Semantic-Guided Generative Image Augmentation Method with Diffusion Models for Image Classification

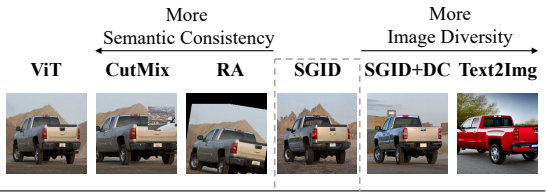
Bohan Li, Xiao Xu, Xinghao Wang, Yutai Hou, Yunlong Feng, Feng Wang, Xuanliang Zhang, Qingfu Zhu*, Wanxiang Che

Harbin Institute of Technology, Harbin, China

{bhli, xxu, xhwang, ythou, ylfeng}@ir.hit.edu.cn, {7203610216, 120l020412}@stu.hit.edu.cn, {qfzhu, car}@ir.hit.edu.cn

Abstract

Existing image augmentation methods consist of two categories: perturbation-based methods and generative methods. Perturbation-based methods apply pre-defined perturbations to augment an original image, but only locally vary the image, thus lacking image diversity. In contrast, generative methods bring more image diversity in the augmented images but may not preserve semantic consistency, thus may incorrectly change the essential semantics of the original image. To balance image diversity and semantic consistency in augmented images, we propose **SGID**, a **S**emantic-guided **G**enerative **I**mage augmentation method with **D**iffusion models for image classification. Specifically, SGID employs diffusion models to generate augmented images with good image diversity. More importantly, SGID takes image labels and captions as guidance to maintain semantic consistency between the augmented and original images. Experimental results show that SGID outperforms the best augmentation baseline by 1.72% on ResNet-50 (from scratch), 0.33% on ViT (ImageNet-21k), and 0.14% on CLIP-ViT (LAION-2B). Moreover, SGID can be combined with other image augmentation baselines and further improves the overall performance. We demonstrate the semantic consistency and image diversity of SGID through quantitative human and automated evaluations, as well as qualitative case studies.



		More Semantic Consistency ←			More Image Diversity →	
	ViT	**CutMix**	**RA**	**SGID**	**SGID+DC**	**Text2Img**
Image						
Cal101	91.20	91.77 (+0.57)	91.92 (+0.72)	**93.91 (+2.71)**	93.44 (+2.24)	91.21 (+0.01)
Cars	82.99	84.20 (+1.21)	85.62 (+2.63)	**86.73 (+3.74)**	86.33 (+3.34)	85.81 (+2.82)
Flowers	95.70	96.53 (+0.83)	96.66 (+0.96)	**97.16 (+1.46)**	97.12 (+1.42)	96.93 (+1.23)
DTD	71.22	71.32 (+0.10)	71.32 (+0.10)	**73.35 (+2.13)**	72.30 (+1.08)	70.57 (-0.65)
Avg.	85.28	85.96 (+0.68)	86.38 (+1.10)	**87.79 (+2.51)**	87.30 (+2.02)	86.13 (+0.85)

Figure 1: A comparison of four baseline methods and our proposed SGID using the ViT (ImageNet-21k) backbone across four datasets. Our SGID strikes a balance between semantic consistency and image diversity, leading to the highest performance improvement.

1 Introduction

The data-hungry problem in deep learning has been a hot topic (Minaee et al. 2021) since a sufficient number of training samples is crucial for unleashing the power of deep networks (Zhang et al. 2022). However, manually collecting and labeling large-scale datasets are both expensive and time-consuming (Li, Hou, and Che 2022), giving rise to the Data Augmentation (DA) methods. Take image classification as an example, generally, a DA method generates augmented images that are diverse from the original training images while preserving the essential semantics of the original images. It has been demonstrated to be effective in improving the model performance on classification tasks in practice (Dunlap et al. 2023).

Typically, DA methods for image classification can be divided into two categories: perturbation-based and generation-based. The former obtains augmented images by modifying the original image with pre-defined perturbations, *e.g.*, image erasing (Zhong et al. 2020) and image mixup (Yun et al. 2019). In this way, most of the semantics remain since the augmented image only locally differs from the original image in a limited area. However, the diversity is quite limited at the same time and becomes the bottleneck of the methods of this category. In contrast, the generation-based methods synthesize augmented images by generative models like diffusion models (Rombach et al. 2022). In this way, they can generate quite diverse images but are inferior to the perturbation-based methods in preserving semantics, which are less diverse but more semantically consistent. The excessive noise or diversity introduced by generative DA methods may **incorrectly change** the essential semantics of the original image. As shown in Figure 1, the last two augmented images incorrectly show two door handles on the same side of the door or change the badge on the rear of the car.

Naturally, we intend to explore how to maintain a bal-

*Corresponding Authors.

ance between image diversity and semantic consistency in a generative image DA method, which has not yet been systematically explored by existing methods, to achieve better performance on image classification tasks. To this end, we propose SGID, a **S**emantic-guided **G**enerative **I**mage augmentation method with **D**iffusion models for image classification. It ensures semantic consistency between the original and augmented images, and achieves good image diversity (see Figure 1). SGID consists of two steps: ***(1)*** Collect the text label of the image and then generate the image caption with the BLIP (Li et al. 2022) model. Both of them convey the essential semantics of the original image. ***(2)*** Concatenate the label and caption to construct a textual prompt and subsequently feed it into the Stable Diffusion (Rombach et al. 2022) model along with the original image. SGID tends to generate diverse augmented images, while the textual prompt as semantic guidance helps to preserve the essential semantics of the original images (see Sec 4.4).

We conduct experiments on seven image classification datasets based on three backbones including ResNet-50 (from scratch) (He et al. 2016), ViT (ImageNet-21k) (Dosovitskiy et al. 2020), and CLIP-ViT (LAION-2B) (Cherti et al. 2022). We compare SGID with seven strong DA baselines including four perturbation-based methods and three generative methods. Our method outperforms the backbones on all datasets and achieves the best or comparable performance to all baselines. Especially, SGID leads to $10.39\%, 2.08\%, 0.85\%$ average accuracy gain across three backbones and seven datasets, respectively. Moreover, we find that combining SGID with standard DA baselines can achieve further improvement on seven datasets. We further compare the semantic consistency and diversity between SGID as well as baselines by human evaluation, automatic similarity evaluation, and case studies.

Our contributions are as follows:

- We propose a **S**emantic-guided **G**enerative **I**mage augmentation method with **D**iffusion models (SGID) for image classification. Human evaluation, automated evaluations, and case studies demonstrate that SGID balances the semantic consistency and image diversity.
- SGID achieves better average performance than the best baseline by 1.72%, 0.33%, and 0.14% across three backbones and seven datasets.
- SGID can be combined with other image augmentation baselines and further improves the overall performance.

2 Background

Image augmentation is widely used in computer vision (Minaee et al. 2021; Algan and Ulusoy 2021). It generates augmented images that are diverse from the original images while preserving their semantics. Common augmentation methods include ***(1)*** perturbation-based methods and ***(2)*** generative methods.

Perturbation-based methods apply pre-defined perturbations to augment an original image and preserve image semantic consistency (Yang et al. 2022). *Random Erasing* (Zhong et al. 2020) generates augmented images by deleting one or more subregions of an image. *CutMix* (Yun et al. 2019) uses the full or partial features and labels of different images to apply some interpolation methods. *RandAugment* (Cubuk et al. 2020) tries to search the space of augmentation methods according to different tasks. *MoEx* (Li et al. 2021) performs the transformation in a learned feature space rather than conducting augmentation only in the input space. Despite their effectiveness in image classification, the pre-defined perturbations of these methods only locally vary the images, thus lacking diversity.

Thanks to the development of generative models like Stable Diffusion (SD), an image generation model pre-trained on large-scale image-text pairs, there are some attempts at generative methods for more diversity (Zhang et al. 2022). *Text2Img* (He et al. 2022) applies a fine-tuned T5 model to generate captions based on image labels, and then employs a text-to-image diffusion model to generate images without using the original image. Dunlap et al. (2023) respectively use diverse captions or editing instructions to modify original images into augmented images. During image generation, they correspondingly employ the Image2Image diffusion model or the InstructPix2Pix diffusion model (Brooks, Holynski, and Efros 2023) as generative models. Inspired by this work, we take its spirit into our SGID to obtain two variants of SGID as generative baselines named *SGID+DiverseCaption* and *SGID+InstructPix2Pix*.[1]

However, existing generative methods may incorrectly change the essential semantics of the original image, which is crucial for image classification (Burg et al. 2023).

This paper aims to improve the semantic consistency of generative methods through guiding the generation of augmented images by explicitly using the essential semantics of original images. SGID balances the image diversity and semantic consistency in augmented images, and achieves consistent performance gains across datasets and backbones.

3 SGID

In this paper, we propose SGID, a semantic-guided generative image augmentation method with diffusion models for image classification. We do not aim to exceed existing image augmentation baselines on various datasets and different backbones, but to explore such a method that balances image diversity and semantic consistency, *i.e.*, preserving the essential semantics of the original images and simultaneously bringing good image diversity. In addition, SGID can be naturally combined with other DA baselines and further improve their performance. It consists of two essential steps, as illustrated in Figure 2:

1. We first collect textual labels for each image, then use BLIP to generate captions, and then use CLIP (Radford et al. 2021) to calculate the similarity between the chosen caption and the original image. Both labels and captions provide the essential semantics of the original images.
2. We construct the prompt based on the label and the caption of each image. The prompt is subsequently fed into Stable Diffusion along with the original image. The semantic guidance contained in the prompt can help gener-

[1]We further discuss this work (Dunlap et al. 2023) and other generative methods in Appendix Sec. F. and D.

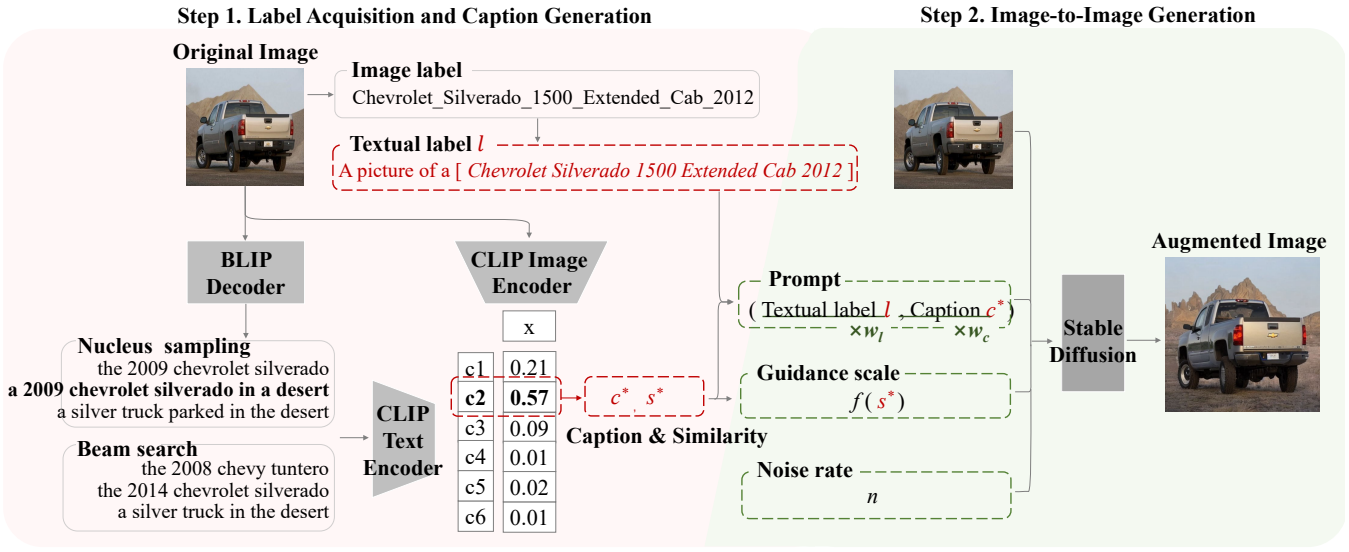


Figure 2: An illustration of our SGID. Step 1 first collects textual labels for each image, then use BLIP to generate captions, and then use CLIP to calculate the similarity between the chosen caption and the original image. Step 2 generates the augmented images through the Stable Diffusion model, utilizing the original image, the prompt consisting of the textual label and caption, the noise rate, and the guidance scale based on the similarity.

ate diverse and semantic-consistent augmented images.[2]

3.1 Label Acquisition and Caption Generation

In **Step 1**, we construct prompts from original images as semantic guidance for the subsequent **Step 2**. The prompt consists of the textual label and the caption of the original image. We argue that the image label contains accurate semantic information and the caption provides the overall description of the image. Thus, the prompt is an explicit constraint to preserve semantic consistency. Specifically, for each image $x \in \mathcal{X}$ in an image classification dataset $\mathcal{D} = (\mathcal{X}, \mathcal{Y})$, we first employ its groundtruth image label to construct a corresponding sentence as the textual label l:

$$l = \text{“A picture of a [label]”}. \tag{1}$$

Then we generate captions for the image by the BLIP model:

$$c = \text{BLIP}(x). \tag{2}$$

We explore different sampling strategies, including beam search (Shen et al. 2017) and nucleus sampling (Holtzman et al. 2019)). Beam search tends to generate common and safe captions, hence offering less extra knowledge, and Nucleus sampling generates more diverse captions (Li et al. 2022). We choose one of the sampling strategies based on the results of the validation set. For each image, we obtain the corresponding caption set $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$ and randomly select one caption from it. Captions can provide further semantic information beyond the object category of the image, such as the description of the background and color. Combining captions with groundtruth but often too-short image labels can provide effective semantic guidance for image-to-image generation.

[2]The introduction to BLIP, CLIP, and SD is Appendix Sec. E.

Moreover, to increase the semantic consistency of the prompt, we also explore potentially higher-quality captions by taking CLIP as an **optional** filter. Given the generated caption set $\mathcal{C}$, we use CLIP to calculate the similarity between the original image with each caption and obtain the similarity set: $\mathcal{S} = \{s_1, s_2, \ldots, s_k\}$. We select the caption with the highest similarity in $\mathcal{S}$. We denote the selected caption and its similarity with c^* and s^*. For example, in Figure 2, the caption c^* = “a 2009 chevrolet silverado in a desert” generated by nucleus sampling gets the highest similarity $s^* = 0.57$.

3.2 Image-To-Image Generation

In **Step 2**, we take the textual label l and the caption c^* as semantic guidance for image-to-image generation. We first concatenate l and c^* to construct the *textual prompt*, *i.e.*, $p = (l, c^*)$. For example, “A picture of a [Chevrolet Silverado 1500 Extended Cab 2012], a 2009 chevrolet silverado in a desert.” in Figure 2. The textual prompt carries not only accurate but brief image semantics from l, but also the overall image description from c^*. It serves as a part of the input to Stable Diffusion to provide semantic guidance for augmented image generation. As for image diversity, we apply Gaussian noise with a noise rate n in Stable Diffusion to make slight changes to the original image based on the above semantic constraint.

Considering the different contributions of l and c^* to p, we adopt the **prompt weighting** strategy. Specifically, we apply different weights w_l, w_c for the label and caption respectively by multiplying the token embeddings of l and c^* by w_l and w_c respectively. Furthermore, to control the extent of semantic guidance on image generation, we adopt the **guidance mapping** strategy that provides a *proper* guidance

		C-100	C-10	Cal101	Cars	Flowers	Pets	DTD	Avg.
ResNet-50	Backbone	75.12	94.81	47.38	82.26	33.02	50.30	19.85	57.53
	+ RE	76.35	95.34	46.26	80.73	34.62	49.30	20.65	57.61
	+ CutMix	**77.56**	**95.52**	49.28	83.55	33.35	51.21	20.14	58.66
	+ MoEx	76.00	95.23	52.74	81.13	36.27	55.79	21.20	59.77
	+ RA	76.40	95.00	58.55	86.85	41.97	57.45	25.02	63.03
	+ Text2Img	74.99	94.97	57.60	85.44	42.06	67.08	31.18	64.76
	+ SGID	75.72	95.48	**59.17***	**88.53***	**45.61***	**73.71***	**37.19***	**67.92***
	+ DC	75.14	95.10	58.29	87.86	43.66	69.73	33.61	66.20
	+ IP	75.09	94.70	56.83	86.32	42.30	68.14	30.22	64.80
	+ SGID & MoEx	77.54	95.68	70.56	87.33	49.20	76.52	38.92	70.82
	+ SGID & RA	77.56	95.49	**75.94**	**91.07**	**55.71**	**82.98**	**49.78**	**75.50**
	+ Text2Img & SGID	75.40	95.06	64.20	87.94	42.36	69.28	35.79	67.15
ViT	Backbone	89.86	98.61	91.20	82.99	95.70	92.04	71.22	88.80
	+ RE	89.87	98.67	91.71	83.94	96.28	92.55	72.07	89.30
	+ CutMix	89.94	98.67	91.77	84.20	96.53	92.45	71.32	89.27
	+ MoEx	90.11	98.67	92.24	85.94	96.78	92.66	70.79	89.60
	+ RA	90.32	98.60	91.92	85.62	96.66	92.91	71.32	89.62
	+ Text2Img	**92.66**	**99.01**	91.21	85.81	96.93	92.67	70.57	89.84
	+ SGID	**92.66**	98.96	**93.91***	**86.73***	**97.16***	**93.38**	**73.35***	**90.88***
	+ DC	92.59	98.66	93.44	86.33	97.12	**93.38**	72.30	90.55
	+ IP	92.53	98.50	92.19	85.82	97.04	93.02	71.70	90.11
	+ SGID & MoEx	92.02	98.83	94.00	87.80	96.69	92.99	73.45	90.83
	+ SGID & RA	91.64	98.72	**94.30**	**88.28**	97.13	**94.00**	**74.14**	**91.17**
	+ Text2Img & SGID	**92.69**	**99.03**	91.35	85.84	**97.17**	93.30	71.27	90.09
CLIP-ViT	Backbone	85.89	95.04	92.96	85.13	91.46	93.69	65.81	87.14
	+ RE	86.26	95.27	94.32	85.26	91.53	93.88	66.33	87.55
	+ CutMix	85.98	95.20	93.85	85.43	91.64	**94.01**	66.20	87.47
	+ MoEx	86.02	95.16	93.81	85.79	91.97	93.76	**67.18**	87.67
	+ RA	86.08	95.32	93.79	86.84	91.88	93.95	66.74	87.80
	+ Text2Img	85.80	94.77	93.72	86.80	91.90	93.71	64.50	87.31
	+ SGID	**86.53***	**95.66***	94.29	**87.19***	**92.04***	93.98	66.25	**87.99***
	+ DC	86.07	95.44	**94.33**	87.12	92.01	93.97	66.04	87.85
	+ IP	86.13	95.20	93.88	87.06	91.93	93.92	66.00	87.73
	+ SGID & MoEx	86.59	95.42	93.88	86.34	92.05	**94.19**	**67.51**	88.00
	+ SGID & RA	**86.66**	**95.68**	**94.52**	**88.06**	**92.75**	94.13	67.45	**88.46**
	+ Text2Img & SGID	85.87	95.16	93.76	86.93	92.04	93.74	64.63	87.45

Table 1: Accuracy of seven image classification datasets and three backbones by four baselines. On each backbone, the performance of the backbones, the perturbation-based methods, the generative methods (including SGID), the integrated methods are provided. The best and second best results in the DA method and integrated methods for each dataset are bolded and underlined. The numbers with * indicate that the im-provement of SGID is statistically significant with $p < 0.05$ under t-test.

scale g. The guidance scale g control how much the image generation process follows the semantic guidance, *i.e.*, the textual prompt p. We apply a function f to map the similarity s^* between the original image x and the chosen caption c^*, to the guidance scale $g = f(s^*)$. Finally, Stable Diffusion generates the augmented images given the above elements:

$$x' = \text{StableDiffusion}(x, p, g, n), \tag{3}$$

where x, x' is the original and augmented image, p is the textual prompt, g is the guidance scale, and n is the noise rate.

4 Experiments

4.1 Experimental Settings

Datasets: We evaluate the effectiveness of our proposed method on seven commonly used datasets, including three *coarse-grained* object classification datasets: CIFAR-10, CIFAR-100 (Krizhevsky 2009), Caltech101 (Cal101) (Fei-Fei, Fergus, and Perona 2004), and four *fine-grained* object classification datasets: Stanford Cars (Cars) (Krause et al. 2013), Flowers102 (Flowers) (Nilsback and Zisserman 2008), OxfordPets (Pets) (Parkhi et al. 2012) and texture classification DTD (Cimpoi et al. 2014).

Backbones: We conduct experiments on three backbones, including a basic model from scratch: ResNet-50 (from scratch) (He et al. 2016), and two pre-trained models: ViT (ImageNet-21k) (Dosovitskiy et al. 2020), CLIP-ViT (LAION-2B) (Cherti et al. 2022). Specifically, ViT (ImageNet-21k) is supervised trained on ImageNet-21k, while CLIP-ViT (LAION-2B) is self-supervised pre-trained on the CLIP paradigm on almost the same pre-trained corpus as the image generation model Stable Diffusion (SD).[3]

Baselines: We apply various DA methods introduced in Sec. 2 as baselines, including four perturbation-based methods: *Random Erasing* (RE), *CutMix*, *MoEx*, and *RandAugment* (RA), and three generative methods: *Text2Img*, *SGID+DiverseCaption* (*SGID+DC*), and *SGID+InstructPix2Pix* (*SGID+IP*). All generative methods employ the same image generation model SD. we re-implement Text2Img, SGID+DC, and SGID+IP to provide extensive experiments across datasets and backbones.

4.2 Implementation Details

We apply nucleus sampling and beam search to respectively generate 10 captions by BLIP. The caption length is between 5 and 20. We use $p = 0.9$ by default in nucleus sampling and num_beams = 3 by default in beam search. The default "CLIP-ViT-B/32" model is used for calculating image-text similarity. We apply the pre-trained "stable-diffusion-v1-5" model and generate one augmented image for each original image. [4] Empirically, we take $f(s^*) = -4 \cdot (s^*)^2 + 2 \cdot s^* + 1$ as the guidance mapping function. We select the noise rate n from $\{0.3, 0.5, 0.7\}$.[5] As for prompt weighting, We assign a weight of 1.50 to the labels because it carries more accurate information for the original image, and a weight of 0.90 to the caption to reduce the interference caused by the potential low-quality captions. We run each method over 5 different random seeds. See Appendix Sec. H for more details including the training of image classifiers.

4.3 Main Results

In this paper, we conduct experiments on three backbones with seven strong DA baselines on seven datasets. Our analysis of these results is based on three perspectives: ***(1)*** overall performance on three backbones; ***(2)*** average performance gain compared to the best baselines; ***(3)*** average performance gain for combined models. We believe the significant performance gain across the above backbones, datasets, and baselines demonstrates the effectiveness and generalizability of SGID. Our main results are shown in Table 1 and Appendix Sec. C.[6]

For ***(1)***, overall, SGID shows positive effects and achieves the highest performance on average across all seven datasets and three backbones. Specifically, our method leads to 10.39% accuracy gains on ResNet-50 (from scratch), 2.08% on ViT (ImageNet-21k), and 0.85% on CLIP-ViT (LAION-2B). This demonstrates that augmenting images with semantic guidance to diffusion models can benefit different backbones. Notably, SGID still shows improvement on CLIP-ViT (LAION-2B), whose pre-training data is almost identical to SD. This demonstrates the effectiveness of the paradigm introduced in SGID: preserving **semantic consistency** in original images and simultaneously bringing good **image diversity**, which will be further discussed in Sec. 4.4.

For ***(2)***, with semantic guidance, SGID performs comparably or better than the best perturbation-based and generative baselines. Specifically, SGID outperforms *RandAugment* and *SGID+DiverseCaption* by 4.89% and 1.72% on average on ResNet-50 (from scratch), by 1.26% and 0.33% on ViT (ImageNet-21k), and by 0.19% and 0.14% on CLIP-ViT (LAION-2B). Higher performance than four strong baselines shows promising capabilities of SGID to generate images with diffusion models under semantic guidance, *i.e.*, balancing diversity and semantic consistency.

For ***(3)***, SGID can be combined with perturbation-based and generative baselines for further improvement. We separately explore applying RandAugment based on SGID and applying SGID based on Text2Img.[7] We find the above inte-

[3]For more details of CLIP-ViT (LAION-2B), see Appendix Sec. H

[4]SD v1-5 is mainly pre-trained on LAION-2B (Schuhmann et al. 2022).

[5]The larger n brings more variation. We choose $n \in [0, 1]$ from $\{0.3, 0.5, 0.7\}$ to preserve image semantics and bring explicit changes in the background, position, etc.

[6]In Table 1, "ResNet-50" means ResNet-50 (from scratch), "ViT" means ViT (ImageNet-21k), and "CLIP-ViT" means CLIP-ViT (LAION-2B). "C-100" and "C-10" mean CIFAR-100 and CIFAR-10. "RE" and "RA" mean Random Erasing and RandAugment. "DC" and "IP" mean DiverseCaption and InstructPix2Pix.

[7]Applying RandAugment based on SGID" indicates conducting the RandAugment transformation based on the augmented images of SGID. "Applying SGID based on Text2Img" indicates conducting SGID based on the augmented images of Text2Img. We provide more results of combined models in Appendix Sec. A.

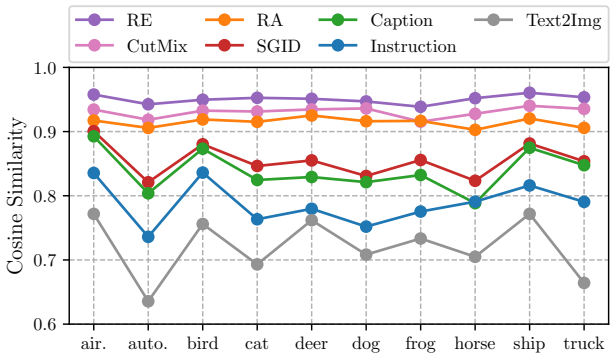


Figure 3: Average cosine similarity between the augmented and original images for each category of CIFAR-10 (air.: airplane, auto.: automobile).

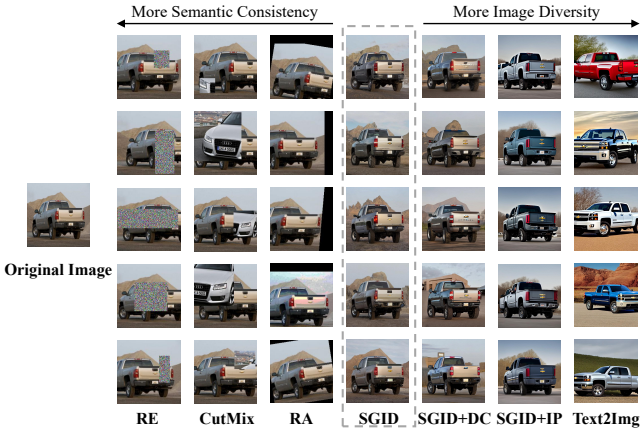


Figure 4: Case study of seven DA methods.

grated methods achieve further improvements, and this conclusion holds on three backbones. For example, as a combination of "perturbation-based & generation-based", "SGID & RA" exceeds RA on three backbones by 12.47%, 1.55% and 0.66%, and exceeds SGID by 7.58%, 0.29% and 0.47%. Consistent and significant performance gains further prove our SGID not only preserves the essential semantics of the original images while bringing good diversity, but also benefits mutually with the perturbation-based method. Interestingly, although "generation-based" methods, Text2Img and our SGID, share the same image generation model, "Text2Img & SGID" still achieves performance gains compared with Text2Img (2.39%, 0.25% and 0.14%), but underperforms SGID (-0.77%, -0.79% and -0.54%). We attribute the gains to the semantic guidance introduced by our SGID, but attribute the reductions to the fact that Text2Img may incorrectly change the essential semantics of original images.

4.4 Image Diversity and Semantic Consistency

In this section, we aim to discuss the image diversity and semantic consistency of SGID and other baselines from three perspectives: ***(1)*** human evaluation; ***(2)*** automatic similarity evaluation; ***(3)*** case study. We try to analyze the potential reason why SGID achieves better performance than existing perturbation-based baselines and generative methods.

Human Evaluation We apply human evaluation on one coarse-grained object classification dataset (Caltech101), one fine-grained object classification dataset (OxfordPets), and the texture classification dataset (DTD).[8] For each dataset, we randomly choose 10 labels and 10 of their corresponding original images. We compare SGID with the best perturbation-based and generative baselines: RandAugment, Text2Img, and SGID+DiverseCaption. We evaluate the augmented images by four DA methods based on the original image. Each augmented image is scored on a scale of $1 \sim 5$ in terms of image diversity and semantic consistency respectively. We employ three experienced annotators. The annotators are trained and pass trial annotations. Each annotator spends an average of 4.5 hours on annotation, and the salary for labeling each piece of data is \$1. Table 2 shows the human evaluation results for four methods and their corresponding average performance on ResNet-50 (from scratch).

We can find that SGID has similar semantic consistency and more image diversity than RandAugment, but more semantic consistency and less image diversity than Text2Img and SGID+DiverseCaption. When both image classification performance and human evaluation results are considered, our SGID achieves the best performance by balancing image diversity and semantic consistency through semantic-guided generative image augmentation.

Automatic Similarity Evaluation We choose CIFAR-10 for automatic similarity evaluation and separately use SGID and six DA baselines to generate five augmented images for each original image.[9] For each DA method, we calculate the average cosine similarity between the original image and its five augmented ones (Zhang et al. 2022). We repeat this process on all original images and calculate the average value for each label as a measure of diversity. The lower the average similarity between the augmented images and the original image, the lower the semantic consistency but the more significant the diversity brought by the corresponding DA method. The results are shown in Figure 3.

Overall, SGID has a lower similarity (0.8548) compared to perturbation-based DA methods, while it has a higher similarity compared to other generative methods. The average similarity of our SGID is between the two categories of DA methods, but SGID performs best in the image classification task. This further demonstrates the significance of balancing image diversity and semantic consistency.

Case Study Figure 4 compares the augmented images generated by SGID and the other six baselines. The three perturbation-based methods bring diversity through transformations. However, these pre-defined transformations could not provide sufficient diversity for augmented images.

[8]The image size of CIFAR-10 and CIFAR-100 are $32 * 32$, which is too small for human evaluation.

[9]We do not include MoEx since it performs the transformation in an implicit feature space instead of the explicit input space.

	Cal101		DTD		Pets		Avg.		Avg.
	Con.	Div.	Con.	Div.	Con.	Div.	Con.	Div.	Performance
RandAugment	4.49	1.30	4.51	1.50	4.70	1.34	4.57	1.38	63.03
SGID	4.07	1.99	4.40	1.76	4.52	1.98	4.33	1.91	**67.92**
SGID+DC	3.56	2.33	3.33	2.71	3.76	2.58	3.55	2.54	<u>66.20</u>
Text2Img	1.67	4.56	1.27	4.87	1.41	4.62	1.45	4.68	64.76

Table 2: Human evaluation results of four DA methods on three datasets from the perspectives of semantic consistency (Con.) and Diversity (Div.). "SGID+DC" indicates SGID+DiverseCaption.

Generative baselines bring more diverse and vivid images than perturbation-based baselines, but they struggle to preserve the semantic consistency of original images. In contrast, SGID preserves the semantic consistency from original images and provides good image diversity, which also results in a performance improvement in downstream tasks.

4.5 Ablation Study

In this section, we study the influence of some essential components in SGID, including ***(1)*** the construction of the textual prompt; ***(2)*** the caption filter and the prompt weighting; ***(3)*** the noise rate and the guidance scale. We further explore augmented image filtering in Appendix Sec. B.

Influence of Textual Prompts We study the influence of textual prompts based on ResNet-50 (from scratch) across four datasets in Table 3. We have the following conclusions: ***(i)*** The semantic guidance is important for the image-to-image paradigm in SGID. In most cases, *i.e.*, *Cars*, *Pets*, and *DTD*, the performance of "w/o Prompt" is the lowest. This means that semantic guidance is essential for image augmentation to preserve semantic consistency. ***(ii)*** The label and the caption both provide semantic constraints. In most cases except for *Cars*, "+ Complete Prompt" shows the best results. As for *Cars*, "+ Label Only" has higher performance than "+ Complete Prompt". We attribute that the dataset is fine-grained and its labels carry very detailed information like "Chevrolet_Silverado_1500_Extended_Cab_2012", which is sufficient for the PMs to generate an augmented image based on the original one. However, BLIP struggles to generate fine-grained captions, thus the generated captions have a counterproductive effect. ***(iii)*** In most cases, "+ Label Only" outperforms "+ Caption Only" since the image label carries groundtruth and accurate image semantics. However, for *CIFAR-100*, "+ Caption Only" shows a better result than "+ Label Only". We attribute to the short label, *e.g.*, bed, forest, etc., provided by *CIFAR-100* cannot provide detailed information in the label like other fine-grained datasets.

Caption Sampling & Prompt Weighting We study the influence of caption sampling and prompt weighting based on ResNet-50 (from scratch) in Table 4. Nucleus sampling, beam search, and the caption filter contribute the best performance on the three datasets respectively. This may be because of the different characteristics of nucleus sampling and beam search, as mentioned in Section 3.2. Nucleus sampling generates more diverse and surprising captions, while beam search tends to generate safe captions (Li et al. 2022). We hypothesize that *Pets* is a fine-grained dataset whose labels are very detailed and a "safe" caption would not influence the semantics of the labels. In contrast, short labels in *CIFAR-100* cannot convey sufficient semantics, thus more diverse captions from nucleus sampling are required. *DTD*, the texture classification dataset, is challenging for BLIP to generate captions ensuring image semantics. Then the caption filter would help to improve caption quality for semantic consistency. As for prompt weighting, its effectiveness is proven in two out of three datasets. We hypothesize that adding different weights to the label and the caption could reconcile their different semantic information.

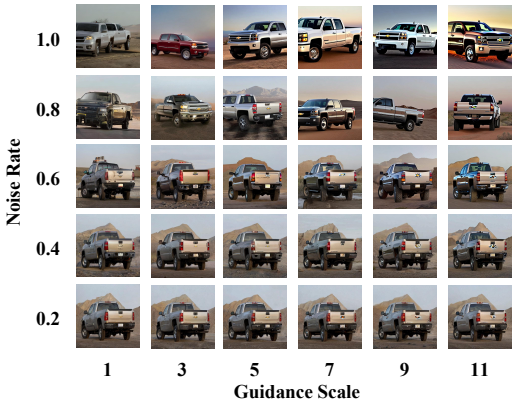


Figure 5: Case study on the influence of noise rate and guidance scale when generating images by SGID on Cars.

Influence of Noise Rate and Guidance Scale. We further explore the effect of adjusting the noise rate and guidance scale on generating augmented images. As shown in Figure 5(b), the generated images show more diversity (*e.g.*, variations in body orientation, color, background, etc.) when the noise rate and the guidance scale are increased. However, the experimental results show that as the noise rate increases, the performance decreases while more diversity is introduced. This suggests that noise rate has a great effect on performance and deserves further exploration in future work.

	CIFAR-100	Cars	Pets	DTD
w/o Prompt	74.11	82.09	55.76	32.39
+ Caption Only	<u>75.65</u>	83.89	63.80	34.78
+ Label Only	73.99	**88.53**	<u>75.90</u>	<u>37.43</u>
+ Complete Prompt	**75.72**	<u>84.31</u>	**76.15**	**37.55**

Table 3: Ablation study of textual prompts. "w/o Prompt" indicates no semantic guidance, "+ Caption Only", "+ Label Only", and "Complete Prompt" use the caption only, the label only, and the complete prompt as the guidance.

	CIFAR-100		Pets		DTD	
	PW	w/o PW	PW	w/o PW	PW	w/o PW
Beam	75.60	74.69	75.06	**76.15**	37.47	35.79
Nucleus	**75.72**	74.93	73.72	75.70	37.16	37.04
Caption Filter	74.80	74.85	70.86	71.95	**37.55**	32.44

Table 4: Ablation study of caption sampling and prompt weighting. "Caption Filter" indicates using CLIP to choose the caption with the highest similarity to the original image among the 20 captions generated by "Beam" and "Nucleus". "PW" indicates the prompt weighting strategy.

5 Related Works

Diffusion Model-Based Methods. There are some works employing diffusion models to generate augmented images or expand datasets. Zhang et al. (2022), He et al. (2022), and Dunlap et al. (2023) generate images based on label-related constraints and/or original images. Li et al. (2023) and Shipard et al. (2023) generate samples for training zero-shot classifiers without relying on original images. While acknowledging the importance of zero-shot learning, this paper focuses on linear probing to improve model performance in image classification. Akrout et al. (2023), Carlini et al. (2023), and Bakhtiarnia, Zhang, and Iosifidis (2023) apply diffusion models for DA in other tasks like skin disease classification, privacy attacks, and crowd counting. These are promising directions we would explore in the future.

Knowledge Distillation (KD). There have been some works using PMs to generate training data (Meng et al. 2022; Wang et al. 2021) and referring to it as a variant of KD (Ye et al. 2022). To some extent, our work could also be categorized as KD. Semantic-guided image generation via pre-trained diffusion models can be seen as a more straightforward and effective form to precisely distill knowledge from pre-trained diffusion models. Moreover, it can be naturally combined with the traditional KD as a new direction.

6 Conclusion

In this paper, we introduce **SGID**, a **S**emantic-guided **G**enerative **I**mage augmentation method with **D**iffusion models for image classification. The proposed method balances image diversity and semantic consistency. Specifically, SGID constructs prompts with image labels and captions as semantic guidance to generate augmented images that preserve the essential semantics in original images and simultaneously bring good image diversity. We demonstrate the effectiveness of SGID by experiments on three backbones with seven strong image augmentation baselines on seven different datasets and SGID outperforms the backbones on all datasets and achieves the best or comparable performance to all baselines. Moreover, SGID can be combined with other image augmentation baselines and further improves the overall performance. We also evaluate the semantic consistency and image diversity of SGID through quantitative human and automated evaluations, as well as qualitative case studies.

Acknowledgments

We gratefully acknowledge the support of the National Natural Science Foundation of China (NSFC) via grant 62236004 and 62206078, and the support of Du Xiaoman (Beijing) Science Technology Co., Ltd.

References

Akrout, M.; Gyepesi, B.; Holló, P.; Poór, A. K.; Kincso, B.; Solis, S.; Cirone, K. D.; Kawahara, J.; Slade, D.; Abid, L.; Kov'acs, M.; and Fazekas, I. 2023. Diffusion-based Data Augmentation for Skin Disease Classification: Impact Across Original Medical Datasets to Fully Synthetic Images. *ArXiv*, abs/2301.04802.

Algan, G.; and Ulusoy, I. 2021. Image classification with deep learning in the presence of noisy labels: A survey. *Knowledge-Based Systems*.

Bakhtiarnia, A.; Zhang, Q.; and Iosifidis, A. 2023. PromptMix: Text-to-image diffusion models enhance the performance of lightweight networks. *ArXiv*, abs/2301.12914.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. arXiv:2211.09800.

Burg, M. F.; Wenzel, F.; Zietlow, D.; Horn, M.; Makansi, O.; Locatello, F.; and Russell, C. 2023. A data augmentation perspective on diffusion models and retrieval. *ArXiv*, abs/2304.10253.

Carlini, N.; Hayes, J.; Nasr, M.; Jagielski, M.; Sehwag, V.; Tramèr, F.; Balle, B.; Ippolito, D.; and Wallace, E. 2023. Extracting Training Data from Diffusion Models. *ArXiv*, abs/2301.13188.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2022. Reproducible scaling laws for contrastive language-image learning. *ArXiv*, abs/2212.07143.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proc. of CVPR*.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proc. of CVPR*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv*, abs/2010.11929.

Dunlap, L.; Umino, A.; Zhang, H.; Yang, J.; Gonzalez, J. E.; and Darrell, T. 2023. Diversify Your Vision Datasets with Automatic Diffusion-Based Augmentation. *ArXiv*, abs/2305.16289.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. *2004 Conference on Computer Vision and Pattern Recognition Workshop*, 178–178.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proc. of CVPR*.

He, R.; Sun, S.; Yu, X.; Xue, C.; Zhang, W.; Torr, P. H. S.; Bai, S.; and Qi, X. 2022. Is synthetic data from generative models ready for image recognition? *ArXiv*, abs/2210.07574.

Holtzman, A.; Buys, J.; Forbes, M.; and Choi, Y. 2019. The Curious Case of Neural Text Degeneration. *arXiv preprint arXiv:1904.09751*.

Krause, J.; Deng, J.; Stark, M.; and Fei-Fei, L. 2013. Collecting a large-scale dataset of fine-grained cars.

Krizhevsky, A. 2009. Learning multiple layers of features from tiny images.

Li, A. C.; Prabhudesai, M.; Duggal, S.; Brown, E. L.; and Pathak, D. 2023. Your Diffusion Model is Secretly a Zero-Shot Classifier. *ArXiv*, abs/2303.16203.

Li, B.; Hou, Y.; and Che, W. 2022. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3.

Li, B.; Wu, F.; Lim, S.-N.; Belongie, S. J.; and Weinberger, K. Q. 2021. On feature normalization and data augmentation. In *Proc. of CVPR*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. C. H. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

Meng, Y.; Huang, J.; Zhang, Y.; and Han, J. 2022. Generating Training Data with Language Models: Towards Zero-Shot Language Understanding. *ArXiv*, abs/2202.04538.

Minaee, S.; Boykov, Y.; Porikli, F. M.; Plaza, A. J.; Kehtarnavaz, N.; and Terzopoulos, D. 2021. Image segmentation using deep learning: A survey. *PAMI*.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *Proc. of ICVGIP*.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. V. 2012. Cats and dogs. In *Proc. of CVPR*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning transferable visual models from natural language supervision. In *Proc. of ICML*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proc. of CVPR*.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. *ArXiv*, abs/2210.08402.

Shen, T.; Lei, T.; Barzilay, R.; and Jaakkola, T. 2017. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30.

Shipard, J.; Wiliem, A.; Thanh, K. N.; Xiang, W.; and Fookes, C. 2023. Diversity is Definitely Needed: Improving Model-Agnostic Zero-shot Classification via Stable Diffusion.

Wang, Z.; Yu, A. W.; Firat, O.; and Cao, Y. 2021. Towards Zero-Label Language Learning. *ArXiv*, abs/2109.09193.

Yang, S.; Xiao, W.-T.; Zhang, M.; Guo, S.; Zhao, J.; and Furao, S. 2022. Image Data Augmentation for Deep Learning: A Survey. *arXiv preprint arXiv:2204.08610*.

Ye, J.; Gao, J.; Li, Q.; Xu, H.; Feng, J.; Wu, Z.; Yu, T.; and Kong, L. 2022. ZeroGen: Efficient Zero-shot Learning via Dataset Generation. In *Conference on Empirical Methods in Natural Language Processing*.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. J. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proc. of ICCV*.

Zhang, Y.; Zhou, D.; Hooi, B.; Wang, K.; and Feng, J. 2022. Expanding Small-Scale Datasets with Guided Imagination. *arXiv preprint arXiv:2211.13976*.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random erasing data augmentation. In *Proc. of AAAI*.