Centroid represents the best achievable performance following the general framework of k-Same-Net [30] leveraging our GAN approaches. We applied all averaging methods to the same set of clusters we constructed using the latent space of k-SALSA. For some comparisons, we also consider the performance based on the non-averaged synthetic images generated from the inversion of each original image, i.e. $G(E(x))$ given an image x (**GAN-Inverted**).

4.3 Fidelity of Synthetic Images

To evaluate the visual quality of synthetic retinal images, we first compare the Fréchet inception distance (FID) [18], a standard performance metric for images generated using GANs, across the methods we considered. Note that, unlike pixel-wise metrics such as PSNR and SSIM [48] (see supplement for additional discussion), FID measures the divergence between the multivariate Gaussian distributions induced by the real vs. synthetic images in the activation of the Inception V3 model [42] trained on ImageNet [9]. Intuitively, FID quantifies how different the synthetic images are from a reference set of real images in a manner that reflects human visual perception. We use the original retinal images from each dataset as the reference to calculate FID on the corresponding synthetic images. As shown in Table 1, k-SALSA consistently obtains the best (the lowest) FID among all averaging methods for different values of k (2, 5, and 10) on both datasets. For all methods, averaging leads to worse FID relative to GAN-Inverted images, with the gap increasing as k becomes larger. This suggests that generating a realistic image becomes more difficult as we average more images. Nevertheless, k-SALSA's FID remains closest to GAN-Inverted even for $k = 10$.

Table 1: Comparison of fidelity of synthetic images

Method	**Metric**	**APTOS**			**EyePACS**		
		$k = 2$	$k = 5$	$k = 10$	$k = 2$	$k = 5$	$k = 10$
GAN-Inverted	FID	11.47	18.12	25.45	9.65	14.65	19.77
k-Same-Pixel	FID	128.23	131.69	138.89	65.21	113.45	144.13
k-Same-PCA	FID	131.691	128.225	138.891	159.57	164.69	173.15
k-Centroid	FID	12.7	22.71	31.09	11.24	20.58	28.49
k-SALSA	FID	**11.84**	**20.09**	**28.4**	**9.95**	**15.07**	**21.28**

In Fig. 2, we provide examples of synthetic averages generated by different methods for $k = 5$ for visual comparison. k-SALSA images more clearly capture the fine-grain, clinically-relevant patterns in the source images, including exudates (appearing as grainy yellow patches) and hemorrhages (appearing as dark spots), both of which are well-established biomarkers of diabetes [32]. k-Centroid generates realistic images, but tend to omit important fine-grain patterns, which initially motivated this work. k-Same-Pixel and k-Same-PCA lead to low-fidelity images that even fail to align the boundaries of the photographs due to their linearity. Examples for other values of k are provided in the supplement.

4.4 Downstream Classification Performance

In addition to generating more realistic and informative summaries of each cluster of retinal images, we are interested in enabling downstream analysis with our synthetic data. We tested whether k-SALSA's synthetic dataset can lead to accurate classifiers of clinical labels, in our case the grading of diabetic retinopathy (DR). In EyePACS, we evaluated normal vs. DR binary classification due to the highly imbalanced number of labels (in contrast to the five-class setting in Kaggle). We tested multi-class prediction with all five labels in APTOS.

The number of images in the training set was 3,000 for APTOS and 10,000 for EyePACS, where the latter was subsampled for efficiency. For each set, we used our clustering approach to obtain same-size clusters for each of $k \in \{2, 5, 10\}$, which were then individually averaged to obtain training images for a classifier.

We also evaluated the classifiers trained on original or GAN-Inverted images, which were subsampled to the same number of training examples as the synthetic datasets for each k for comparison. We provide experimental details and a comparison without subsampling in the supplement.

Table 2: Comparison of diabetic retinopathy classification performance

Method	**Metric**	**APTOS**			**EyePACS**		
		$k = 2$	$k = 5$	$k = 10$	$k = 2$	$k = 5$	$k = 10$
Original	Accuracy	0.771	0.752	0.700	0.794	0.767	0.725
	Cohen's κ	0.903	0.888	0.856	0.414	0.327	0.297
GAN-Inverted	Accuracy	0.744	0.702	0.651	0.731	0.730	0.708
	Cohen's κ	0.865	0.828	0.814	0.140	0.140	0.058
k-Same-Pixel	Accuracy	0.502	0.318	0.366	0.361	0.296	0.434
	Cohen's κ	0.663	0.273	0.140	0.029	0.000	0.043
k-Same-PCA	Accuracy	0.572	0.394	0.293	0.3360	0.361	0.657
	Cohen's κ	0.651	0.559	0.253	0.010	0.029	0.096
k-Centroid	Accuracy	**0.688**	0.688	0.646	0.680	0.664	0.611
	Cohen's κ	**0.786**	0.745	0.647	**0.268**	0.185	0.160
***k*-SALSA**	Accuracy	0.687	**0.712**	**0.673**	**0.704**	**0.688**	**0.705**
	Cohen's κ	0.773	**0.769**	**0.710**	0.254	**0.222**	**0.225**

The results in Table 2 show that the k-SALSA synthetic datasets generally outperform the alternative approaches with respect to both accuracy and Cohen's κ statistic (with quadratic weighting) on the test set. k-Centroid achieves slightly better performance for $k = 2$, but remains comparable to our approach. Since Euclidean averaging may introduce greater distortions for larger values of k, we expect the advantage of k-SALSA to be more pronounced for moderate to large k, which is consistent with our results. As expected, performance based on original images is higher than the synthetic dataset, but part of this gap is ascribed to the limitations of the current GAN models as suggested by the lower performance of the non-averaged, GAN-inverted images compared to the original. Interestingly, for EyePACS $k = 10$, k-SALSA outperforms GAN-Inverted, suggesting that summarizing salient features may even be beneficial for classification when the data is limited. We include in the supplement additional results illustrating the impact of k and a promising extension of k-SALSA which uses data augmentation to mitigate dataset reduction due to averaging.

4.5 Mitigation of Membership Inference Attacks

To compare the privacy properties of the methods, we implemented a membership inference attack (MIA), where an adversary holding a synthetic dataset attempts to infer whether a target person was part of a specific cluster. We trained ResNet18 [17] on the training set to classify cluster membership using the synthetic averages generated by each method. We then evenly divided the test set into two parts, generated cluster averages on the first, then evaluated the performance of the classifier in ranking the images in *both* test sets for membership in each cluster, based only on its synthetic average. For each cluster size k, we calculated the top-K accuracy (i.e., mean fraction of top K samples in the ranking that correspond to correct guesses) with $K = k$.

Table 3: Membership inference attack top-K accuracy (%) with $K = k$.

Method	**APTOS**			**EyePACS**		
	$k = 2$	$k = 5$	$k = 10$	$k = 2$	$k = 5$	$k = 10$
k-Same-Pixel	100.0	91.76	84.21	97.96	94.28	86.67
k-Same-PCA	98.98	84.29	2.5	98.98	86.21	2.63
k-Centroid	78.57	41.42	2.1	88.63	45.92	2.63
k-SALSA	**77.55**	**40.0**	**1.0**	**71.42**	**35.17**	**0.52**

The results are summarized in Table 3. Note that an adversary with access to the encoder would achieve an expected accuracy of 50% for all values of k, since in any neighborhood a random half of the samples correspond to negative matches that were not included in the private dataset. This represents a realistic scenario where the attacker does not have *a priori* knowledge of individuals in the private dataset. For a worst-case evaluation, we assume that the target's image is identical to the one in the dataset; any protection offered by noisy re-acquisition of images is likely to be bypassed with more sophisticated MIA (e.g., using vessel structures). We observed that pixel-averaging provides little to no privacy. k-Same-PCA and k-Centroid lower the risks, the latter to a greater extent. k-SALSA results in the strongest mitigation with MIA accuracy of 1% and 0.52% for APTOS and EyePACS, respectively, for $k = 10$. Our improvement over k-Centroid is likely due to the fact that the addition of our local style loss prioritizes similarity in high-level visual patterns over low-level content, potentially reducing the amount of identity-related information that can be exploited by the attack. None of the methods provide strong privacy at $k = 2$, which reflects an insufficient amount of variability between the two source images that could be leveraged for privacy; however, we expect our approach to provide meaningful privacy protection for larger values of k as our results show.

4.6 Ablation Study

We conducted an ablation study to evaluate the importance of individual components of our methodology. Recall that the loss function of k-SALSA includes the content loss and the local style loss (see Eq. 5). We considered four alternative models with: only style loss, only content loss, both but using global style features computed over the whole image, and both without the flexible alignment (i.e., each local style is directly compared to that of the corresponding patch in the other image at the same location). All of these alternatives performed considerably worse than k-SALSA in downstream classification performance (Table 4). The especially poor performance without alignment suggests that enforcing style preservation without spatial flexibility can in fact be harmful for the method.

Table 4: Ablation study (APTOS, $k = 5$)

Method	**Accuracy**	**Cohen's κ**
Style loss only (local, with alignment)	0.685	0.735
Content loss only	0.673	0.712
Content loss, Global style loss	0.687	0.761
Content loss, Local style loss, No alignment	0.57	0.417
***k*-SALSA**	**0.712**	**0.769**

5 Discussion and Conclusions

We presented k-SALSA, an end-to-end pipeline for synthesizing a k-anonymous retinal image dataset given a private input dataset. We leverage local style alignment, our new approach for summarizing source images in a cluster while preserving local texture information. Our results demonstrate that k-anonymization of retinal images, preserving both privacy and clinical utility, is feasible.

We would like to address several limitations of the current method in future work. First, k-SALSA's performance is dependent on the quality of the underlying GAN and GAN inversion models. We plan to devise strategies tailored to retinal images (e.g., separately modelling different parts of the image) to further improve GAN models. Next, we plan to explore more rigorous frameworks for privacy such as differential privacy (DP) [8]. While it is generally difficult to apply DP to high-dimensional data such as images, certain relaxations of DP [25] may lead to a practical solution. Lastly, we plan to explore the application of our methodology to other imaging modalities for the retina, including the OCT.

Our work demonstrates that domain-inspired techniques can be combined with the state-of-the-art GAN techniques to design effective approaches to privatizing sensitive data. The methodological insights introduced by our work is of general interest to other domains (e.g. genomics), where privacy-aware aggregation of sensitive data may overcome challenges in data sharing.

Acknowledgements. M.J. is supported by the Ministry of Trade, Industry, and Energy in Korea, under Human Resource Development Program for Industrial Innovation (Global) (P0017311) supervised by the Korea Institute for Advancement of Technology. H.C. is supported by NIH DP5 OD029574-01 and by the Schmidt Fellows Program at Broad Institute.

References

1. Abramian, D., Eklund, A.: Refacing: Reconstructing anonymized facial features using GANS. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) (2019)
2. Alaluf, Y., Patashnik, O., Cohen-Or, D.: Restyle: A residual-based stylegan encoder via iterative refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
3. Bischoff-Grethe, A., Ozyurt, I.B., Busa, E., Quinn, B.T., Fennema-Notestine, C., Clark, C.P., Morris, S., Bondi, M.W., Jernigan, T.L., Dale, A.M., et al.: A technique for the deidentification of structural brain mr images. Human brain mapping (2007)
4. Burlina, P., Paul, W., Liu, T.Y.A., Bressler, N.M.: Detecting Anomalies in Retinal Diseases Using Generative, Discriminative, and Self-supervised Deep Learning. JAMA Ophthalmology (2022)
5. Burlina, P.M., Joshi, N., Pacheco, K.D., Liu, T.Y.A., Bressler, N.M.: Assessment of Deep Generative Models for High-Resolution Synthetic Retinal Image Generation of Age-Related Macular Degeneration. JAMA Ophthalmology (2019)
6. Chen, Y., Long, J., Guo, J.: Rf-gans: A method to synthesize retinal fundus images based on generative adversarial network. Computational intelligence and neuroscience (2021)
7. Coyner, A.S., Chen, J., Campbell, J.P., Ostmo, S., Singh, P., Kalpathy-Cramer, J., Chiang, M.F.: Diagnosability of Synthetic Retinal Fundus Images for Plus Disease Detection in Retinopathy of Prematurity. AMIA Symposium (2020)
8. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. (2014)
9. Fei-Fei, L., Deng, J., Li, K.: Imagenet: Constructing a large-scale image database. Journal of vision (2009)
10. Garfinkel, S., et al.: De-identification of Personal Information:. US Department of Commerce, National Institute of Standards and Technology (2015)
11. 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 (2016)
12. Gkoulalas-Divanis, A., Loukides, G., Sun, J.: Publishing data from electronic health records while preserving privacy: A survey of algorithms. Journal of biomedical informatics (2014)
13. Goodfellow, I.: Nips 2016 tutorial: Generative adversarial networks. arXiv preprint arXiv:1701.00160 (2016)
14. der Goten, V., Alexander, L., Hepp, T., Akata, Z., Smith, K.: Conditional de-identification of 3d magnetic resonance images. arXiv preprint arXiv:2110.09927 (2021)
15. Gui, J., Sun, Z., Wen, Y., Tao, D., Ye, J.: A review on generative adversarial networks: Algorithms, theory, and applications. IEEE Transactions on Knowledge and Data Engineering (2021)
16. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2020)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016)

18. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems (2017)
19. Jakob, C.E., Kohlmayer, F., Meurers, T., Vehreschild, J.J., Prasser, F.: Design and evaluation of a data anonymization pipeline to promote open science on covid-19. Scientific data (2020)
20. Jeong, Y., Choi, J., Kim, S., Ro, Y., Oh, T.H., Kim, D., Ha, H., Yoon, S.: Ficgan: Facial identity controllable gan for de-identification. arXiv preprint arXiv:2110.00740 (2021)
21. Jourabloo, A., Yin, X., Liu, X.: Attribute preserved face de-identification. In: 2015 International conference on biometrics (ICB) (2015)
22. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. Advances in Neural Information Processing Systems (2020)
23. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2019)
24. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2020)
25. Kifer, D., Machanavajjhala, A.: Pufferfish: A framework for mathematical privacy definitions. ACM Transactions on Database Systems (TODS) (2014)
26. Korot, E., Pontikos, N., Liu, X., Wagner, S.K., Faes, L., Huemer, J., Balaskas, K., Denniston, A.K., Khawaja, A., Keane, P.A.: Predicting sex from retinal fundus photographs using automated deep learning. Scientific reports (2021)
27. Long, Y., Wang, B., Yang, Z., Kailkhura, B., Zhang, A., Gunter, C., Li, B.: G-pate: Scalable differentially private data generator via private aggregation of teacher discriminators. Advances in Neural Information Processing Systems (2021)
28. Mariño, C., Penedo, M.G., Penas, M., Carreira, M.J., Gonzalez, F.: Personal authentication using digital retinal images. Pattern Analysis and Applications (2006)
29. Maximov, M., Elezi, I., Leal-Taixé, L.: Ciagan: Conditional identity anonymization generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
30. Meden, B., Emeršič, Ž., Štruc, V., Peer, P.: k-same-net: k-anonymity with generative deep neural networks for face deidentification. Entropy (2018)
31. Milchenko, M., Marcus, D.: Obscuring surface anatomy in volumetric imaging data. Neuroinformatics (2013)
32. Mohamed, Q., Gillies, M.C., Wong, T.Y.: Management of diabetic retinopathy: a systematic review. Jama (2007)
33. Newton, E.M., Sweeney, L., Malin, B.: Preserving privacy by de-identifying face images. IEEE transactions on Knowledge and Data Engineering (2005)
34. Niu, Y., Gu, L., Lu, F., Lv, F., Wang, Z., Sato, I., Zhang, Z., Xiao, Y., Dai, X., Cheng, T.: Pathological evidence exploration in deep retinal image diagnosis. In: Proceedings of the AAAI conference on artificial intelligence (2019)
35. Paul, W., Cao, Y., Zhang, M., Burlina, P.: Defending medical image diagnostics against privacy attacks using generative methods. arXiv preprint arXiv:2103.03078 (2021)
36. Poplin, R., Varadarajan, A.V., Blumer, K., Liu, Y., McConnell, M.V., Corrado, G.S., Peng, L., Webster, D.R.: Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nature Biomedical Engineering (2018)

37. Ravindra, V., Grama, A.: De-anonymization attacks on neuroimaging datasets. In: Proceedings of the 2021 International Conference on Management of Data (2021)
38. Ribaric, S., Pavesic, N.: An overview of face de-identification in still images and videos. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG) (2015)
39. Schimke, N., Kuehler, M., Hale, J.: Preserving privacy in structural neuroimages. In: IFIP annual conference on data and applications security and privacy (2011)
40. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
41. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (2002)
42. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016)
43. Taylor, R.: AI and the Retina: Finding Patterns of Systemic Disease. EyeNet Magazine (2021)
44. Tom, E., Keane, P.A., Blazes, M., Pasquale, L.R., Chiang, M.F., Lee, A.Y., Lee, C.S., Force, A.A.I.T.: Protecting data privacy in the age of ai-enabled ophthalmology. Translational Vision Science & Technology (2020)
45. U.S. Dept. of Health and Human Services: Standards for privacy of individually identifiable health information, Final Rule. Federal Registrar (2002)
46. Wagner, S.K., Fu, D.J., Faes, L., Liu, X., Huemer, J., Khalid, H., Ferraz, D., Korot, E., Kelly, C., Balaskas, K., et al.: Insights into systemic disease through retinal imaging-based oculomics. Translational vision science & technology (2020)
47. Wang, X., Zhang, R., Shen, C., Kong, T., Li, L.: Dense contrastive learning for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
48. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing (2004)
49. Wisely, C.E., Wang, D., Henao, R., Grewal, D.S., Thompson, A.C., Robbins, C.B., Yoon, S.P., Soundararajan, S., Polascik, B.W., Burke, J.R., et al.: Convolutional neural network to identify symptomatic alzheimer's disease using multimodal retinal imaging. British Journal of Ophthalmology (2022)
50. Wu, Y., Yang, F., Xu, Y., Ling, H.: Privacy-protective-gan for privacy preserving face de-identification. Journal of Computer Science and Technology (2019)
51. Xia, W., Zhang, Y., Yang, Y., Xue, J.H., Zhou, B., Yang, M.H.: Gan inversion: A survey. arXiv preprint arXiv:2101.05278 (2021)
52. Xu, C., Ren, J., Zhang, D., Zhang, Y., Qin, Z., Ren, K.: Ganobfuscator: Mitigating information leakage under gan via differential privacy. IEEE Transactions on Information Forensics and Security (2019)
53. Yu, Z., Xiang, Q., Meng, J., Kou, C., Ren, Q., Lu, Y.: Retinal image synthesis from multiple-landmarks input with generative adversarial networks. Biomedical engineering online (2019)
54. Zhou, Y., Wang, B., He, X., Cui, S., Shao, L.: Dr-gan: Conditional generative adversarial network for fine-grained lesion synthesis on diabetic retinopathy images. IEEE Journal of Biomedical and Health Informatics (2020)