Label-Efficient Learning to See

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 [1–3, 5], or by completely removing them [4].

Contributions. To address the challenge of label-efficiency, we have designed specific solutions along several directions: meta-learning [2,3,5] and augmentation [1]. These solutions 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 [3], object detection [2], and video localization [5]. 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 [3], feature-level in object detection [2], and frame/tempo-level in video [5]. In the second direction, we apply Mixup on point cloud and use optimal transport to seek correspondence. This approach has been applied to various point cloud tasks, achieving impressive results in several benchmarks [1]. Finally, we explore label-efficient learning from the perspective of generative image diffusion [4]. 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.

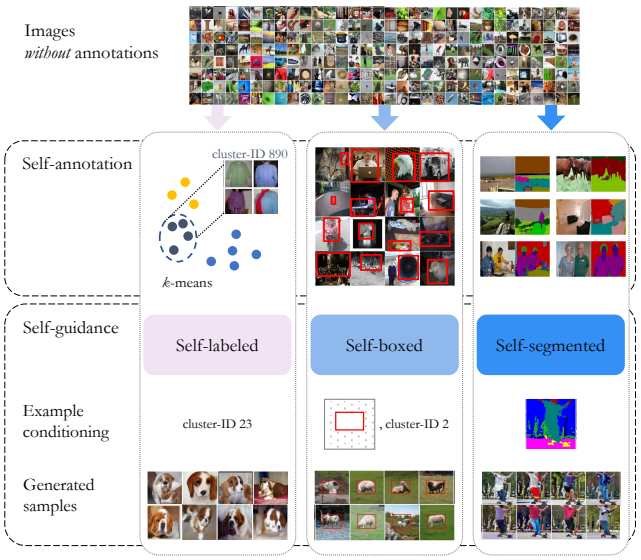

Figure 1. **Research highlight: self-guided diffusion [4].** Our method can leverage large and diverse image datasets *without* any annotations for training guided diffusion models. Starting from a dataset without ground-truth annotations, we apply a self-supervised feature extractor to create self-annotations. Using these, we train diffusion models with either self-labeled, self-boxed, or self-segmented guidance that enable controlled generation and improved image fidelity.

2. Research Progress

AAAI2019: Meta-learn to segment [3]. To address label-efficient learning in image segmentation, we propose the Attention-based Multi-Context Guiding (A-MCG) network. This network consists of three branches: a support branch, a query branch, and a feature fusion branch. A key differentiator of A-MCG is its integration of multi-scale context features between the support and query branches. This integration enforces better guidance from the support set. In addition, we adopt a spatial attention mechanism along the fusion branch to highlight context information from several scales, thereby enhancing self-supervision in one-shot learning. To address the fusion problem in multi-shot learning, we adopt ConvLSTM to collaboratively integrate sequential support features and elevate the final accuracy.

ICCV2019: Meta-learn to locate box [2]. Further-

more, we consider exploring the label-efficient learning in object detection, we call few-shot common-localization. Given a few weakly-supervised support images, we aim to localize the common object in the query image without any box annotation. This task differs from standard few-shot settings, since we aim to address the localization problem, rather than the global classification problem. To tackle this new problem, we propose a network that aims to get the most out of the support and query images. To that end, we introduce a spatial similarity module that searches the spatial commonality among the given images. We furthermore introduce a feature reweighting module, which balances the influence of different support images through graph convolutional networks. To evaluate few-shot common-localization, we repurpose and reorganize the well-known Pascal VOC and MS-COCO datasets, as well as a video dataset from Imagenet VID. Experimental evaluation on the new settings for few-shot common-localization shows the importance of searching for spatial similarity and feature reweighting, outperforming baselines from related tasks.

ECCV2020: Meta-learn to locate temporally in video [5]. Aside from image-level meta-learning, I have also investigated the label-efficient learning in video action localization by localizing the temporal extent of an action in a long untrimmed video. While existing work uses many examples with their start, end, and/or class of the action during training, we propose few-shot common action localization. This method determines the start and end of an action in a long untrimmed video based on just a handful of trimmed video examples containing the same action, without knowing their common class label. To accomplish this task, we introduce a new 3D convolutional network architecture that can align representations from the support videos with the relevant query video segments. The aligned feature is then sent to the remaining logic of video action localization. We evaluate few-shot common action localization in untrimmed videos containing single or multiple action instances and demonstrate the effectiveness and general applicability of our proposal.

ECCV2020: Point Cloud Mixup [1]. Besides label-efficient learning for visual grid structures, i.e., image and video, I have also researched it for 3D point clouds. Data augmentation by interpolation has proven to be a simple and effective approach in the image domain. However, such a mixup cannot be directly transferred to point clouds as there is no one-to-one correspondence between the points of two different objects. In this work, we define data augmentation between point clouds as a shortest path linear interpolation. To accomplish this, we introduce PointMixup, an interpolation method that generates new examples through an optimal assignment of the path function between two point clouds. We prove that PointMixup finds the shortest path between two point clouds and that the interpolation is assignment invariant and linear. With the definition of interpolation, PointMixup allows the introduction of strong interpolation-based regularizers such as mixup and manifold mixup to the point cloud domain. Experimentally, we demonstrate the potential of PointMixup for point cloud classification, especially when examples are scarce, as well as increased robustness to noise and geometric transformations of points.

CVPR2023: Self-guided Diffusion Models [4]. 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.

The thesis is awaiting approval from the PhD committee, I expect to graduate before the summer.

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, 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.

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

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