Expanding Small-Scale Datasets with Guided Imagination

Yifan Zhang, Daquan Zhou, Bryan Hooi, Kai Wang, Jiashi Feng

National University of Singapore, ByteDance

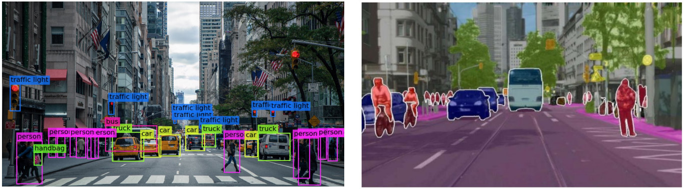
October 4, 2023

Contents

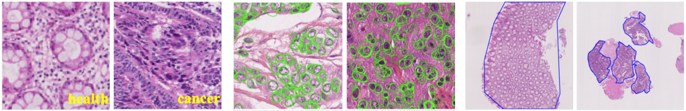
Contents

Background: data scarcity

- The power of DNNs depends heavily on the quantity of training data
- Data scarcity: there are many real-world scenarios where only a limited amount of data is accessible for training



(a) Automatic driving

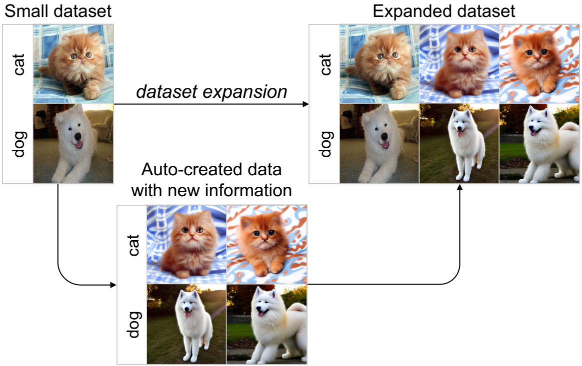


cancer diagnosis nuclei segmentation nidus segmentation

(b) Medical diagnosis

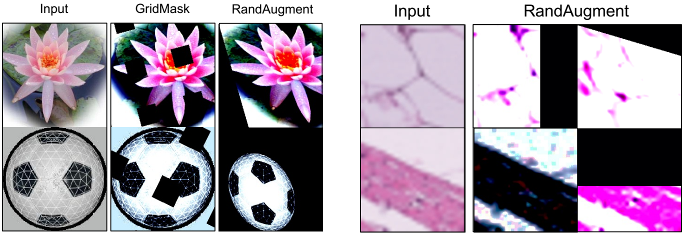
Dataset expansion: a new task

- Collecting and annotating data on a large scale is often costly and time-consuming in such applications
- Dataset expansion: an automatic data generation pipeline to expand a small dataset into a larger & more informative one for model training



Preliminary explorations of previous techniques

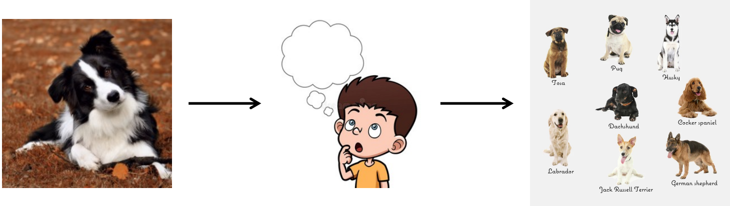
- Naive applications of existing methods cannot address this task
- Data augmentation mainly varies the surface visual characteristics of an image, but cannot create images with new content



- Direct synthesis with pre-trained generative models: those models are class-agnostic to the target dataset, and cannot ensure the synthetic samples have the correct labels and are beneficial to model training

Our motivation

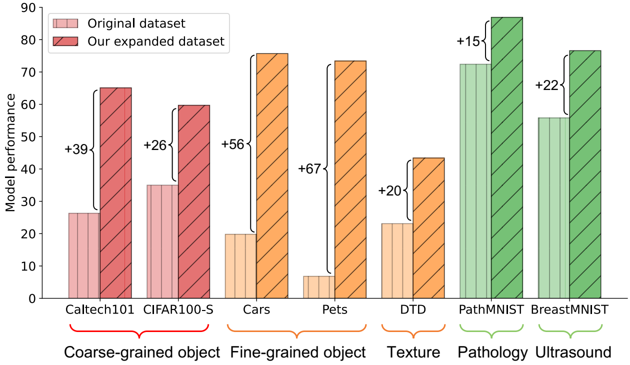
- Motivation: different from the above methods, our solution is inspired by human learning with imagination



- Such an imagination process is highly useful for dataset expansion, since it does not simply perturb the object's appearance but applies rich prior knowledge to create object variants with new information

Our solution

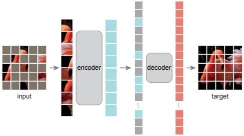
- In light of this, we design a new guided imagination framework (GIF) for dataset expansion
- GIF expands datasets effectively in various small-data scenarios, boosting model accuracy by 36.9% on average over six natural image datasets and by 13.5% on average over three medical datasets



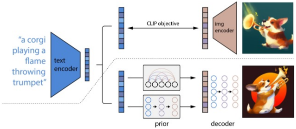
Contents

Our idea

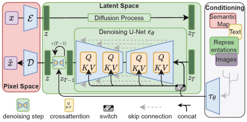
- We attempt to build a computational model to simulate the imagination process, based on prior models, for dataset expansion
- Prior model: deep generative models are trained to capture the entire distribution of a training dataset, and thus can be used as prior models to generate samples with new content



MAE



DALLE-2



Stable Diffusion

Computational imagination models and Challenges

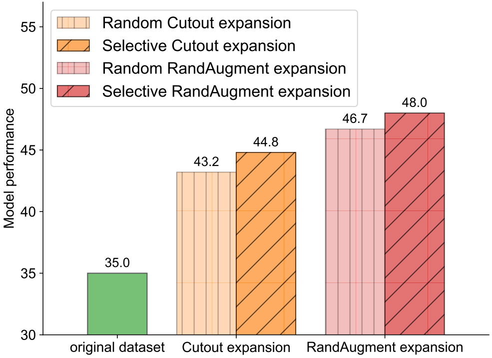
- Given a prior model G, and a seed example (x, y) from the small dataset to expand, we formulate the imagination as $x' = G(f(x) + \delta)$
- Here, $f(\cdot)$ is an image encoder to transform the raw image into an embedding for imagination, and δ is a perturbation applied to $f(x)$ such that G can generate x' different from x

Key questions:

1. How to optimize δ to provide useful guidance: ensure the generated samples with correct labels and is helpful for model training?
2. How to conduct effective expansion: sample-agnostic vs sample-wise expansion? pixel-level vs channel-level update?

Class-maintained informativeness boosting

- Key insight: the generated sample x' should bring new information compared to x, while retaining the same class semantics
- This is difficult to achieve after perturbation in the latent space
- We find that using CLIP zero-shot abilities to maintain class labels and boost informativeness can lead to better expansion effectiveness

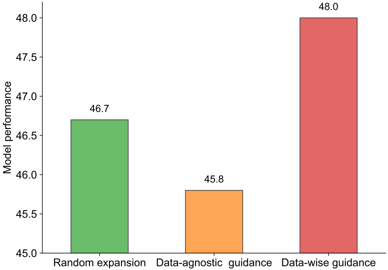


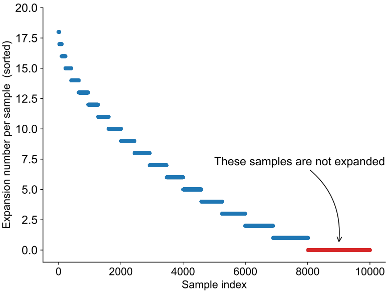
Sample diversity promotion

- To avoid "imagination collapse" where generative models generate excessively similar data, we further promote sample diversity
- The generated images with diversity guidance are more diversified
- This can lead to 1.4% additional accuracy gains on CIFAR100-Subset

Sample-wise expansion

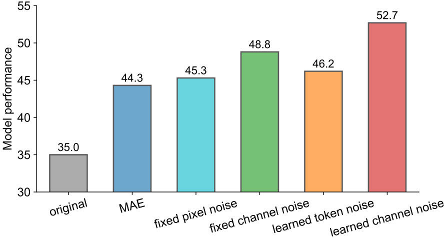
- We find that sample-wise expansion performs much better
- Given a fixed expansion ratio, the sample-agnostic expansion strategy tends to select more expanded samples for easy-to-augment images
- This leads sample-agnostic expansion to waste valuable original samples for expansion and also incurs a class-imbalance problem





Pixel-level optimization vs channel-level optimization?

- We first explore pixel-level noise optimization to vary latent features in MAE, which, however, does not perform well



- We find that the generated image based on pixel-level noise variation is analogous to adding pixel-level noise to the original images

(a) original image

(b) RandAugment

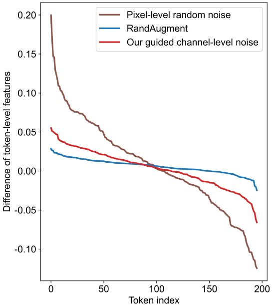
(c) MAE reconstruction

(d) noised-added MAE

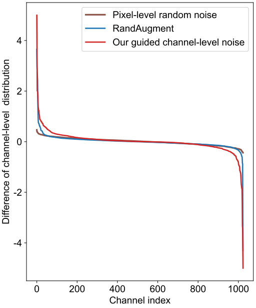
(e) our Guided MAE

Channel-level noise optimization

- MAE with pixel noise variation may harm the integrity and smoothness of image content, while RandAugment slightly changes the content of images but their styles and geometric positions
- This difference inspires us to factorize the influences on images into two perspectives: image styles (i.e., channel dimension of latent feature) and image content (i.e., token dimension of latent feature)



(a) Difference of token-level feature distribution



(b) Difference of channel-level feature distribution

Summary of preliminary studies

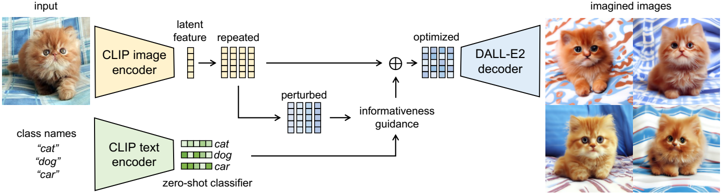
- How to optimize δ to provide useful guidance: ensure the generated samples with correct labels and is helpful for model training?
  1. Class-maintained informativeness boosting
  2. Sample diversity promotion

- How to conduct effective expansion: sample-agnostic vs sample-wise expansion? pixel-level vs channel-level update?
  1. Sample-wise expansion
  2. Channel-level noise optimization

Contents

Guided imagination framework (GIF)

- To detail GIF, we use DALL-E2 as a prior model for illustration



- For each latent feature f, we inject residual multiplicative perturbation with randomly initialized noise $z \sim \mathcal{U}(0,1)$ and bias $b \sim \mathcal{N}(0,1)$ and enforce an ε-ball constraint $\mathcal{P}_{f,\epsilon}(\cdot)$:

$$f' = \mathcal{P}_{f,\epsilon}((1+z)f + b),$$

- In light of our explored criteria, GIF optimizes z and b over the latent feature space as follows:

$$z', b' \leftarrow \arg\max_{z,b} \mathcal{S}_{inf} + \mathcal{S}_{div},$$

Guided imagination framework (GIF)

- GIF optimizes z and b over the latent feature space as follows:

$$z', b' \leftarrow \arg\max_{z,b} \mathcal{S}_{inf} + \mathcal{S}_{div},$$

- Class-maintained informativeness: we design $\mathcal{S}_{inf}$ to improve the information entropy of the perturbed feature while maintaining its class semantics as the seed sample

$$\mathcal{S}_{inf} = s'_j + (s \log(s) - s' \log(s')), \quad \text{s.t.} \quad j = \arg\max(s),$$

- Sample diversity: To promote the diversity of the generated samples, we design $\mathcal{S}_{div}$ as the Kullback–Leibler (KL) divergence among all perturbed latent features of a seed sample

$$\mathcal{S}_{div} = \mathcal{D}_{KL}(f' \| \bar{f}),$$

Theoretical analysis

- We theoretically find our method benefits model generalization
- We resort to δ-cover, and define the dataset diversity by δ-diversity as the inverse of the minimal δ_{min}, i.e., $\delta_{div} = \frac{1}{\delta_{min}}$

Theorem

Let A denote a learning algorithm that outputs a set of parameters given a dataset $\mathcal{D} = \{x_i, y_i\}_{i \in [n]}$ with n i.i.d. samples drawn from distribution $\mathcal{P}_{\mathcal{Z}}$. Assume the hypothesis function is λ^η-Lipschitz continuous, the loss function $\ell(x, y)$ is λ^ℓ-Lipschitz continuous for all y, and is bounded by L, with $\ell(x_i, y_i; A) = 0$ for all $i \in [n]$. If $\mathcal{D}$ constitutes a δ-cover of $\mathcal{P}_{\mathcal{Z}}$, then with probability at least $1 - \gamma$, the generalization error bound satisfies:

$$|\mathbb{E}_{x,y\sim\mathcal{P}_{\mathcal{Z}}}[\ell(x, y; A)] - \frac{1}{n} \sum_{i \in [n]} \ell(x_i, y_i; A)| \overset{c}{\leq} \frac{\lambda^\ell + \lambda^\eta LC}{\delta_{div}},$$

where C is a constant and $\overset{c}{\leq}$ indicates "smaller than" up to a constant.

Theoretical analysis

- This theorem shows that the more diverse samples are created, the more improvement of generalization performance:

$$|\mathbb{E}_{x,y\sim\mathcal{P}_{\mathcal{Z}}}[\ell(x,y;A)] - \frac{1}{n}\sum_{i\in[n]}\ell(x_i,y_i;A)| \overset{c}{\leq} \frac{\lambda^{\ell}+\lambda^{\eta}LC}{\delta_{div}}$$

1. In real small-data applications, the data limitation issue leads the covering radius δ to be very large and thus the δ-diversity is low
2. Simply increasing the data number (e.g., via data repeating) does not help generalization since it does not increase δ-diversity
3. GIF applies two key criteria to create informative and diversified new samples. The expanded dataset thus has higher data diversity, leading to higher δ-diversity and boosting model generalization

Implementation of GIF-DALLE

- GIF-DALLE follows exactly the above pipeline for guided imagination

Algorithm 1 GIF-DALLE Algorithm

Input: Original small dataset $\mathcal{D}_o$; CLIP image encoder $f_{\text{CLIP-I}}(\cdot)$; DALL-E2 diffusion decoder $G(\cdot)$; CLIP zero-shot classifier $w(\cdot)$; Expansion ratio K; Perturbation constraint ε.
Initialize: Synthetic data set $\mathcal{D}_s = \emptyset$
for $x \in \mathcal{D}_o$ **do**
  $\mathcal{S}_{inf} = 0$
  $f = f_{\text{CLIP-I}}(x)$; `// latent feature encoding for seed sample`
  $s = w(f)$; `// CLIP zero-shot prediction for seed sample`
  for *i=1,...,K* **do**
    Initialize noise $z_i \sim \mathcal{U}(0,1)$ and bias $b_i \sim \mathcal{N}(0,1)$
    $f'_i = \mathcal{P}_{f,\epsilon}((1+z_i)f + b_i)$; `// noise perturbation (Eq.(1))`
    $s' = w(f'_i)$; `// CLIP zero-shot prediction`
    $\mathcal{S}_{inf}$ += $s'_j + (s\log(s) - s'\log(s'))$, s.t. $j = \arg\max(s)$; `// class-maintained informativeness (Eq.(5))`
  end
  $\bar{f} = mean(\{f'_i\}_{i=1}^K)$
  $\mathcal{S}_{div} = \sum_i \{\mathcal{D}_{KL}(\sigma(f'_i) \| \sigma(\bar{f}))\}_{i=1}^K = \sum_i \sigma(f'_i)\log(\sigma(f'_i)/\sigma(\bar{f}))$; `// sample diversity (Eq.(6))`
  $\{z'_i, b'_i\}_{i=1}^K \leftarrow \arg\max_{z,b} \mathcal{S}_{inf} + \mathcal{S}_{div}$; `// guided latent feature optimization (Eq.(2))`
  for *i=1,...,K* **do**
    $f''_i = \mathcal{P}_{f,\epsilon}((1+z'_i)f + b'_i)$; `// guided noise perturbation (Eq.(1))`
    $x''_i = G(f''_i)$; `// sample creation`
    Add $x''_i \to \mathcal{D}_s$.
  end
end
Output: Expanded dataset $\mathcal{D}_o \cup \mathcal{D}_s$.

Implementation of GIF-MAE

- Dislike GIF-DALLE, GIF-MAE first generates the latent feature via its encoder, and then conducts channel-wise latent optimization

Algorithm 3 GIF-MAE Algorithm

Input: Original small dataset $\mathcal{D}_o$; MAE image encoder $f(\cdot)$ and image decoder $G(\cdot)$; CLIP image encoder $f_{\text{CLIP-I}}(\cdot)$; CLIP zero-shot classifier $w(\cdot)$; Expansion ratio K; Perturbation constraint ε.
Initialize: Synthetic data set $\mathcal{D}_s = \emptyset$
for $x \in \mathcal{D}_o$ **do**
  $\mathcal{S}_{inf} = 0$
  $f = f(x)$; // latent feature encoding for seed sample
  $s = w(f_{\text{CLIP-I}}(x))$; // CLIP zero-shot prediction for seed sample
  for i=1,...,K **do**
    Initialize noise $z_i \sim \mathcal{U}(0,1)$ and bias $b_i \sim \mathcal{N}(0,1)$
    $f'_i = \mathcal{P}_{f,\epsilon}((1+z_i)f + b_i)$; // channel-level noise perturbation (Eq.(1))
    $x'_i = G(f'_i)$; // intermediate image generation
    $s' = w(f_{\text{CLIP-I}}(x'_i))$
    $\mathcal{S}_{inf}$ += $s'_j + (s\log(s) - s'\log(s'))$, s.t. $j = \arg\max(s)$; // class-maintained informativeness (Eq.(5))
  end
  $\bar{f} = mean(\{f'_i\}_{i=1}^K)$
  $\mathcal{S}_{div} = \sum_i \{\mathcal{D}_{KL}(\sigma(f'_i)\|\sigma(\bar{f}))\}_{i=1}^K = \sum_i \sigma(f'_i)\log(\sigma(f'_i)/\sigma(\bar{f}))$; // sample diversity (Eq.(6))
  $\{z'_i, b'_i\}_{i=1}^K \leftarrow \arg\max_{z,b} \mathcal{S}_{inf} + \mathcal{S}_{div}$; // guided latent feature optimization (Eq.(2))
  for i=1,...,K **do**
    $f''_i = \mathcal{P}_{f,\epsilon}((1+z'_i)f + b'_i)$; // guided channel-wise noise perturbation (Eq.(1))
    $x''_i = G(f''_i)$; // sample creation
    Add $x''_i \to \mathcal{D}_s$.
  end
end
Output: Expanded dataset $\mathcal{D}_o \cup \mathcal{D}_s$.

Implementation of GIF-SD

- GIF-SD has one more step than GIF-MAE before noise perturbation, i.e., conducting prompt-guided diffusion for the latent feature

Algorithm 2 GIF-SD Algorithm

Input: Original small dataset $\mathcal{D}_o$; SD image encoder $f(\cdot)$ and image decoder $G(\cdot)$; SD diffusion module $f_{\text{diff}}(\cdot; [prompt])$; CLIP image encoder $f_{\text{CLIP-I}}(\cdot)$; DALL-E2 diffusion decoder $G(\cdot)$; CLIP zero-shot classifier $w(\cdot)$; Expansion ratio K; Perturbation constraint ε.

Initialize: Synthetic data set $\mathcal{D}_s = \emptyset$

for $x \in \mathcal{D}_o$ **do**

```
    S_inf = 0
    f = f(x) ;                                   // latent feature encoding for seed sample
    Randomly sample a [prompt] ;                 // Prompt generation (Eq.(7))
    f = f_diff(f; [prompt]) ;                    // SD latent diffusion
    s = w(f_CLIP-I(x)) ;                         // CLIP zero-shot prediction for seed sample
    for i=1,...,K do
        Initialize noise z_i ~ U(0,1) and bias b_i ~ N(0,1)
        f'_i = P_{f,ε}((1 + z_i)f + b_i) ;       // noise perturbation (Eq.(1))
        s' = w(f'_i) ;                           // CLIP zero-shot prediction
        S_inf += s'_j + (s log(s) − s' log(s')), s.t. j = arg max(s) ; // class-maintained informativeness (Eq.(5))
    end
    f̄ = mean({f'_i}_{i=1}^K)
    S_div = Σ_i {D_KL(σ(f'_i) ‖ σ(f̄))}_{i=1}^K = Σ_i σ(f'_i) log(σ(f'_i)/σ(f̄)) ; // sample diversity (Eq.(6))
    {z'_i, b'_i}_{i=1}^K ← arg max_{z,b} S_inf + S_div ; // guided latent feature optimization (Eq.(2))
    for i=1,...,K do
        f''_i = P_{f,ε}((1 + z'_i)f + b'_i) ;    // guided noise perturbation (Eq.(1))
        x''_i = G(f''_i) ;                       // sample creation
        Add x''_i → D_s.
    end
end
```

Math form of the loop body:

- $\mathcal{S}_{inf} = 0$
- $f = f(x)$; // latent feature encoding for seed sample
- Randomly sample a $[prompt]$; // Prompt generation (Eq.(7))
- $f = f_{\text{diff}}(f; [prompt])$; // SD latent diffusion
- $s = w(f_{\text{CLIP-I}}(x))$; // CLIP zero-shot prediction for seed sample
- **for** i=1,...,K **do**
  - Initialize noise $z_i \sim \mathcal{U}(0,1)$ and bias $b_i \sim \mathcal{N}(0,1)$
  - $f'_i = \mathcal{P}_{f,\epsilon}((1+z_i)f + b_i)$; // noise perturbation (Eq.(1))
  - $s' = w(f'_i)$; // CLIP zero-shot prediction
  - $\mathcal{S}_{inf} \mathrel{+}= s'_j + (s\log(s) - s'\log(s'))$, s.t. $j = \arg\max(s)$; // class-maintained informativeness (Eq.(5))
- **end**
- $\bar{f} = mean(\{f'_i\}_{i=1}^K)$
- $\mathcal{S}_{div} = \sum_i \{\mathcal{D}_{KL}(\sigma(f'_i) \| \sigma(\bar{f}))\}_{i=1}^K = \sum_i \sigma(f'_i)\log(\sigma(f'_i)/\sigma(\bar{f}))$; // sample diversity (Eq.(6))
- $\{z'_i, b'_i\}_{i=1}^K \leftarrow \arg\max_{z,b} \mathcal{S}_{inf} + \mathcal{S}_{div}$; // guided latent feature optimization (Eq.(2))
- **for** i=1,...,K **do**
  - $f''_i = \mathcal{P}_{f,\epsilon}((1+z'_i)f + b'_i)$; // guided noise perturbation (Eq.(1))
  - $x''_i = G(f''_i)$; // sample creation
  - Add $x''_i \to \mathcal{D}_s$.
- **end**

end

Output: Expanded dataset $\mathcal{D}_o \cup \mathcal{D}_s$.

Contents

Contents

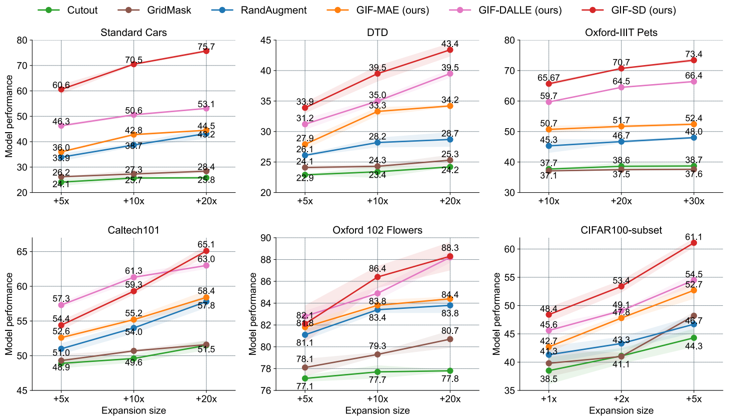
Expansion effectiveness

- GIF is more effective in expanding small-scale datasets
- Compared with the model trained on original datasets, GIF-SD leads to 36.9% accuracy gains on average over six natural image datasets and 13.5% gains over three medical datasets

Dataset	Natural image datasets							Medical image datasets			
	Caltech101	Cars	Flowers	DTD	CIFAR100-S	Pets	Average	PathMNIST	BreastMNIST	OrganSMNIST	Average
Original	26.3	19.8	74.1	23.1	35.0	6.8	30.9	72.4	55.8	76.3	68.2
CLIP	82.1	55.8	65.9	41.7	41.6	85.4	62.1	10.7	51.8	7.7	23.4
Distillation of CLIP	33.2	18.9	75.1	25.6	37.8	11.1	33.6	77.3	60.2	77.4	71.6
Expanded											
Cutout	51.5	25.8	77.8	24.2	44.3	38.7	43.7 (+12.8)	78.8	66.7	78.3	74.6 (+6.4)
GridMask	51.6	28.4	80.7	25.3	48.2	37.6	45.3 (+14.4)	78.4	66.8	78.9	74.7 (+6.5)
RandAugment	57.8	43.2	83.8	28.7	46.7	48.0	51.4 (+20.5)	79.2	68.7	79.6	75.8 (+7.6)
MAE	50.6	25.9	76.3	27.6	44.3	39.9	44.1 (+13.2)	81.7	63.4	78.6	74.6 (+6.4)
DALL-E2	61.3	48.3	84.1	34.5	52.1	61.7	57.0 (+26.1)	82.8	70.8	79.3	77.6 (+9.4)
SD	51.1	51.7	78.8	33.2	52.9	57.9	54.3 (+23.4)	85.1	73.8	78.9	79.3 (+11.1)
GIF-MAE (ours)	58.4	44.5	84.4	34.2	52.7	52.4	54.4 (+23.5)	82.0	73.3	80.6	78.6 (+10.4)
GIF-DALLE (ours)	63.0	53.1	88.2	39.5	54.5	66.4	60.8 (+29.9)	84.4	76.6	80.5	80.5 (+12.3)
GIF-SD (ours)	**65.1**	**75.7**	**88.3**	**43.4**	**61.1**	**73.4**	**67.8** (+36.9)	**86.9**	**77.4**	**80.7**	**81.7** (+13.5)

Expansion efficiency

- GIF is more *sample efficient* than data augmentations
- 5× expansion by GIF-SD and GIF-DALLE even outperforms 20× expansion by various data augmentations, implying our methods are at least 4× more efficient than them on Cars



Benefit to model generalization

- GIF significantly boosts out-of-distribution (OOD) generalization, bringing 11+% gain on average over 15 types of OOD corruption

Table: CIFAR100-C with the severity level 1

	Noise			Blur				Weather				Digital				
Dataset	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Average
Original	25.6	29.3	25.0	34.2	**32.2**	31.7	30.9	32.3	28.3	31.8	33.7	29.2	31.7	34.1	30.9	30.7
5×-expanded by GIF-SD	50.3	54.6	50.8	59.2	29.4	53.7	51.9	53.1	54.0	58.7	59.5	57.1	52.5	57.9	**54.7**	53.2 (+22.5)
20×-expanded by GIF-SD	**55.0**	**60.5**	**54.8**	**66.1**	30.2	**56.0**	**58.0**	**61.1**	**62.2**	**65.1**	**66.2**	**64.3**	**59.2**	**63.8**	**60.8**	**58.9** (+27.2)

Table: CIFAR100-C with the severity level 3

	Noise			Blur				Weather				Digital				
Dataset	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Average
Original	12.8	17.0	12.5	30.5	31.7	25.2	28.6	26.5	19.0	18.6	28.3	11.5	29.5	33.6	28.8	23.6
5×-expanded by GIF-SD	29.7	36.4	32.7	51.9	32.4	39.2	46.0	45.3	38.1	47.1	55.7	37.3	48.6	53.2	49.4	43.3 (+19.3)
20×-expanded by GIF-SD	**31.8**	**39.2**	**34.7**	**58.4**	**33.4**	**43.1**	**51.9**	**51.7**	**47.4**	**55.0**	**63.3**	**46.5**	**54.9**	**58.0**	**53.6**	**48.2** (+24.6)

Table: CIFAR100-C with the severity level 5

	Noise			Blur				Weather				Digital				
Dataset	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Average
Original	9.4	10.7	5.5	24.9	**28.9**	22.3	25.9	19.4	16.6	8.2	18.3	2.7	29.0	31.8	27.3	18.7
5×-expanded by GIF-SD	21.4	23.8	10.8	31.8	22.8	33.1	37.6	38.1	31.1	24.7	43.7	8.6	38.6	**36.0**	45.6	29.8 (+11.1)
20×-expanded by GIF-SD	**22.9**	**25.5**	**11.1**	**33.5**	24.1	**36.2**	**41.8**	**46.4**	**38.4**	**32.1**	**53.5**	**13.9**	**40.4**	32.0	**48.8**	**33.4** (+14.7)

Applicability to various model architectures

- The expanded datasets are readily used for training various model architectures, bringing consistent gains for all the architectures

Dataset	Cars				
	ResNet-50	ResNeXt-50	WideResNet-50	MobilteNet-v2	Avg.
Original dataset	$19.8_{\pm 0.9}$	$18.4_{\pm 0.5}$	$32.0_{\pm 0.8}$	$26.2_{\pm 4.2}$	24.1
5×-expanded by GIF-DALLE	$53.1_{\pm 0.2}$	$43.7_{\pm 0.2}$	$60.0_{\pm 0.6}$	$47.8_{\pm 0.6}$	51.2 (+27.1)
5×-expanded by GIF-SD	$\mathbf{60.6}_{\pm 1.9}$	$\mathbf{64.1}_{\pm 1.3}$	$\mathbf{75.1}_{\pm 0.4}$	$\mathbf{60.2}_{\pm 1.6}$	**65.0** (+40.9)

Dataset	CIFAR100-S				
	ResNet-50	ResNeXt-50	WideResNet-50	MobilteNet-v2	Avg.
Original dataset	$35.0_{\pm 3.2}$	$36.3_{\pm 2.1}$	$42.0_{\pm 0.3}$	$50.9_{\pm 0.2}$	41.1
5×-expanded by GIF-DALLE	$54.5_{\pm 1.1}$	$52.4_{\pm 0.7}$	$55.3_{\pm 0.3}$	$56.2_{\pm 0.2}$	54.6 (+13.5)
5×-expanded by GIF-SD	$\mathbf{61.1}_{\pm 0.8}$	$\mathbf{59.0}_{\pm 0.7}$	$\mathbf{64.4}_{\pm 0.2}$	$\mathbf{62.4}_{\pm 0.1}$	**61.4** (+20.3)

Dataset	Pets				
	ResNet-50	ResNeXt-50	WideResNet-50	MobilteNet-v2	Avg.
Original dataset	$6.8_{\pm 1.8}$	$19.0_{\pm 1.6}$	$22.1_{\pm 0.5}$	$37.5_{\pm 0.4}$	21.4
5×-expanded by GIF-DALLE	$46.2_{\pm 0.1}$	$52.3_{\pm 1.5}$	$66.2_{\pm 0.1}$	$60.3_{\pm 0.3}$	56.3 (+34.9)
5×-expanded by GIF-SD	$\mathbf{65.8}_{\pm 0.6}$	$\mathbf{56.5}_{\pm 0.6}$	$\mathbf{70.9}_{\pm 0.4}$	$\mathbf{60.6}_{\pm 0.5}$	**63.5** (+42.1)

Effectiveness in long-tailed datasets

- Compared to training on the original CIFAR100-LT dataset, 20× expansion by our GIF-SD leads to a 13.5% model accuracy gain
- GIF boosts the performance of few-shot classes more than many-shot classes, which means that GIF helps to address class imbalance

CIFAR100-LT	Training losses	Many-shot classes	Medium-shot classes	Few-shot classes	Overall
Original	Cross-entropy	70.5	41.1	8.1	41.4
20×-expanded by GIF-SD	Cross-entropy	79.5 (+9.0)	54.9 (+13.8)	26.4 (+18.3)	54.9 (+13.5)
Original	Balanced Softmax	67.9	45.8	17.7	45.1
20×-expanded by GIF-SD	Balanced Softmax	73.7 (+5.8)	59.2 (+13.4)	44.5 (+26.8)	59.9 (+14.8)

Contents

Comparisons to CLIP

- GIF has two advantages over CLIP in real small-data applications:
  1. GIF has better applicability to the scenarios of different image domains, like medical image domains
  2. GIF creates expanded datasets ready for training various architectures, which is more applicable to the scenario with hardware constraints

Dataset	PathMNIST	BreastMNIST	OrganSMNIST
Original dataset	$72.4_{\pm 0.7}$	$55.8_{\pm 1.3}$	$76.3_{\pm 0.4}$
Linear-probing of CLIP	$74.3_{\pm 0.1}$	$60.0_{\pm 2.9}$	$64.9_{\pm 0.2}$
fine-tuning of CLIP	$78.4_{\pm 0.9}$	$67.2_{\pm 2.4}$	$78.9_{\pm 0.1}$
distillation of CLIP	$77.3_{\pm 1.7}$	$60.2_{\pm 1.3}$	$77.4_{\pm 0.8}$
5×-expanded by GIF-MAE	$82.0_{\pm 0.7}$	$73.3_{\pm 1.3}$	$80.6_{\pm 0.5}$
5×-expanded by GIF-DALLE	$84.4_{\pm 0.3}$	$76.6_{\pm 1.4}$	$80.5_{\pm 0.2}$
5×-expanded by GIF-SD	$\mathbf{86.9}_{\pm 0.3}$	$\mathbf{77.4}_{\pm 1.8}$	$\mathbf{80.7}_{\pm 0.2}$

Effectiveness of guidance

- With our guidance, GIF obtains consistent performance gains compared to unguided expansion with SD, DALL-E2, or MAE

Dataset	Natural image datasets							Medical image datasets			
	Caltech101	Cars	Flowers	DTD	CIFAR100-S	Pets	Average	PathMNIST	BreastMNIST	OrganSMNIST	Average
Original	26.3	19.8	74.1	23.1	35.0	6.8	30.9	72.4	55.8	76.3	68.2
MAE	50.6	25.9	76.3	27.6	44.3	39.9	44.1 (+13.2)	81.7	63.4	78.6	74.6 (+6.4)
GIF-MAE (ours)	58.4	44.5	84.4	34.2	52.7	52.4	54.4 (+23.5)	82.0	73.3	80.6	78.6 (+10.4)
DALL-E2	61.3	48.3	84.1	34.5	52.1	61.7	57.0 (+26.1)	82.8	70.8	79.3	77.6 (+9.4)
GIF-DALLE (ours)	63.0	53.1	88.2	39.5	54.5	66.4	60.8 (+29.9)	84.4	76.6	80.5	80.5 (+12.3)
SD	51.1	51.7	78.8	33.2	52.9	57.9	54.3 (+23.4)	85.1	73.8	78.9	79.3 (+11.1)
GIF-SD (ours)	**65.1**	**75.7**	**88.3**	**43.4**	**61.1**	**73.4**	**67.8** (+36.9)	**86.9**	**77.4**	**80.7**	**81.7** (+13.5)

Ablation of guidance

- Boosting the class-maintained informativeness $\mathcal{S}_{inf}$ is important for GIF-DALLE expansion

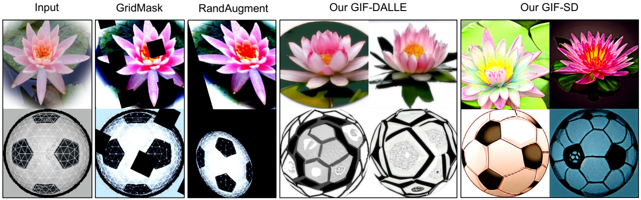
Method	$\mathcal{S}_{inf}$	$\mathcal{S}_{div}$	CIFAR100-Subset
GIF-DALLE			$52.1_{\pm 0.9}$
	✓		$53.1_{\pm 0.3}$
		✓	$51.8_{\pm 1.3}$
	✓	✓	$54.5_{\pm 1.1}$

- Both the class-maintained informativeness guidance $\mathcal{S}_{inf}$ and the diversity promotion guidance $\mathcal{S}_{div}$ contribute to model performance

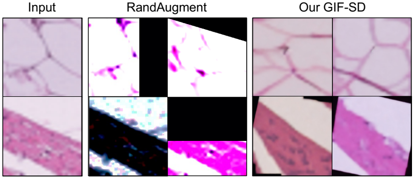
Method	Designed prompts	$\mathcal{S}_{inf}$	$\mathcal{S}_{div}$	CIFAR100-Subset
GIF-SD				$52.9_{\pm 0.8}$
	✓			$56.2_{\pm 1.0}$
	✓	✓		$59.6_{\pm 1.1}$
	✓		✓	$59.4_{\pm 1.2}$
	✓	✓	✓	$61.1_{\pm 0.8}$

Visualization

- GIF can create images with new content from the seed images



- RandAugment randomly varies the medical images and may crop the lesion areas. Hence, it cannot ensure the created samples are informative, and even generates noisy samples

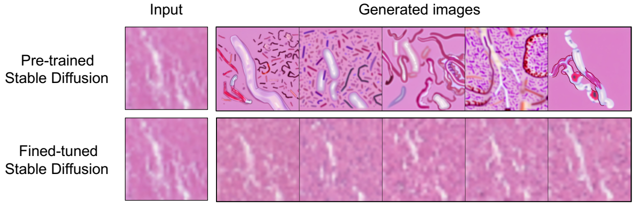


Need we fine-tune generative models on medical datasets?

- Directly applying pre-trained SD and DALL-E2 performs limited, compared to GIF-MAE

Dataset	PathMNIST	BreastMNIST	OrganSMNIST	Average
Original	$72.4_{\pm 0.7}$	$55.8_{\pm 1.3}$	$76.3_{\pm 0.4}$	68.2
GIF-MAE	$82.0_{\pm 0.7}$	$73.3_{\pm 1.3}$	$80.6_{\pm 0.5}$	78.6
GIF-DALLE (w/o tuning)	$78.4_{\pm 1.0}$	$59.3_{\pm 2.5}$	$76.4_{\pm 0.3}$	71.4
GIF-DALLE (w/ tuning)	$84.4_{\pm 0.3}$	$76.6_{\pm 1.4}$	$80.5_{\pm 0.2}$	80.5
GIF-SD (w/o tuning)	$80.8_{\pm 1.6}$	$59.4_{\pm 2.2}$	$79.5_{\pm 0.4}$	73.2
GIF-SD (w/ tuning)	$86.9_{\pm 0.6}$	$77.4_{\pm 1.8}$	$80.7_{\pm 0.2}$	81.7

- Pre-trained SD suffers from domain shifts between natural and medical images, and cannot generate informative medical samples
- Fine-tuning prior generative models is necessary for medical domains



Comparison to infinite data augmentation

- RandAugment with more epochs leads to better performance but gradually converges
- GIF-SD achieves better performance when training only 100 epochs

Methods	Epochs	Consumption	Accuracy
Original			
Standard training	100	1 million	$35.0_{\pm 1.7}$
Training with RandAugment	100	1 million	$39.6_{\pm 2.5}$
Training with RandAugment	200	2 million	$46.9_{\pm 0.9}$
Training with RandAugment	300	3 million	$48.1_{\pm 0.6}$
Training with RandAugment	400	4 million	$49.6_{\pm 0.4}$
Training with RandAugment	500	5 million	$51.3_{\pm 0.3}$
Training with RandAugment	600	6 million	$51.1_{\pm 0.3}$
Training with RandAugment	700	7 million	$50.6_{\pm 1.1}$
Expanded			
5×-expanded by GIF-SD	100	6 million	$\mathbf{61.1}_{\pm 0.8}$

Comparison to picking related samples from larger datasets

- Picking and labeling data from larger image datasets with CLIP has the potential for dataset expansion
- However, a large-scale related dataset may be unavailable while selecting data from different image domains is unhelpful

CIFAR100-Subset	Accuracy
Original dataset	$35.0_{\pm 1.7}$
Expanded dataset	
5×-expanded by picking data from ImageNet with CLIP	$50.9_{\pm 1.1}$
5×-expanded by GIF-DALLE	$54.5_{\pm 1.1}$
5×-expanded by GIF-SD	$\mathbf{61.1}_{\pm 0.8}$

Relation analysis between domain gap and model accuracy

- We compute the Fréchet Inception Distance (FID) between the synthetic images and the original images of CIFAR100-S
- One might assume that a lower FID indicates higher quality in the expanded data, but in reality, it's not always the case
- The effectiveness depends on how much additional information and class consistency the generated data can provide, rather than the distribution similarity between those samples and the original data

Datasets	FID	Accuracy (%)
CIFAR100-S	-	35.0
RandAugment	24.3	46.7
Cutout	104.7	44.3
Gridmask	104.8	48.2
GIF-MAE	72.3	52.7
GIF-DALLE	39.5	54.5
GIF-SD	81.7	61.1

Safety check

- We employ the Google Cloud Vision API[1] to perform a safety check on the 50,000 images generated by GIF-SD
- The synthetic images by our method are safe and harmless

Metrics	Very unlikely	Unlikely	Neutral	Likely	Very likely
Adult	96%	4%	0%	0%	0%
Spoof	82%	15%	3%	0%	0%
Medical	86%	14%	0%	0%	0%
Violence	69%	31%	0%	0%	0%
Racy	66%	25%	9%	0%	0%

[1] https://cloud.google.com/vision/docs/detecting-safe-search

Contents

Summary

1. A new task of dataset expansion that contributes to boosting DNN training in real small-data scenarios
2. Two key criteria for effective expansion: <u>class-maintained informativeness boosting</u> and <u>sample diversity promotion</u>
3. A new Guided Imagination Framework for effective expansion: leading to promising performance improvement on both small-scale natural and medical image datasets

Future directions

1. Huge headroom of dataset expansion: the expanded samples are still less informative than the real ones. For example, $5\times$-expanded CIFAR100-S (61.1 ± 0.8) vs CIFAR100 (71.0 ± 0.6)
2. Computational efficiency: although it is not our focus, exploring how to conduct more computationally efficient expansion is important
3. More tasks: it is also exciting to conduct dataset expansion for object detection and semantic segmentation

Thanks

More visualization of GIF-SD

More visualization of GIF-DALLE

More visualization of GIF-MAE