Research History and Plan

Tao Hu
University of Amsterdam
Supervisors: Cees Snoek, Pascal Mettes
taohu620@gmail.com

1. Research Statement

Problem Statement. My research focuses on achieving label-efficient learning in computer vision. When label annotation is required, several challenges need to be considered: (1) the labor-intensive process of annotation, (2) the inherent lack of labels in most data, and (3) the challenge of dealing with long-tail label distribution problems. Therefore, our approach aims to bypass these challenges by focusing on label efficiency. This can be achieved either by reducing the need for labels [1–4, 6, 7], or by completely removing them [5].

Contributions. To tackle the challenge of label efficiency, we have developed specific solutions that aim to reduce the need for labor-intensive label annotation. In meta-learning, we follow the paradigm of few-shot learning. This paradigm consists of support and query sets and has been generalized into segmentation [4], object detection [2], and video localization [7]. Technically, these methods involve feature extraction of query and support sets using corresponding backbones and similarity matching. The similarity matching between support and query features can be multi-context in segmentation [4], feature-level in object detection [2], and frame/tempo-level in video [7]. In the second direction, we apply Mixup on point clouds and use optimal transport to seek correspondence. This approach has been applied to various point cloud tasks, achieving impressive results in several benchmarks [1]. Thirdly, we explore label-efficient learning from the perspective of generative image diffusion [5]. Rather than relying on human-annotated guidance, we seek the guidance signal in a self-supervised principle to achieve various granularities of guidance in the diffusion model, as shown in Figure 1. Finally, we explore flow-matching, a new superclass for the diffusion model, to enhance label-efficiency in image editing [6] and text generation [3].

2. Research Progress

AAAI2019: Meta-learn to segment [4]. To address label-efficient learning in image segmentation, we propose

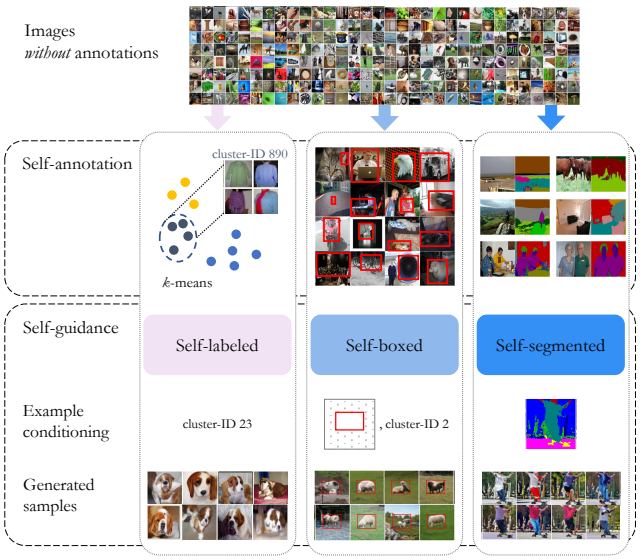

Figure 1. **Research highlight: self-guided diffusion [5].** Our method can leverage large and diverse image datasets *without* any annotations for training guided diffusion models. Starting from a dataset without ground-truth annotations, we apply a self-supervised feature extractor to create self-annotations. Using these, we train diffusion models with either self-labeled, self-boxed, or self-segmented guidance that enable controlled generation and improved image fidelity.

the Attention-based Multi-Context Guiding (A-MCG) network. It consists of three branches: support, query, and feature fusion. A-MCG integrates multi-scale context features for better guidance from the support set. We use spatial attention to enhance self-supervision in one-shot learning and employ ConvLSTM for improved fusion in multi-shot learning, and achieve competitive results on several benchmarks.

ICCV2019: Meta-learn to locate box [2]. We explore label-efficient learning in object detection through the task of few-shot common-localization, where we aim to localize the common object in a query image using only a few weakly-supervised support images, without box anno-

tations. Our proposed network leverages a spatial similarity module to identify spatial commonalities among the images and a feature reweighting module that balances the influence of different support images through graph convolutional networks. We evaluate our approach on repurposed Pascal VOC, MS-COCO, and Imagenet VID datasets, demonstrating the importance of spatial similarity search and feature reweighting, outperforming baselines from related tasks.

ECCV2020: Meta-learn to locate temporally in video [7]. In video action localization, we tackle the challenge of label-efficient learning by proposing a few-shot common action localization approach. Instead of relying on labeled examples with action details, we localize actions in untrimmed videos using just a few trimmed video examples of the same action, without knowing their class label. We introduce a novel 3D convolutional network architecture that aligns support video representations with query video segments for accurate localization. Our approach demonstrates effectiveness and general applicability in various scenarios, showcasing the potential of label-efficient learning in video action localization.

ECCV2020: Point Cloud Mixup [1]. In my research, I have explored label-efficient learning for 3D point clouds. While interpolation-based data augmentation is effective in images, it cannot be directly applied to point clouds due to the absence of one-to-one point correspondence. To address this, I propose PointMixup, a shortest path linear interpolation method for point clouds. PointMixup optimally assigns the path function between two point clouds, allowing the introduction of strong interpolation-based regularizers. Experimental results demonstrate the potential of PointMixup for point cloud classification, particularly in scenarios with limited examples, and its improved robustness to noise and geometric transformations.

CVPR2023: Self-guided Diffusion Models [5]. Going beyond discriminative modelling, diffusion models have made remarkable progress in generating high-quality images, especially when guidance controls the generative process. However, guidance requires numerous annotated image pairs for training and is therefore dependent on their availability, accuracy, and unbiasedness. To achieve label-efficient guidance in diffusion models, we propose a framework for self-guided diffusion models. This is illustrated in Figure 1. Instead of relying on annotations, we leverage the flexibility of self-supervision signals to provide guidance. Our method uses a feature extraction function and a self-annotation function to provide guidance signals at various image granularities, from the level of the entire image to object boxes and even segmentation masks. Our experiments on single-label and multi-label image datasets demonstrate that self-labeled guidance consistently outperforms diffusion models without guidance and may even surpass guidance based on ground-truth labels, especially on unbalanced data. When equipped with self-supervised box or mask proposals, our method further generates visually diverse yet semantically consistent images without the need for any class, box, or segment label annotation. Self-guided diffusion is simple, flexible, and expected to be beneficial when deployed at scale.

***Submitted:* Flow Matching for label-efficiency.** Flow Matching presents several advantages as a candidate for diffusion models. I leverage Flow Matching to improve label-efficiency in both image editing [6] and text generation [3]. By Flow Matching, I explore various approaches to data generation in image editing and expedite text generation through sampling. These methods aim to enhance generation efficiency and promote label-efficiency by utilizing Flow Matching. In the first work [6], we enhance image editing using Flow Matching with a transformer-based U-ViT structure. We introduce a controllable editing space called *u-space* and propose an efficient sampling solution for adaptive step-size ODE solvers. Additionally, we present a method for fine-grained image editing using text prompts while preserving the original content. In the second work [3], we introduce FlowSeq for conditional text generation, addressing limitations in current diffusion models. FlowSeq achieves single-step text generation through a novel anchor loss, eliminating the need for costly hyperparameter optimization. Extensive evaluations demonstrate competitive performance in tasks such as question generation, open-domain dialogue, and paraphrasing. Notably, our approach achieves a remarkable 2000-fold boost in sampling speed compared to baseline methods while maintaining high quality.

3. Research Plan

In my future research plan, I will focus my energy on generative AI, with an emphasis on developing realistic, large-scale, deployable applications that can have a high impact on industry and academia. In addition to understanding the basics of different generative models, I will concentrate on solving the issues of sampling efficiency and interpretability of generative AI. My ultimate goal is to use the progress from generative models to make discriminative models better. I also aim to expand the boundaries of generative AI to more diverse areas, such as audio and remote sensing.

References

[1] Yunlu Chen*, Tao Hu*, Efstratios Gavves, Thomas Mensink, Pascal Mettes, Pengwan Yang, and Cees G. M. Snoek. Pointmixup: Augmentation for point clouds. In *ECCV*, 2020.

[2] Tao Hu, Pascal Mettes, Jia-Hong Huang, and Cees G. M. Snoek. Silco: Show a few images, localize the common object. In *ICCV*, 2019.

[3] Tao Hu, Di Wu, Yuki M. Asano, Pascal Mettes, Basura Fernando, and Cees G.M. Snoek. Flow matching for conditional text generation in a single sampling step. In *submission*, 2023.

[4] Tao Hu, Pengwan Yang, Chiliang Zhang, Gang Yu, Yadong Mu, and Cees G. M. Snoek. Attention-based multi-context guiding for few-shot semantic segmentation. In *AAAI*, 2019.

[5] Tao Hu*, David W Zhang*, Yuki M. Asano, Gertjan J. Burghouts, and Cees G. M. Snoek. Self-guided diffusion models. In *CVPR*, 2023.

[6] Tao Hu, David W Zhang, Pascal Mettes, Meng Tang, Deli Zhao, and Cees G.M. Snoek. Latent space editing in transformer-based flow matching. In *ICML Frontiers4lcd Workshop*, 2023.

[7] Pengwan Yang*, Tao Hu*, Pascal Mettes, and Cees G. M. Snoek. Localizing the common action among a few videos. In *ECCV*, 2020.