# Strategic Reduction of Training and Annotation in Computer Vision: Leveraging Pre-trained Models and Auxiliary Tasks for Efficient Learning

Yuhao Lin

Submitted for the Degree of
Doctor of Philosophy
from the
University of Adelaide




THE UNIVERSITY
*of* ADELAIDE

Australian Institute for Machine Learning
The University of Adelaide
Adelaide, South Australia, Australia.

December 2024

© Yuhao Lin 2024

Thank you, self, for not letting the dream depart.

# Abstract

In the ever-evolving field of computer vision, advancements have been largely driven by deep learning technologies and the use of large-scale datasets. However, the labor-intensive and costly nature of large-scale data annotation and training processes presents significant challenges for broader application. This thesis addresses the following key questions: 1) How can we improve existing semi-supervised learning models to maintain relatively high performance with fewer training annotations? 2) With the emergence of powerful pretrained models, can they reduce the demands for training time and annotation in downstream tasks? 3) How can we use these pretrained models to reduce or even avoid manual annotation, and how can the feature of these models themselves assist tasks that lack annotations?

Semantic segmentation has long been a labor-intensive task, requiring the precise classification of every pixel. As a result, semi-supervised learning, a well-established approach, has found widespread application in this domain. To address the first question, we revisit the use of classical autoencoding techniques in semi-supervised semantic segmentation. By incorporating foreground-only image reconstruction as an auxiliary task within a modern segmentation framework, we achieve significant performance improvements across multiple datasets.

With the advent of powerful pretrained models, fundamental tasks like image segmentation have been greatly improved. However, in class-agnostic object counting tasks, not only is class and quantity information required, but location annotations must also be provided, which demands significant human and financial resources. To tackle this challenge, we explore the application of pretrained models, including SAM and the high-dimensional self-supervised learning model (DINO). By combining these models with the transductive prototype mechanism from semi-supervised learning, we establish a new performance benchmark that not only competes with traditional training-dependent methods but occasionally surpasses them, effectively bridging the performance gap in dynamic environments without extensive training.

For the third issue, we delve deeper into the capabilities of these powerful pretrained models to generate high-quality supplementary annotations, such as keypoints, using pretrained diffusion models, thereby reducing and even avoiding the dependency on manual annotation. Additionally, we introduce a novel approach named Categorical Keypoint Positional Encoding (CKPE), which ingeniously enhances the performance of the Vision Transformer architecture. This strategy significantly decreases reliance on manual annotations while improving the accuracy and robustness of animal re-identification across diverse wildlife datasets.

Overall, this thesis demonstrates the transformative potential of hybrid learning paradigms that integrate auxiliary tasks and pretrained models to significantly improve the efficiency of computer vision applications. It lays a solid foundation for future advancements in handling precise vision tasks with limited data resources, emphasizing the crucial role of innovative learning strategies in advancing computer vision systems.

# Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

The author acknowledges that copyright of published works contained within the thesis resides with the copyright holder(s) of those works.

I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

# Publications

This thesis contains the following works that have been published or prepared to be submitted:

- **Yuhao Lin**, Haiming Xu, Lingqiao Liu, Jinan Zou, Javen Qinfeng Shi. "Revisiting Image Reconstruction for Semi-supervised Semantic Segmentation." 2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA). IEEE, 2023.
- **Yuhao Lin**, Haiming Xu, Lingqiao Liu, Javen Qinfeng Shi. "A Simple-but-effective Baseline for Training-free Class-Agnostic Counting." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2025.
- **Yuhao Lin**, Lingqiao Liu, Javen Qinfeng Shi. "Categorical Keypoint Positional Embedding for Robust Animal Re-Identification." CVPR 2025 Under Review.

In addition, I have the following papers not included in this thesis:

- Qiaoyang Luo, Lingqiao Liu, **Yuhao Lin**, and Wei Zhang. 2021. Don't Miss the Labels: Label-semantic Augmented Meta-Learner for Few-Shot Text Classification. In Findings of the Association for Computational Linguistics: ACL-IJCNLP, 2021.
- Jinan Zou, Haiyao Cao, Lingqiao Liu, **Yuhao Lin**, Ehsan Abbasnejad, Javen Qinfeng Shi. Astock: A New Dataset and Automated Stock Trading based on Stock-specific News Analyzing Model. In Fourth Workshop on Financial Technology and Natural Language Processing, 2022.
- Jinan Zou, Maihao Guo, Yu Tian, **Yuhao Lin**, Haiyao Cao, Lingqiao Liu, Ehsan Abbasnejad, Javen Qinfeng Shi. Semantic Role Labeling Guided Out-of-distribution Detection. Proceedings of the 30th International Conference on Computational Linguistics, 2024.
- Nojoud Mushabab H Alshehri, **Yuhao Lin**. GPT-4 Attempting to Attack AI-Text Detectors. Australasian Joint Conference on Artificial Intelligence. Singapore: Springer Nature Singapore, 2024.

# Acknowledgements

This thesis is the culmination of over three years of rigorous scholarship and heartfelt engagement at the University of Adelaide, particularly within the Australian Institute for Machine Learning (AIML).

Firstly, I extend my deepest appreciation to my principal supervisors, Prof.Javen Qinfeng Shi and Prof.Lingqiao Liu, who have been my scholarly guides and personal beacons throughout this journey. Their boundless patience and insightful guidance not only fostered my academic growth but also nurtured my development as an individual with integrity and resilience. Their light has guided me through challenges, for which I am profoundly grateful.

I am also thankful to Dr.Haiming Xu, whose constructive advice and steadfast support fueled my passion for research. His profound knowledge and diligent approach to research have profoundly shaped my scholarly perspective.

My journey was also made richer and more enjoyable by the company of my supportive friends and colleagues at the University of Adelaide. Their camaraderie and encouragement have been invaluable.

Lastly, my heartfelt gratitude goes to my parents. Their unconditional love and generous financial support have been the bedrock of my academic journey. Thank you for creating a nurturing environment that allowed me to pursue my studies without hindrance.

To all of you, I owe a debt of gratitude that words alone cannot express. Thank you for walking this path with me.

# Contents

**Dedication** **iii**

**Declaration** **vii**

**Publications** **ix**

**Acknowledgment** **xi**

**1 Introduction** **1**

1.1 Overview . . . . . 1

1.2 Problem Formulation . . . . . 3

1.2.1 Semi-supervised semantic segmentation . . . . . 3

1.2.2 Class-Agnostic Counting . . . . . 4

1.2.3 Wild Animal ReID . . . . . 5

1.3 Main contribution . . . . . 6

1.4 Thesis outline . . . . . 7

**2 Literature Review** **9**

2.1 Semi-Supervised Learning in Computer Vision . . . . . 9

2.2 Integration of Auxiliary Tasks in Enhancing Model Performance . . . . 10

2.3 Adapting Pretrained Models for Data-Efficient Learning . . . . . 11

**3 Revisiting Image Reconstruction for Semi-supervised Semantic Segmentation** **13**

3.1 Abstract . . . . . 16

3.2 Introduction . . . . . 17

3.3 Related Work . . . . . 20

3.3.1 Autoencoders for Semi-Supervised Learning . . . . . 20
3.3.2 Supervised Semantic Segmentation . . . . . 20
3.3.3 Semi-supervised Semantic Segmentation . . . . . 21
3.4 Image Reconstruction . . . . . 22
3.5 Image Reconstruction for Semi-supervised Semantic Segmentation . . . 23
3.5.1 Preliminary . . . . . 23
3.5.2 Semi-supervised Segmentation Baseline . . . . . 23
3.5.3 Reconstruction as an Auxiliary Task . . . . . 24
3.5.4 Improving the Reconstruction Task by Object-Background Disentanglement . . . . . 26
3.6 Experiments . . . . . 27
3.6.1 Experimental Setup . . . . . 27
3.6.2 Comparison with State-of-the-Arts . . . . . 28
3.6.3 Analysis . . . . . 31
3.7 Conclusion . . . . . 35

**4 A Simple-but-effective Baseline for Training-free Class-Agnostic Counting 37**

4.1 Abstract . . . . . 40
4.2 Introduction . . . . . 40
4.3 Related Work . . . . . 43
4.3.1 Class-Agnostic Counting Methods . . . . . 43
4.3.2 Superpixel Methods . . . . . 44
4.3.3 Vision Foundation Models . . . . . 44
4.4 Method . . . . . 45
4.4.1 Object-prior Point Prompt Matters . . . . . 46
4.4.2 Semantic-rich Feature Matters . . . . . 47
4.4.3 Multi-scale Segmentation Helps . . . . . 50
4.4.4 Transductive Prototype Updating Helps . . . . . 51
4.5 Experiment . . . . . 51
4.5.1 Experimental Setup . . . . . 51
4.5.2 Comparison with State-of-the-Arts . . . . . 52
4.6 Ablation study . . . . . 56
4.7 Conclusion . . . . . 58

**5 Categorical Keypoint Positional Embedding for Robust Animal Re-Identification 59**

5.1 Abstract . . . 61

5.2 Introduction . . . 61

5.3 Related Work . . . 65

5.3.1 Animal ReID . . . 65

5.3.2 ViT for ReID . . . 65

5.3.3 Visual foundation model with correspondences . . . 66

5.4 Method . . . 67

5.4.1 Task forming . . . 67

5.4.2 Keypoint Detection and Propagration . . . 68

5.4.3 Keypoint Positional Embedding (KPE) . . . 69

5.4.4 Categorical Keypoint Positional Embedding (CKPE) . . . 71

5.4.5 Training and Testing . . . 72

5.5 Experiment . . . 74

5.5.1 Experimental Setup . . . 74

5.5.2 Comparison with State-of-the-Arts . . . 75

5.6 Ablation study . . . 75

5.6.1 Different approaches to find keypoints . . . 75

5.6.2 Different Timesteps to Propagate the Keypoints . . . 78

5.6.3 Different ViT Scales . . . 79

5.6.4 Different numbers of keypoints . . . 79

5.6.5 Keypoints vs Random points . . . 80

5.7 Conclusion . . . 81

**6 Conclusion and Future Work 83**

6.1 Conclusion . . . 83

6.2 Future Work . . . 85

6.2.1 Semi-Supervised Semantic Segmentation . . . 85

6.2.2 Class-Agnostic Object Counting . . . 85

6.2.3 Categorical Keypoint Positional Embedding . . . 86

**Bibliography 87**

# List of Figures

3.1 Some visualization results from Pascal VOC 2012 **validation set** (the reconstruction branch is only trained on the training set). (a) input images, (b) feature maps from the reconstruction-only model, (c) feature maps from the reconstruction-segmentation model, and (d) the segmentation results of the same epoch as c. We observe that the activation areas in some feature maps are focused on the objects (i.e., birds in the first row and goats in the fourth row). . . . . 17

3.2 (a) illustrates a simple diagram of our joint training process. The model is based on the traditional teacher-student structure, and we add another decoder with the same architecture as the segmentation decoder to reconstruct the original image. When training, the optimization goal is to minimize the total loss from both segmentation CE loss and reconstruction MSE loss. In (b), we compare the performance of baseline with and without original image reconstruction on Pascal VOC 2012 [26] under the standard partition protocols. It is worth noting that even by simply reconstructing the input image, the baseline outperforms current SOTA results in all partitions. . . . . 18

3.3 Illustration of our approach. The left side is the basic MT structure which employs two segmentation networks, the student $h_s, g_s$ and the teacher $h_t, g_t$. The right side $g_{rec}$ is our Foreground-Only reconstruction module, which takes the output of the student's encoder $h_s$ and reconstructs a monotonous background image. Reconstruction loss masks are applied for loss calculation. Specifically, for labeled images, we set the background to 0 according to the ground truth, and the one-hot mask $M_{rec}^{sup}$ in shape $[H \times W \times 1]$ is all ones. For unlabeled images, the mask $M_{rec}^{unsup}$ is derived from the teacher's ignoring uncertain areas. Then the teacher is updated by the student's exponential moving average (EMA). 21

3.4 Visualizations of the latent images obtained from FOrec and standard reconstruction, from left to right, are (a) input images, (b) latent images obtained from standard reconstruction, (c) latent images obtained from FOrec, (d) segmentation results from standard reconstruction, (e) segmentation results from FOrec, and (f) segmentation ground truth. Note that the (b) and (c) are slices from the same position for a fair comparison. 33

4.1 Visualizing the performance, our method notably narrows the gap in counting accuracy between training-free and training-based approaches. Impressively, it even surpasses the top-performing training-based methods on the CARPK dataset, underscoring the potential of training-free approaches. . . . . 41

4.2 **Top** row illustrates the creation process of object-prior point prompts. **Bottom** row depicts a pipeline overview of our proposed training-free object counting approach. Details of the transductive update module are provided in Eq. 4.4. Reference objects are marked with yellow boxes in the input image (zoom in for clarity). . . . . 45

4.3 Effectiveness demonstration of the use of superpixel on the quality of mask proposals generated by SAM. **Left:** an image with reference exemplars (hot-air balloon) in black boxes. **Right:** a bubble chart illustrates the trade-off between the recall rate of the interested object and the time cost. Without increasing the number of point prompts, SAM with superpixel[1]can significantly improve the recall rate of hot-air balloons in the output mask proposals, thus avoiding the potential computational burden caused by the demand for denser point grids. Please find visualizations of the generated mask proposal in the appendix. . . . . 46

4.4 Visualization of the similarity mappings of image features to reference object features (marked in black boxes) using SAM and DINOv2. It distinctly shows that SAM's similarity mapping erroneously highlights numerous areas unrelated to the target object, whereas DINOv2's mapping accurately encompasses the objects of interest. This demonstrates that DINOv2's features possess more semantically relevant knowledge. 48

4.5 Effectiveness demonstration of our multi-scale mechanism in scenarios involving extremely tiny counting objects. Merely increasing the quantity of point prompts[2]in SAM fails to yield accurate instance-level mask proposals. However, integrating SAM with our multi-scale mechanism (using 32*32 points for aerial photography of sheep in this demonstration) effectively achieves better mask quality. . . . . 49

4.6 Qualitative results on the FSC-147 dataset are displayed. Exemplars are marked with black (white) boxes, and counting values are noted in the bottom-right corner. Error predictions are indicated by red boxes. Our method accurately predicts masks in both densely and sparsely populated scenes. Best viewed by zooming in. . . . . 55

5.1 Comparison of model performance between the previous state-of-the-art (SOTA) and our proposed method across four datasets. Our approach significantly outperforms the previous SOTA, demonstrating notable improvements in accuracy for all datasets. . . . . 62

5.2 Comparison of keypoint similarity heatmaps between a pre-trained Vision Transformer (ViT) model and a pre-trained diffusion model. The most salient feature in the heatmap, indicating the keypoint, is highlighted. (a) is the image and the keypoint, above is the mouth and bottom is the right eye. (b) is the similarity heatmaps from ViT and (c) is the similarity heatmaps from Stable Diffusion. It is obvious that the pre-trained diffusion model exhibits stronger semantic correspondence compared to the pre-trained ViT model. . . . . 63

5.3 Architecture Overview of Our Proposed Categorical Keypoint Positional Embedding for Wild Animal ReID: Initially, GPT-4 identifies critical keypoints for the ReID task from a single image. Subsequently, a diffusion model is employed to propagate these keypoints across the entire dataset. The identified keypoints then inform our ViT-based ReID module through Keypoint Positional Embedding or Categorical Keypoint Positional Embedding, focusing on the content at these specific locations and semantic information to enhance feature discrimination and accuracy. . . . . 67

# Chapter 1

# Introduction

## 1.1 Overview

The field of computer vision has experienced exponential growth, primarily driven by significant advancements in deep learning and the increased availability of large-scale datasets. Despite these technological advances, the labor-intensive and costly process of data annotation continues to be a major bottleneck, limiting the broader application and scalability of sophisticated vision technologies.

Historically, the reliance on extensive annotated datasets for training computer vision models has posed substantial challenges, necessitating significant human effort and resource investment [96]. Over time, there has been a shift from exclusively supervised learning paradigms to more data-efficient methodologies, including semi-supervised, unsupervised, and transfer learning techniques. These approaches significantly reduce dependence on large annotated corpora [11].

Recent innovations have focused on alleviating the burdens of training and annotation. One notable advancement is the use of synthetic data and data augmentation techniques, which significantly enhance the diversity of training samples without requiring additional manual annotation [105]. Additionally, the rise of self-supervised learning represents a significant step towards minimizing annotation requirements. By leveraging the inherent structure of unlabeled data, models can undergo pre-training on vast

datasets, thus reducing the need for extensively annotated data for fine-tuning specific tasks [47].

In the realm of multi-task learning (MTL), auxiliary tasks play a crucial role in enhancing model performance in computer vision by providing additional regularization and aiding in the refinement of generalized representations. By training on auxiliary tasks alongside the main task, a model can utilize shared representations to improve generalization across tasks. This integration allows the transfer of useful features and inductive biases from the auxiliary tasks to the main task, which is particularly beneficial when the primary task has limited training data. For instance, tasks like depth estimation or contour detection can serve as auxiliary tasks that enhance the feature learning process for primary tasks such as object detection or semantic segmentation[79]. These auxiliary tasks encourage the model to develop a more robust and comprehensive understanding of the visual world, capturing underlying structures or patterns that, while not explicitly targeted by the main task, are beneficial for overall performance. Implementing such tasks in computer vision not only improves accuracy but also enhances the model's ability to generalize from training data to real-world scenarios, a key challenge in the field [50].

Pretrained fundamental vision models such as SAM (Segment Anything Model) [52], DINO (Knowledge Distillation with No Labels) [83], CLIP (Contrastive Language-Image Pre-training) [88], and Stable Diffusion [93] have also made significant strides in the field by minimizing the need for extensive training datasets for downstream tasks. Trained on vast, diverse datasets, these models learn rich, generalized features that can be efficiently transferred to new tasks with minimal fine-tuning. For example, SAM has been effectively used for instance segmentation tasks across various objects, demonstrating the ability to generalize across categories without additional training [10]. DINO has achieved success in self-supervised image classification, enhancing feature extraction without labeled data. CLIP has facilitated applications in zero-shot learning, effectively categorizing images in settings it was never explicitly trained on, which illustrates its robust cross-modal understanding [88]. Similarly, Stable Diffusion supports creative tasks such as generating realistic images from textual descriptions, which has been applied in design and art generation, facilitating novel

ways of visual content creation [93]. These examples underscore how robust pretrained models reduce resource consumption and training time, making sophisticated computer vision tasks more accessible and scalable across industries.

This thesis specifically addresses the strategic reduction of training and annotation demands in computer vision. It explores the utilization of pretrained models and the integration of auxiliary tasks to enhance learning efficiency and extend practical applicability across various scenarios.

## 1.2 Problem Formulation

### 1.2.1 Semi-supervised semantic segmentation

**Problem Formulation for Semi-Supervised Semantic Segmentation**

Let $X = \{x_1, x_2, \ldots, x_N\}$ be a dataset of $N$ images where each image $x_i$ belongs to either a labeled subset $X_L$ or an unlabeled subset $X_U$, such that $X_L \cup X_U = X$ and $X_L \cap X_U = \emptyset$. The labeled subset $X_L$ has associated ground truth segmentation maps $Y = \{y_1, y_2, \ldots, y_M\}$, where $y_i$ is a pixel-wise label map for the image $x_i$ and $M \leq N$ (since not all images are labeled).

The goal of semi-supervised semantic segmentation is to learn a function $f_\theta : X \to Y$, parameterized by $\theta$, that predicts a label map $\hat{y}_i$ for each image $x_i$, minimizing the difference between $\hat{y}_i$ and $y_i$ for labeled images and making sensible predictions for unlabeled images.

**Model Training**

Training involves optimizing $\theta$ such that the model performs well on both labeled and unlabeled data. This is typically achieved through a combination of supervised and unsupervised losses:

1. **Supervised Loss ( $\mathcal{L}_{sup}$ )**: This is calculated on the labeled dataset $X_L$ using a standard segmentation loss, such as cross-entropy loss or Dice loss, computed

between the predicted label maps $\hat{y}_i$ and the ground truth label maps $y_i$:

$$\mathcal{L}_{sup}(\theta) = \frac{1}{M} \sum_{i=1}^{M} \text{Loss}(f_\theta(x_i), y_i)$$

2. **Unsupervised Loss ( $\mathcal{L}_{unsup}$ )**: This is computed on the unlabeled dataset $X_U$. Various strategies can be employed, such as consistency regularization, entropy minimization, or pseudo-labeling. A common approach is to enforce consistency under different data augmentations or model perturbations:

$$\mathcal{L}_{unsup}(\theta) = \frac{1}{N-M} \sum_{j=M+1}^{N} \text{Consistency}(f_\theta(\tilde{x}_j), f_\theta(\tilde{\tilde{x}}_j))$$

where $\tilde{x}_j$ and $\tilde{\tilde{x}}_j$ are differently augmented versions of the same image $x_j$ from $X_U$.

#### Objective Function

The overall objective function to be minimized can be a weighted combination of the supervised and unsupervised losses:

$$\mathcal{L}(\theta) = \mathcal{L}_{sup}(\theta) + \lambda \mathcal{L}_{unsup}(\theta)$$

where $\lambda$ is a hyperparameter that balances the two losses.

The semi-supervised learning framework aims to leverage the unlabeled data to provide additional training signals that help improve the generalization of the model $f_\theta$ beyond the limited labeled data, effectively utilizing the entire dataset $X$ for training.

### 1.2.2 Class-Agnostic Counting

## Formulations for Class-Agnostic Counting

#### Traditional Formulation

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a dataset of $n$ images. The objective is to train a function $f_\theta : X \to \mathbb{N}$ that predicts the count $c_i$ of objects in each image $x_i$, irrespective of object classes. Here, $\theta$ represents the model parameters, designed to be class-agnostic.

#### Objective Function

The training objective is to minimize the Mean Squared Error (MSE) between the predicted counts and the true counts:

$$\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} (f_\theta(x_i) - c_i)^2$$

### Training-Free Formulation

In a training-free approach, $f : X \to \mathbb{N}$ is used to predict the count $c_i$ of objects in each image $x_i$ without further training. This function utilizes pre-existing models trained on related tasks like segmentation.

The model $f$ is applied directly to segment or highlight distinct objects, followed by a counting mechanism that tallies these identified segments or features.

## 1.2.3 Wild Animal ReID

Animal Re-Identification (ReID) is a computer vision task aimed at identifying and matching individual animals across different non-overlapping camera captures. We formally define this task as follows:

**Given:**

- A **training set** $\mathcal{T}$ used to train the model, which includes images and their ID labels.
- A **query set** $\mathcal{Q}_{\text{test}} = \{q_1, q_2, \ldots, q_n\}$ and a **gallery set** $\mathcal{G}_{\text{test}} = \{g_1, g_2, \ldots, g_m\}$ for testing. The training and testing sets are distinct, ensuring no overlap to simulate real-world scenarios where the model encounters completely unseen samples. This separation tests the model's generalization capabilities and effectiveness in practical applications.

**Objective:**

- Learn a mapping function $f : \mathcal{I} \to \mathbb{R}^d$ that maps an image $I$ to a point in a $d$-dimensional Euclidean space. This function is optimized so that images of the same animal are closer together, and those of different animals are farther apart.
- For each query image $q_i \in \mathcal{Q}_{\text{test}}$, find the set $\{g_j \in \mathcal{G}_{\text{test}} : \text{identity}(q_i) = \text{identity}(g_j)\}$, using the trained model to predict and match identities based on learned representations.

## 1.3 Main contribution

This thesis makes significant advancements in the field of data-efficient learning by tackling three distinct challenges in computer vision: semi-supervised semantic segmentation, class-agnostic object counting, and animal re-identification. Each project contributes novel approaches that not only enhance the theoretical understanding of their respective areas but also provide practical solutions to real-world problems. The key contributions are outlined as follows:

- Enhancement of Semi-Supervised Semantic Segmentation: This research revitalizes the use of autoencoding, a classical technique, by integrating it with a contemporary semi-supervised learning framework for semantic segmentation. The innovative aspect of this project lies in demonstrating that image reconstruction, when used as an auxiliary task, can significantly leverage unlabeled data. This integration enhances the model's ability to extract and interpret semantic features, resulting in improved segmentation accuracy. The study introduces a modification to the reconstruction task that more effectively separates object features from background, which is shown to improve performance across diverse datasets.
- Innovative Approach to Class-Agnostic Object Counting: The thesis introduces a pioneering approach to object counting that operates independently of object class knowledge and without the need for specific training. By adapting pre-existing models with advanced image processing technologies and employing techniques

like superpixel algorithms and enriched encoding methods, this work establishes a robust baseline for training-free, class-agnostic counting. This approach not only simplifies the deployment of counting technologies but also matches, and in some cases surpasses, the performance of traditional training-based systems, broadening its applicability in various dynamic environments.

- Advancements in Animal Re-Identification: Focusing on ecological conservation efforts, this project significantly enhances the accuracy and efficiency of animal re-identification systems. By developing a keypoint propagation mechanism that utilizes a pre-trained diffusion model, the research improves the Vision Transformer architecture with the introduction of Keypoint Positional Encoding and Categorical Keypoint Positional Embedding. These enhancements facilitate more detailed and accurate feature representations, substantially improving re-identification accuracy across multiple wildlife datasets, thereby aiding in the tracking and study of animal populations more effectively.

Collectively, these contributions not only address pressing challenges in the field but also set new directions for future research, underscoring the potential of data-efficient learning strategies to transform the capabilities of computer vision systems in diverse applications.

## 1.4 Thesis outline

In Chapter 2, we review previous state-of-the-art methods in semi-supervised semantic segmentation, class-agnostic counting, and animal re-identification. We also provide a detailed literature review on the basics of data-efficient learning methods and their relevance in computer vision.

In Chapter 3, we explore the problem of semi-supervised semantic segmentation and make two major contributions to this field: (1) We reintroduce and adapt the classical autoencoding method as an auxiliary task within a modern semi-supervised learning framework to enhance model performance with limited labeled data. (2) We propose

modifications to the image reconstruction task that significantly disentangle object features from background noise, demonstrating improvements across various datasets.

In Chapter 4, we develop a novel approach to class-agnostic object counting that does not rely on prior specific training. Our proposed solution includes adapting pre-existing models with advanced image processing techniques such as superpixel algorithms and enriched encoding methods, establishing a new baseline for training-free, class-agnostic counting that performs on par with traditional methods.

In Chapter 5, we address the challenge of animal re-identification crucial for ecological research. We propose an innovative keypoint propagation mechanism that utilizes a pre-trained diffusion model to reduce the need for manual annotation. We enhance the Vision Transformer architecture with Keypoint Positional Encoding and Categorical Keypoint Positional Embedding to improve the accuracy and robustness of the identification process across multiple wildlife datasets.

In Chapter 6, the conclusion and potential future works are discussed. We reflect on the contributions of the thesis to the field of data-efficient learning in computer vision and outline possible directions for further research that could continue to push the boundaries of what is achievable with minimal data inputs in various computer vision tasks.

# Chapter 2

# Literature Review

## 2.1 Semi-Supervised Learning in Computer Vision

Semi-supervised learning (SSL) in computer vision leverages both labeled and unlabeled data to train models, addressing the challenge of insufficient labeled data. This approach is increasingly relevant as the demand grows for large models trained on extensive datasets, as indicated by scaling laws suggesting that bigger data and models enhance performance [112, 128].

A key technique in SSL is pseudo-labeling, where models generate their own labels (pseudo-labels) for unlabeled data, facilitating training when labeled examples are scarce. Originally detailed by Lee[57], pseudo-labeling is akin to entropy minimization, which enhances classification by positioning similar labels closely on the data manifold, separated by low-density regions. This concept underscores two principal assumptions in SSL: the low density separation and the cluster assumptions, indicating that closely clustered points likely belong to the same class [11].

The literature has evolved to incorporate methods like consistency regularization, where models ensure that slightly altered versions of the same data point (via augmentations or perturbations) are similar in embedding or output space. This technique complements entropy minimization by promoting stable predictions across variations of the same input, typically using hard pseudo-labels to sharpen the learning focus [8, 80].

Pseudo-labeling also intersects with unsupervised sub-fields such as self-supervised learning and knowledge distillation, where it aids in transferring knowledge between models or leveraging inherent data structures without explicit labels. In these scenarios, neural networks either approximate a classification function or transfer knowledge from a 'teacher' model to a 'student' model, essentially providing a form of supervision that supports model training in the absence of labels [40, 72, 122].

This synthesis of SSL and unsupervised learning methods points to a broader integration of strategies that optimize model performance by harnessing unlabeled data, shaping future directions in machine learning research across various domains.

## 2.2 Integration of Auxiliary Tasks in Enhancing Model Performance

Multi-task learning (MTL) is a machine learning paradigm designed to handle multiple related tasks simultaneously, thereby improving model performance and learning efficiency. By integrating auxiliary tasks, MTL leverages the inherent interconnections between tasks to enhance generalization across all tasks, which is particularly effective in addressing the challenge of data sparsity. This approach aggregates sparse data from various tasks, functioning as an advanced form of data augmentation to improve the accuracy of each task [140]. In the era of "big data", especially in fields like computer vision and natural language processing, deep MTL models are more effective than single-task models at utilizing large datasets, resulting in more robust and generalized representations.

Evidence suggests that integrating auxiliary tasks within the MTL framework significantly enhances learning efficiency. Although auxiliary tasks may not be the primary focus, they provide supplementary information that often boosts the performance of the main tasks. For instance, [68] highlights that depth estimation and semantic segmentation are frequently used together in multi-task setups. When depth estimation serves as an auxiliary task, it provides spatial context that significantly improves object detection systems by refining object boundary accuracy in images. This is particularly

beneficial in complex environments where depth cues play a critical role in distinguishing between foreground and background objects.

Moreover, the relationship between MTL and other learning paradigms, such as transfer learning [85], multi-label learning [138], and multi-output regression, underscores its unique capability to facilitate efficient knowledge transfer and learning across tasks. Unlike transfer learning, which primarily focuses on enhancing a specific target task, MTL treats all tasks equally, encouraging comprehensive knowledge flow between tasks.

## 2.3 Adapting Pretrained Models for Data-Efficient Learning

In the evolving field of computer vision, significant strides have been made through the adaptation of pretrained models for data-efficient learning. Notably, the development of DINO (self-DIstillation with NO labels) and its more advanced iteration, DINOv2, showcases the potential of self-supervised learning to master rich semantic visual representations without relying on labeled data. Through a teacher-student distillation process, DINO and DINOv2 refine the architecture and training mechanisms, enhancing the model's efficiency and the accuracy of learned representations [10, 83].

Simultaneously, the Segment Anything Model (SAM) has specialized in image segmentation, generating precise object masks from simple prompts such as points and boxes. This model's proficiency in handling diverse segmentation tasks and its zero-shot transfer capabilities across various datasets highlight its adaptability and the practical efficacy of pretrained models in real-world applications [18, 22, 52].

Moreover, advancements in visual foundation models like Vision Transformers (ViTs) and Stable Diffusion (SD) further underscore the versatility of pretrained features. ViTs and Stable Diffusion have been pivotal in enhancing semantic matching, often surpassing specialized methods previously dominant in the field. Stable Diffusion, in particular, trained on billions of text-image pairs, supports the generation of images from textual descriptions, facilitating the development of robust text-to-image applications. This model also illustrates the emergence of unsupervised learning capabilities,

such as keypoint propagation, which are valuable for tasks requiring detailed and precise image understanding without extensive labeled datasets [4, 39, 99, 109].

Collectively, these foundational models redefine the capabilities of computer vision systems, providing a flexible framework that allows for tackling new tasks and adapting to unfamiliar data distributions without bespoke training. The continued integration of such models promises to advance the field, reducing the reliance on extensive data annotation and training while broadening the scope of computer vision applications across industries.

# Chapter 3

# Revisiting Image Reconstruction for Semi-supervised Semantic Segmentation

Semantic segmentation has traditionally been a labor-intensive task, demanding precise pixel-wise annotations, which can be both time-consuming and costly, particularly in large-scale applications such as medical imaging, remote sensing, and autonomous driving. Reducing annotation overhead while maintaining model performance is a critical challenge in this field. Semi-supervised learning has emerged as a powerful solution by leveraging the synergy between a small set of labeled samples and a vast pool of unlabeled data, enabling models to learn meaningful representations and capture complex contextual relationships with minimal supervision. To address the first question, we revisit the use of classical autoencoding techniques in semi-supervised semantic segmentation.

# Statement of Authorship

| | |
|---|---|
| Title of Paper | Revisiting Image Reconstruction for Semi-supervised Semantic Segmentation |
| Publication Status | ☑ Published ☐ Accepted for Publication<br>☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style |
| Publication Details | Lin, Yuhao, et al. "Revisiting Image Reconstruction for Semi-supervised Semantic Segmentation." *2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2023. |

## Principal Author

| | |
|---|---|
| Name of Principal Author (Candidate) | Yuhao Lin |
| Contribution to the Paper | Proposed the initial idea, developed the research concept, led data collection and analysis, and drafted and revised the manuscript. |
| Overall percentage (%) | 70% |
| Certification: | This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this pap |
| Signature | Date: 27 Dec 2024 |

## Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

| | |
|---|---|
| Name of Co-Author | Haiming Xu |
| Contribution to the Paper | Assisted in conducting the experiment and revised the manuscript. |
| Signature | Date: 12.27.2024 |

| | |
|---|---|
| Name of Co-Author | Lingqiao Liu |
| Contribution to the Paper | Oversaw the project development and revised the manuscript. |
| Signature | Date: 6/1/2025 |

Please cut and paste additional co-author panels here as required.

| | | | |
|---|---|---|---|
| Name of Co-Author | Jinan Zou | | |
| Contribution to the Paper | Assisted with visualizations and revised the manuscript. | | |
| Signature | | Date | 27/12/2024 |

| | | | |
|---|---|---|---|
| Name of Co-Author | Javen Qinfeng Shi | | |
| Contribution to the Paper | Oversaw the project development and revised the manuscript. | | |
| Signature | | Date | 13/01/2025 |

## 3.1 Abstract

Autoencoding, which aims to reconstruct the input images through a bottleneck latent representation, is one of the classic feature representation learning strategies. While it has proven effective as an auxiliary task for semi-supervised learning, its popularity has waned with the advent of more sophisticated methods in recent years. In this chapter, we revisit the idea of using image reconstruction as the auxiliary task and incorporate it with a modern semi-supervised semantic segmentation framework. Surprisingly, we discover that such an old idea in semi-supervised learning can produce results competitive with state-of-the-art semantic segmentation algorithms. By visualizing the intermediate layer activations of the image reconstruction module, we show that the feature map channels exhibit a strong correlation with semantic concepts. This observation explains why joint training with the reconstruction task proves beneficial for the segmentation task. Motivated by our observation, we further proposed a modification to the image reconstruction task, aiming to further disentangle the object clue from the background patterns. From experiment evaluation on various datasets, we show that using reconstruction as auxiliary loss can lead to consistent improvements in various datasets and methods. The proposed method can further lead to significant improvement in object-centric segmentation tasks.

## 3.2 Introduction

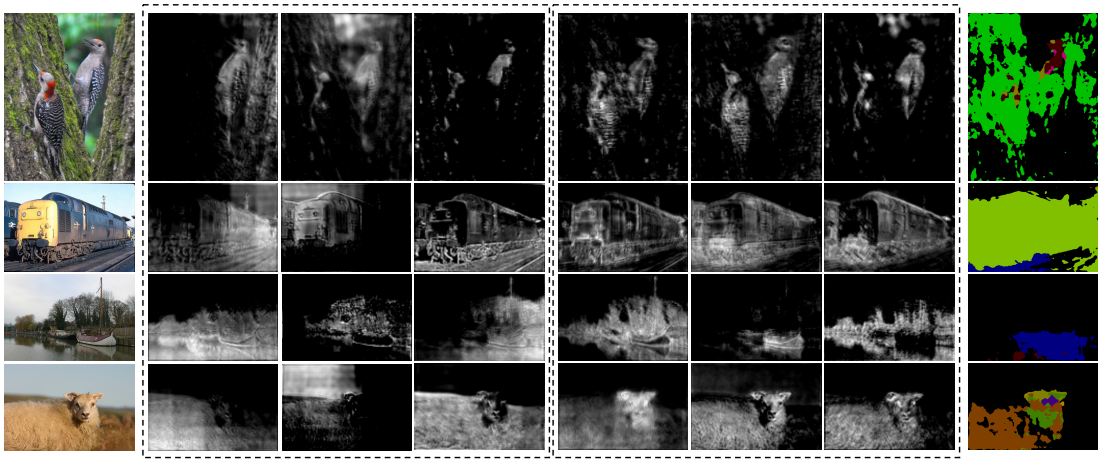
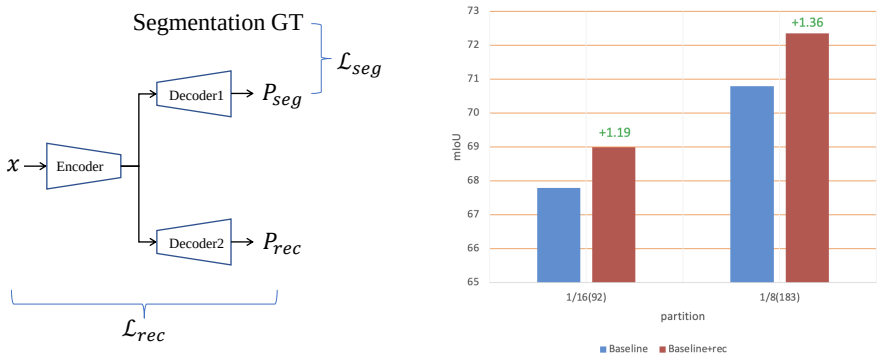



Figure 3.1: Some visualization results from Pascal VOC 2012 **validation set** (the reconstruction branch is only trained on the training set). (a) input images, (b) feature maps from the reconstruction-only model, (c) feature maps from the reconstruction-segmentation model, and (d) the segmentation results of the same epoch as c. We observe that the activation areas in some feature maps are focused on the objects (i.e., birds in the first row and goats in the fourth row).

Autoencoding aims to reconstruct inputs as outputs with the least possible amount of distortion [7] through an information bottleneck created with low dimension or low-resolution latent variables. Because of its simplicity and effectiveness, it has attracted researchers' attention since it was first introduced in the 1980s [95]. Autoencoders again enter the visions of the researchers when the deep stacked-autoencoder architectures [41] have shown state-of-the-art results as a feature extractor. Nowadays, as one of the most classic representation learning strategies, autoencoding has been widely applied in different downstream tasks in computer vision field, such as clustering [34], and classification [74]. It has also been discovered [31, 56] that an auto-encoder style reconstruction task could be an excellent auxiliary task for semi-supervised learning. In semi-supervised learning settings, we have access to a large volume of unlabelled data and only a small number of labeled training samples. The reconstruction objective, however, can be applied without any class labels. Such a scheme has become less

popular as more sophisticated methods [19, 43, 70] have been proposed recently.

Semi-supervised semantic segmentation is a challenging yet important topic in computer vision, with many real-world applications. It requires utilizing both labeled and unlabeled data to improve segmentation results. As an important application of semi-supervised learning, many semi-supervised learning methods [45, 107] have been applied and extended to solve the semi-supervised segmentation problem [43, 119, 136]. However separating the fuzzy margin between foreground and background is still challenging.




(a) (b)

Figure 3.2: (a) illustrates a simple diagram of our joint training process. The model is based on the traditional teacher-student structure, and we add another decoder with the same architecture as the segmentation decoder to reconstruct the original image. When training, the optimization goal is to minimize the total loss from both segmentation CE loss and reconstruction MSE loss. In (b), we compare the performance of baseline with and without original image reconstruction on Pascal VOC 2012 [26] under the standard partition protocols. It is worth noting that even by simply reconstructing the input image, the baseline outperforms current SOTA results in all partitions.

This work explores the use of an autoencoder-style reconstruction task to improve semi-supervised segmentation. Perhaps surprisingly, we find that if we incorporate the image reconstruction task with a commonly used semi-supervised segmentation baseline method [57, 136], the final performance can be improved, especially when the number of training images is small (see Figure 3.2). This motivates us to understand how reconstruction helps the segmentation task. By visualizing the intermediate activation

maps of the reconstruction branch (see Figure 3.1), we find that the latent activations of a reconstruction branch have already uncovered the semantics of objects if the reconstruction branch is jointly trained with a semi-supervised segmentation loss. This explains the benefit of the reconstruction task for semi-supervised segmentation, as both tasks share some similarities.

From further observation of the latent activations, we noticed that the object and part of its background could often co-occur in one feature map, suggesting potential entanglement of the object and background clue. Thus, we propose a strategy to further disentangle those two clues and expect to align the reconstruction task and segmentation task better. Specifically, we propose to reconstruct foreground-region-only images for the labeled image and perform a similar reconstruction task but apply the loss to partial pixels, as guided by the pseudo-label generated from the segmentation head. Through our experimental evaluation on various datasets, we show that joint reconstruction can be used as a strong semi-supervised segmentation baseline that achieves consistent improvement under different scenarios. In summary, the main contributions of this chapter are highlighted as follows:

- We revisit image construction as an auxiliary task for semi-supervised segmentation and show that it can be very effective when working together with existing semi-supervised segmentation methods.

- We visualize the intermediate activations of the reconstruction decoder and shed light on why it is beneficial for the segmentation task.

- We further propose a method that modifies the reconstruction task and makes the reconstruction tasks more suited for the object-centric segmentation problems.

## 3.3 Related Work

### 3.3.1 Autoencoders for Semi-Supervised Learning

As one classic unsupervised representation learning approach, Autoencoders (AE) are widely used for unsupervised learning and as a regularization scheme in semi-supervised learning [7, 48]. Because of the simplicity, they have been attracting researchers' attention since it was first introduced in the 1980s [95]. In semi-supervised learning settings [111, 126], they have a considerable volume of unlabelled data but only a small number of labeled training samples. In this setting, some study [90] shows that skip connections and layer-wise unsupervised targets effectively turn autoencoders into hierarchical latent variable models, which are well suited for semi-supervised learning. Because of its nature of clustering, it has been shown effective as an auxiliary task for many semi-supervised learning tasks such as regression [34], and classification [74]. Based on the success of AE, denoising autoencoders (DAE) [115] and variational autoencoders (VAE) [51] are proposed for better representation learning. However, it has become less popular as more sophisticated methods have been proposed in recent years.

### 3.3.2 Supervised Semantic Segmentation

As the fundamental task in computer vision, semantic segmentation has witnessed an explosion of progress in architecture design during recent decades. Starting from the FCN [71], which modifies the end-to-end architecture to be fully convolutional layers. After that, some extensions were explored: 1) encoder-decoder structure [6, 16, 94], 2) multi-scale aspects of the image [15, 66], 3) pyramidal feature maps [141], 4) dilated convolutions [13, 16, 134]. Recently, attention mechanisms [15, 30, 44] are popular among researchers because of their strength in global context communication. However, these fully supervised segmentation networks are data-hungry, laborious, and time-consuming.

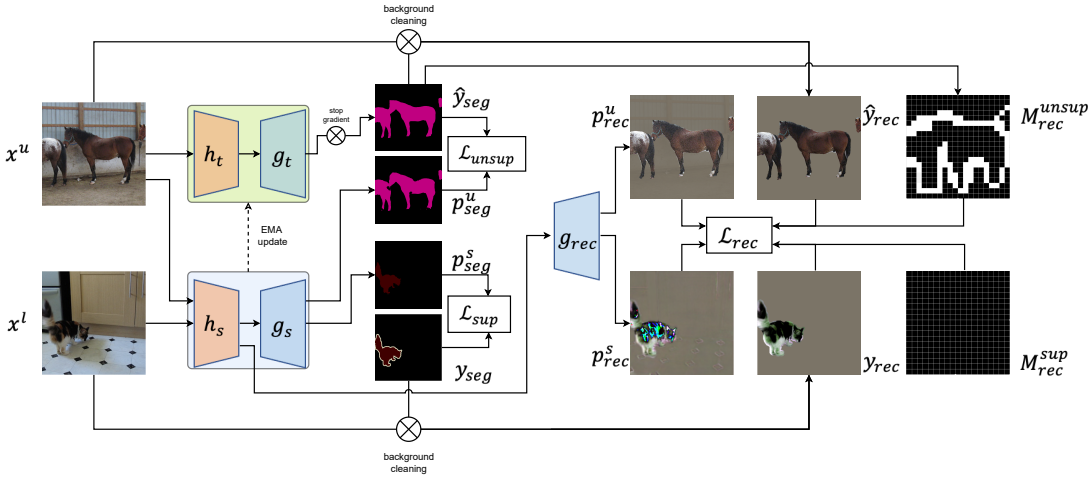



Figure 3.3: Illustration of our approach. The left side is the basic MT structure which employs two segmentation networks, the student $h_s, g_s$ and the teacher $h_t, g_t$. The right side $g_{rec}$ is our Foreground-Only reconstruction module, which takes the output of the student's encoder $h_s$ and reconstructs a monotonous background image. Reconstruction loss masks are applied for loss calculation. Specifically, for labeled images, we set the background to 0 according to the ground truth, and the one-hot mask $M_{rec}^{sup}$ in shape $[H \times W \times 1]$ is all ones. For unlabeled images, the mask $M_{rec}^{unsup}$ is derived from the teacher's ignoring uncertain areas. Then the teacher is updated by the student's exponential moving average (EMA).

### 3.3.3 Semi-supervised Semantic Segmentation

Semi-supervised semantic segmentation aims to utilize the tremendous unlabeled and small amount of labeled data fully. Though the emergence of attention mechanisms for the model attracts much attention in the segmentation field, SOTA semi-supervised semantic segmentation still relies on the CNN model Deeplab V3+ [14] and PSPNet [141]. One of the most existing studies used approaches is consistent learning with various perturbations [49, 70, 84, 136]. For example, cutmix-seg [136] validates the effectiveness of image-level perturbation with cutmix data augmentation. Cross-consistency

training (CCT) [84] introduces a feature-level perturbation and constrains the outputs of different decoders. Similarly, guided collaborative training (GCT) [49] proposed a model-level perturbation with different network initialization and enforced the consistency between models. Most recently, PS-MT [70] introduced a new adversarial perturbation for double teachers for better prediction accuracy.

Another frequently used technique is self-training [19, 130, 135], which generates pseudo labels with unlabeled data and trains the model with labeled and pseudo-label data. Furthermore, considering the class imbalance and unreliable pseudo labels, recent studies [33, 43, 119] were proposed and achieved state-of-the-art performance.

## 3.4 Image Reconstruction

Autoencoder [41] is one of the oldest unsupervised/self-supervised learning approaches. It is usually implemented by an encoder-decoder pair. The encoder encodes the input image information into latent variables that are often low-dimensional or low-resolution. Then the decoder decodes the latent variables into an image of the same size as the input. A loss function is used to ensure the reconstructed output image is as close as possible to the input image, that is,

$$\min_{\theta_e,\theta_d} \mathbb{E}\Big(\|x - f_{\theta_d}(f_{\theta_e}(x))\|_2^2\Big), \tag{3.1}$$

where $\mathbb{E}$ is a dissimilarity or distortion function and $f_{\theta_d}$, $f_{\theta_e}$ are the decoder and encoder, respectively.

Modern image segmentation neural networks, such as the DeepLab family [13, 28, 86] can also be seen as an encoder-decoder structure, although the decoder is usually lightweight compared to the encoder. The encoder encodes the images into a $H' \times W' \times d$ dimensional feature map, and the decoder, e.g., ASPP module decodes the feature maps into the predicted segmentation mask. Therefore, the image reconstruction task can be naturally applied to the existing image segmentation neural networks by sharing the same encoder but using a different decoder (branch) to produce reconstructed images.

## 3.5 Image Reconstruction for Semi-supervised Semantic Segmentation

### 3.5.1 Preliminary

Semi-supervised semantic segmentation task is defined as: given labeled images $X_l \in \mathbb{R}^{H\times W\times 3}$ , the corresponding pixel-wise semantic labels $y \in (1, C)^{H\times W}$, and unlabelled images $X_u \in \mathbb{R}^{H\times W\times 3}$ (W, H, C denotes the width, height, and the number of classes respectively). The goal is to learn a model $F$ from both label data $D^l = \{X_l, y\}$ and unlabelled data $D^u = \{X_u\}$. In most work [49, 70, 84, 119, 136], the overall optimization target is formalized as

$$\mathcal{L} = \mathcal{L}_s + \lambda \mathcal{L}_{ul},$$

where $\mathcal{L}_s$ and $\mathcal{L}_{ul}$ are loss functions for labeled and unlabeled images respectively.

### 3.5.2 Semi-supervised Segmentation Baseline

Most of the state-of-the-art semi-supervised semantic segmentation methods [19, 130, 135] are built around a simple baseline, as we call robust pseudo-labeling. Specifically, the model is firstly trained on the small number of labeled images and then produces the posterior probability estimation for each pixel on unlabeled images. Pseudo-labels are then generated if the highest posterior probability exceeds a predefined threshold. In robust pseudo-labeling, certain data augmentation, e.g., cutout [23], cutmix [136], is applied to the input image, and the pseudo-labels will be used to update the model with the augmented input images.

Our semi-supervised segmentation baseline is based on a particular version of the robust pseudo-labeling approach. Specifically, our baseline follows the typical student-teacher framework in semi-supervised semantic segmentation [49, 70, 84, 119], with the teaching network parameters being the exponential moving average [110] of the parameters of the student network.

Each network consists of a convolutional feature encoder $h$ and a segmentation decoder $g$. We denote the student version and teacher version of the encoder and decoder as

$h_s$,$g_s$, $h_t$ $g_t$, respectively. At each training step, we equally sample $b$ labeled images $\mathcal{B}_l$ and $b$ unlabeled images $\mathcal{B}_{ul}$. For $\mathcal{B}_l$ and $\mathcal{B}_{ul}$, we apply strong augmentations [19, 49] (e.g., color jitter, randomize grayscale, blur, CutMix [136] and zoom in/out [15, 65]) to the student model. The teacher model will generate posterior distribution $P(y_{i,j} = c|I^n_{i,j})$, indicating the likelihood of each pixel $(i, j)$ being assigned to class $c$. A pseudo-label for pixel $(i, j)$ is generated if $\max_c P(y_{i,j} = c|I^n_{i,j}) > \tau$. Then the pseudo-labels will be used to train a student network with augmented input images.

### 3.5.3 Reconstruction as an Auxiliary Task

Based on the framework mentioned in the Section 3.5.2, we further incorporate image reconstruction as an auxiliary task, which is shown in Figure 2(a). Different from the baseline, our framework has two decoders ($g_s$,$g_{rec}$) in the student network, and the two decoders share the encoder part ($h_s$).

Therefore, the outputs of the model are segmentation prediction: $P_{seg} = g_s \circ h_s(x) \in \mathbb{R}^{H\times W\times C}$ and the image reconstruction pixel value prediction $I_{rec} = g_{rec} \circ h_s(x) \in \mathbb{R}^{H\times W\times 3}$. Same as [17], we adopt the Mean Squared Error (MSE) loss for image reconstruction, and the image reconstruction module does not need additional annotations. The overall loss is defined as

$$\mathcal{L} = \mathcal{L}_s + \lambda_1\mathcal{L}_{ul} + \lambda_2\mathcal{L}_{rec}.$$

Surprisingly, this embarrassingly simple baseline achieves quite good performance. Figure 3.2 shows the performance before and after adding the reconstruction task. As seen, the benefit of using a reconstruction task is especially pronounced when the number of labeled images is small.

To understand the improvement, we perform visualization analysis on the reconstruction decoder. In particular, we consider the latent activations (feature maps) before the last layer of the reconstruction decoder. Recall that this last layer is a (kernel size $= 1 \times 1$) convolutional layer, which maps a $\mathbf{Z} \in \mathbf{R}^{H\times W\times d}$ feature map into the reconstructed image $I_{rec} \in \mathbb{R}^{H\times W\times 3}$. There are three convolutional filters $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3 \in \mathbb{R}^d$,

one for each color channel. Now consider one output channel from $I_{rec}$, denoted as $I^c_{rec}$, it is clear that it can be written as

$$I^c_{rec} = \sum_{k=1}^{d} \mathbf{Z}^k w^k_c, \tag{3.2}$$

where $w^k_c$ denotes the $k$-th dimension of $\mathbf{w}_c$ and $\mathbf{Z}^k \in \mathbb{R}^{H \times W}$ denotes the $k$-th slice of $\mathbf{Z}$. Intuitively, the above equation suggests that the reconstructed image is the weighted average of $d$ slices $\mathbf{Z}^k$, where each $\mathbf{Z}^k$ is equivalent to an image, and we call it "latent image". An interesting discovery is that those "latent images" could correspond to the semantic concepts in images if the encoder is jointly trained with a segmentation model. Figure 3.1 shows some example "latent images" from an autoencoder trained with reconstruction loss only and trained jointly with semi-supervised segmentation loss. Also, to show the relative progress of the segmentation decoder and the reconstruction decoder, we choose an epoch that the training process has not converged yet. From Figure 3.1, we can make the following observations:

- If the autoencoder is trained without the semi-supervised segmentation loss, the "Latent Images" have a weak correlation to the semantic concepts.
- When the autoencoder is trained with the semi-supervised segmentation loss, some "Latent Images" can correspond to some semantic concepts. For example, in the second and fourth row of Figure 3.1, the activation areas in some feature maps are focused on the train and the sheep. It seems that the semantic segmentation loss provides an inductive bias to make the image reconstructed through semantically meaningful "latent images".
- Surprisingly, some "Latent Images" recover the object contour better than the segmentation decoder at the same training epoch. It seems that the "Latent Images" are leading the segmentation decoder, which might explain why the reconstruction task could help segmentation.
- Finally, we find the "Latent Images" are far from perfect. Some background pixels, especially those that are the context of the object, tend to co-occur with the object in the latent image.

### 3.5.4 Improving the Reconstruction Task by Object-Background Disentanglement

The last observation discussed in Section 3.5.3 suggests a potential object-background entanglement may exist in the current reconstruction task. Thus, we propose the following strategy to disentangle the object and its context background. More specifically, we let the reconstruction decoder only reconstruct the foreground-only images at the labeled set. In other words, the output from the reconstruction decoder only contains pixels belonging to the object parts while the background pixels are set to zero, that is,

$$y_{rec\ (i,j)} = \begin{cases} x^l_{(i,j)}, & \text{if } y_{(i,j)} \in \text{foreground} \\ 0, & \text{otherwise} \end{cases} \tag{3.3}$$

Examples of foreground-only images are shown in Figure 3.3. For unlabeled images, we do not have access to the class labels, those we cannot directly generate foreground-only images. Thus we recourse to pseudo-labels. For an unlabeled image, we consider the following three scenarios for a pixel $(i, j)$: (1) the current pixel can generate a pseudo-label, and the pseudo-label corresponds to the foreground. In other words, $max_{c' \in \mathcal{O}} P(y_{i,j} = c' | x_{i,j}) > \tau$, where $\mathcal{O}$ is the set of classes that belong to objects. (2) the current pixel can generate a pseudo-label, and the pseudo-label corresponds to the background. (3) no pseudo-label can be generated from the current pixel and the segmentation decoder is uncertain about the class of the pixel. We will ignore the loss of pixels from (3) and only perform the foreground-only reconstruction for pixels from (1) and (2). We call this modified reconstruction method as Foreground-Only reconstruction (FOrec). The scheme is illustrated in Figure 3.3.

**Discussion:** The FOrec method is mainly for object centric semantic segmentation, where the aim is to segment different objects from the background. For generic scene segmentation, i.e., segmenting both things and stuff, one can choose a category that often co-occurs with other categories as the background. Applying FOrec in that case could potentially alleviate the entanglement of those semantic concepts.

## 3.6 Experiments

In this section, we compare our approach with several semi-supervised semantic segmentation methods.

### 3.6.1 Experimental Setup

**Datasets:** Our experiments are mainly conducted on the Pascal VOC 2012 [26], and Cityscapes [20], which are widely used in semi-supervised semantic segmentation tasks [49, 70, 84, 136]. The **classic** Pascal VOC 2012 consists of 1,464/1,449/1,556 images covering twenty classes for training, validation, and testing, respectively. Due to the demand for data for the semi-supervised semantic segmentation scene, some researchers [49, 70, 84] adapt the additional labels from [36], which means the training data is augmented up to 10,528 images. In the augmented training set, 1,464 labeled data is selected among 1,464 samples in the **classic** setting, while the remainings are of low quality, containing noise. The augmented data selection setting is named **blender**. Both settings are evaluated for the measurement of performance. Cityscapes [20] is the urban driving scene dataset, consisting of 2,975, 500, and 1,525 images covering 19 classes for training, validation, and testing, respectively.

In this chapter, we follow the same data splitting protocol from $U^2$PL [119] and experiment with four kinds of label partition: 1/16, 1/8, 1/4, and 1/2. Our code will be released after the anonymity period.

**Evaluation metrics:** Following the previous works [70, 119], we adopt the mean Intersection-over-Union (mIoU) as the evaluation metrics.

**Implement detail:** Following the prior work [19, 70, 125], the network structure of our method is based on Deeplab V3+ [14], with pretrained ResNet-101 as the backbone. The segmentation head and the auxiliary task head are the default pixel-level linear classifier.

For all experiments on both datasets, we employ the stochastic gradient descent (SGD) as the optimizer and polynomial learning rate decay: $(1 - \frac{iter}{total_iter})^{0.8}$ for model opti-

mization. For reconstruction loss, we set the $\lambda_1$ as 0.5 and $\lambda_2$ as 1 for the unsupervised and supervised parts, respectively.

On Pascal VOC 2012 [26], the images are cropped into $512 \times 512$ pixels and trained with initial learning rate $1.0 \times 10^{-3}$ , weight decay $1.0 \times 10^{-4}$, and 80 training epochs. On Cityscapes [20] , we crop the images into $712 \times 712$ pixels and trained our model with an initial learning rate $1.0 \times 10^{-2}$ , weight decay $5.0 \times 10^{-4}$ and 200 training epochs. Our experiments were run with batch size 16 on 8 NVIDIA Tesla V100 GPUs.

For the reconstruction decoder, we simply adapt the same structure as the segmentation decoder. The only difference between them is the channel number of the last layer.

### 3.6.2 Comparison with State-of-the-Arts

**Pascal VOC 2012**

| Method | 92 | 183 | 366 | 732 | 1464 |
|---|---|---|---|---|---|
| Supervised | 45.8 | 54.9 | 65.9 | 71.7 | 72.5 |
| MT [110] | 51.7 | 58.9 | 63.9 | 69.5 | 71.0 |
| PseudoSeg [143] | 57.6 | 65.5 | 69.1 | 72.4 | 73.2 |
| CPS [19]$_{\text{[CVPR 21']}}$ | 64.1 | 67.4 | 71.7 | 75.9 | - |
| PS-MT [70]$_{\text{[CVPR 22']}}$ | 65.8 | 69.6 | 76.6 | 78.4 | 80.0 |
| ST++ [130]$_{\text{[CVPR 22']}}$ | 65.2 | 71.0 | 74.6 | 77.3 | 79.1 |
| U$^2$PL [119]$_{\text{[CVPR 22']}}$ | 68.0 | 69.1 | 73.6 | 76.2 | 79.5 |
| FOrec (Ours) | **71.0** | **74.7** | **77.5** | **78.7** | **81.1** |

Table 3.1: Comparing results of state-of-the-art algorithms on PASCAL VOC 2012 [26] val set with mIoU (%) metric. Methods are trained on the classic setting, i.e., the labeled images are selected from the original VOC train set, which consists of 1, 464 samples in total. Best results are in bold.

| Method | 662 | 1323 | 2646 | 5290 |
|---|---|---|---|---|
| Supervised | 67.87 | 71.55 | 75.80 | 77.13 |
| MT [110] | 70.51 | 71.53 | 73.02 | 76.58 |
| CPS [19][CVPR 21'] | 74.48 | 76.44 | 77.68 | 78.64 |
| AEL [43] [NIPS 21'] | 77.20 | 77.57 | 78.06 | 80.29 |
| ST++ [130][CVPR 22'] | 74.70 | 77.90 | 77.90 | - |
| PS-MT [70][CVPR 22'] | 75.50 | 78.20 | 78.72 | 79.76 |
| UCC [27][CVPR 22'] | 76.49 | 77.06 | 79.09 | 79.54 |
| $U^2$PL [119] [CVPR 22'] | 77.21 | 79.01 | 79.30 | 80.50 |
| FOrec (Ours) | **78.84** | **80.52** | **80.92** | **80.99** |

Table 3.2: Comparing results of state-of-the-art algorithms on PASCAL VOC 2012 [26] val set with mIoU (%) metric. Methods are trained on the blender setting, i.e., the labeled images are selected from the augmented VOC train set, which consists of 10, 582 samples in total. Best results are in bold.

Table 3.1 and Table 3.2 illustrated the results on Pascal VOC 2012 validation set, Table 3.1 is under **classic** setting and Table 3.2 is under **blender** setting.

For **classic** setting, Table 3.1 illustrates that our approach successfully exploits unlabelled data, with a dramatic performance boost from the fully supervised training. Specifically, in the smaller partition like 92 and 183, our approach surpasses the fully supervised baseline with 25.2% and 19.8%, respectively. Meanwhile, compared with the current SOTA methods, our approach performs consistently better than all other methods for all partition protocols (using the ResNet-101 as the backbone). Taking the $U^2$PL [119] as the instance, our approach improves the performance by 1.6% to 5.6% in all cases. Furthermore, in some partitions, our approach is better than all the current SOTA methods with fewer labeled samples. For example, compared with all current SOTA methods, our approach trained with 92 labeled images outperforms trained with 183 labeled images. This demonstrates that when the number of labeled

data is extremely small (92, 183), our approach achieves a significant improvement in performance.

For **blender** setting, Table 3.2 indicates that our approach outperforms the supervised baseline by a large gap 3.86% to 10.97% from the fully supervised training. Compared with the current SOTA methods, our approach beats all other methods for all partition protocols. Even comparing with one of the currently best performed approaches U$^2$PL[119], our approach improves the performance by 1.63%, 1.51%, 1.62% in 662, 1323, and 2646, respectively. The impressive boost proves that our method is not only useful for accurately annotated data, but also compatible with the noisy annotation.

**Cityscapes**

| Method | 1/16 | 1/8 | 1/4 | 1/2 |
|---|---|---|---|---|
| Supervised | 65.74 | 72.53 | 74.43 | 77.83 |
| MT [110] | 69.03 | 72.06 | 74.20 | 78.15 |
| CCT [84]$_{\text{[CVPR 20']}}$ | 69.32 | 74.12 | 75.99 | 78.10 |
| CPS [19]$_{\text{[CVPR 21']}}$ | 69.78 | 74.31 | 74.58 | 76.81 |
| AEL [43]$_{\text{[NIPS 21']}}$ | **74.45** | 75.55 | 77.48 | 79.01 |
| U$^2$PL [119] $_{\text{[CVPR 22']}}$ | 70.30 | 74.37 | 76.47 | 79.05 |
| FOrec (Ours) | 72.42 | **75.76** | **77.65** | **79.18** |

Table 3.3: Comparing results of state-of-the-art algorithms on Cityscapes [20] val set with mIoU (%) ↑ metric. Methods are trained on identical label partitions, and the labeled images are selected from the Cityscapes train set, which consists of 2, 975 samples in total. Best results are in bold.

Table 3.3 demonstrates the results of our method against several current state-of-the-art algorithms on Cityscapes validation set. Compared to the fully supervised results, our method successfully exploits unlabelled data, with an obvious performance boost for all partitions. e.g., under the 1/16 label partition, our approach surpasses the fully supervised result by 6.68%. Then, compared to the state-of-the-art algorithm

$U^2$PL[119], ours performs better than $U^2$PL in all cases by 2.12%, 1.39% and 1.18% under the 1/16, 1/8 and 1/4 label partition, respectively.

Note that the performance of our method on the 1/16 label partition is slightly lower than that of AEL [43]. The reason is that the class imbalance problem of this partition is more serious and AEL especially aims to deal with class imbalance problems. However, our methods focus on separating the foreground objects from the background patterns in semi-supervised semantic segmentation tasks, we do not explicitly consider processing label imbalance problems. Technically, there is a high probability that merging both ideas is a to optimize overall performance.

### 3.6.3 Analysis

| | Standard Reconstruction | FOrec | 1/16 (92) | 1/8 (183) |
|---|---|---|---|---|
| ① | ✗ | ✗ | 67.82 | 70.78 |
| ② | ✓ | ✗ | 68.99 | 72.35 |
| ③ | ✗ | ✓ | **70.99** | **74.67** |

Table 3.4: Ablation study on the effectiveness of the standard reconstruction task and the proposed foreground-only reconstruction scheme (FOrec) on pascal VOC 2012. ✓and ✗ represent the variant containing or not containing the sub module at each row respectively.

| | std. Recon. | FOrec | 1/8 (372) | 1/4 (744) |
|---|---|---|---|---|
| ① | ✗ | ✗ | 72.23 | 76.19 |
| ② | ✓ | ✗ | 74.14 | 76.56 |
| ③ | ✗ | ✓ | **75.76** | **77.65** |

Table 3.5: Ablation study on the effectiveness of the standard reconstruction task and the proposed foreground-only reconstruction scheme (FOrec) on CityScapes.

In the following part, we perform a series of experiments to analyze the proposed method. Specially, we consider the following analyses: (1) The comparison between FOrec and standard reconstruction on both the object-centric segmentation task, i.e., PASCAL VOC, and scene-understanding segmentation task, i.e., CityScapes. (2)The comparison between the foreground-background segmenta- tion and FOrec. (3) The latent images created from FOrec. (4) The applicability of the proposed method to other semi-supervised segmentation approaches.

**The comparison of the standard reconstruction and FOrec**

| Method | 92 | 183 | 366 | 732 |
|---|---|---|---|---|
| fore/back-ground seg. | 67.7 | 71.9 | 76.1 | 78.1 |
| FOrec (ours) | **71.0** | **74.7** | **77.5** | **78.7** |

Table 3.6: Comparison between foreground-background (saliency estimation) segmentation and foreground-only reconstruction (Forec) on PASCAL VOC 2012 (classic setting).

In Table 3.4 and 3.5, we compare the standard reconstruction task and the proposed foreground-only reconstruction scheme. As seen, FOrec achieves a significant improvement over the standard reconstruction task on PASCAL VOC 2012 clean setting and also on Cityscapes, the improvement over the standard reconstruction is around 2%. It is clear that even the standard reconstruction could provide stable benefit in segmentation task. And also FOrec could be applicable and boost the benefit.

**The comparison of the foreground-background segmentation and FOrec**

Table 3.6 shows that the naive foreground-background segmentation performs worse than our approach, the improvement benefit from the foreground-only reconstruction is ranging from 0.5% to 3.3%. We think that unifying all kinds of foregrounds into one class is not conducive to the semantic segmentation of different objects. Same experiments are conducted on another codebase called PS-MT and conclusions are exactly the same.

**The impact of FOrec on the "latent images"**

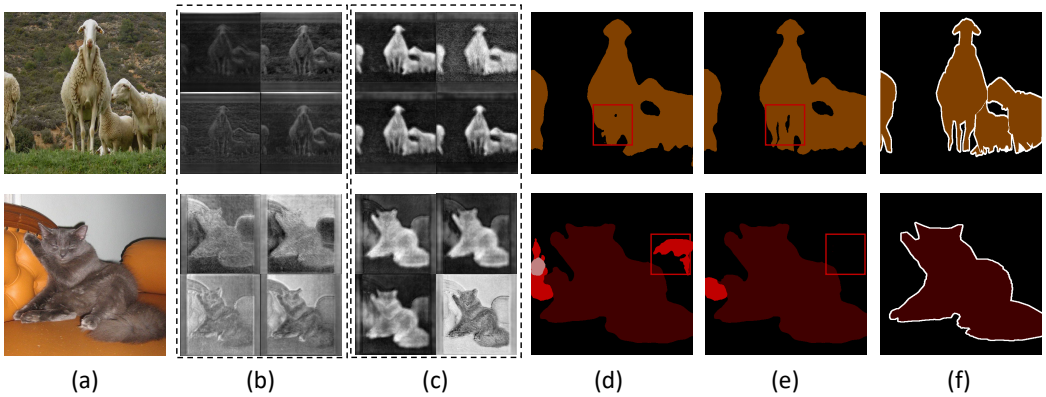



Figure 3.4: Visualizations of the latent images obtained from FOrec and standard reconstruction, from left to right, are (a) input images, (b) latent images obtained from standard reconstruction, (c) latent images obtained from FOrec, (d) segmentation results from standard reconstruction, (e) segmentation results from FOrec, and (f) segmentation ground truth. Note that the (b) and (c) are slices from the same position for a fair comparison.

The FOrec is proposed to address the issue that background pixels tend to co-occur with foreground pixels. To verify this design, we visualize the latent images obtained from FOrec and standard reconstruction. We conduct an experiment by using two models, one trained by FOrec and another trained by standard reconstruction (jointly trained with semi-supervised segmentation loss for both cases with the same architecture). Then we use the trained model to generate latent images for input images from the validation set. *Note that although FOrec is trained to reconstruct foreground-only images, the Foreground-only images are only used at the training time as the target. Once trained, it can be used without knowing the foreground mask.* The results are shown in Figure 3.4. As seen, by applying FOrec, the latent images tend to capture more object regions. We can observe more "latent images" that only include the objects. For example, in our figures, the goats and the cat in (c) are highlighted a lot compared with the latent images of standard reconstruction. When mapping to the results of the segmentation task, these natures are kept. For instance, in the first example

in Figure 3.4, the gaps between goats' legs are captured, which did not appear in the results from the standard reconstruction. Moreover, in the second example, the sofa next to the cat is faded in the FOrec feature maps. In the segmentation results, it disappears. These results in Figure 3.4 clearly validates the effectiveness of the proposed method.

**The applicability of the proposed method on other semi-supervised segmentation methods**

| Method | 92 | 183 | 366 | 732 |
|---|---|---|---|---|
| PS-MT [70] | 65.8 | 69.6 | 76.6 | 78.4 |
| PS-MT+rec | 68.4 | 71.0 | 77.2 | 78.9 |
| PS-MT+FOrec | **70.3** | **71.9** | **77.9** | **79.8** |

Table 3.7: Comparing results of another SOTA codebase (PS-MT[70]) on classic setting PASCAL VOC 2012 [26] val set with mIoU (%) metric. Experiments are conducted in the same settings as in the chapter.

Finally, we apply both the reconstruction task and FOrec to another semi-supervised learning approach, PS-MT [70]. The results are shown in Table 3.7. As seen, the reconstruction is still effective. Compared with the PS-MT baseline, applying FOrec leads to a significant increase, especially when the number of training examples is small. The advantage of FOrec over standard reconstruction is also seen. Again, we observe FOrec tends to produce a superior performance on the low-supervision regime, e.g., when only 92 labeled images are used.

## 3.7 Conclusion

In this chapter, we revisit the idea of using image reconstruction as an auxiliary task for semi-supervised semantic segmentation. We find that this old idea is able to boost the performance of current state-of-the-art semantic segmentation algorithms. By visualizing the intermediate layer activations of the image reconstruction module, we show that the feature map channel can correlate well with the semantic concept, which explains why joint training with the reconstruction task is helpful for the segmentation task. Motivated by this observation, we further proposed a modification to the original image reconstruction task, aiming to further disentangle the object clue from the background patterns. From experiment evaluation on various datasets, we show that using foreground only reconstruction as an auxiliary task leads to consistent improvements in various datasets and methods. The proposed method can further lead to significant improvement in object-centric segmentation tasks. For these datasets without background class, our proposed method can still provide slight improvements. More investigations will be made to improve scene understanding ability.

# Chapter 4

# A Simple-but-effective Baseline for Training-free Class-Agnostic Counting

With the advent of powerful pretrained models, fundamental tasks like image segmentation have been greatly improved. However, class-agnostic object counting (CAC) presents unique challenges, as it requires not only class-agnostic quantity estimation but also precise object localization—both of which traditionally demand substantial human effort and financial resources for annotation. This annotation burden is particularly costly in applications such as ecological monitoring, traffic control, and industrial inspection, where large-scale, fine-grained object counting is essential. To tackle this challenge, we explore the application of pretrained models, including SAM and the high-dimensional self-supervised learning model (DINO), demonstrating that a CAC model can achieve performance comparable to state-of-the-art training-based approaches without requiring task-specific training data.

# Statement of Authorship

| | |
|---|---|
| Title of Paper | A Simple-but-effective Baseline for Training-free Class-Agnostic Counting |
| Publication Status | ☐ Published ☑ Accepted for Publication<br>☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style |
| Publication Details | Lin, Yuhao, et al. "A Simple-but-effective Baseline for Training-free Class-Agnostic Counting." arXiv preprint arXiv:2403.01418 (2024). |

## Principal Author

| | |
|---|---|
| Name of Principal Author (Candidate) | Yuhao Lin |
| Contribution to the Paper | Proposed the initial idea, developed the research concept, led results collection and analysis, and drafted and revised the manuscript. |
| Overall percentage (%) | 70% |
| Certification: | This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper. |
| Signature | Date: 27 Dec 2024 |

## Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

| | |
|---|---|
| Name of Co-Author | Haiming Xu |
| Contribution to the Paper | Assisted in drafting the manuscript and responding to reviews; also helped organize the experiments. |
| Signature | Date: 12.27.2024 |

| | |
|---|---|
| Name of Co-Author | Lingqiao Liu |
| Contribution to the Paper | Oversaw the project development and helped organize the experiments. |
| Signature | Date: 6/1/2025 |

Please cut and paste additional co-author pa

| | | | | |
|---|---|---|---|---|
| Name of Co-Author | Javen Qinfeng Shi | | | |
| Contribution to the Paper | Oversaw the project development and revised the manuscript. | | | |
| Signature | | | Date | 13/01/2025 |

## 4.1 Abstract

Class-Agnostic Counting (CAC) seeks to accurately count objects in a given image with only a few reference examples. While previous methods achieving this relied on additional training, recent efforts have shown that it's possible to accomplish this without training by utilizing pre-existing foundation models, particularly the Segment Anything Model (SAM), for counting via instance-level segmentation. Although promising, current training-free methods still lag behind their training-based counterparts in terms of performance. In this research, we present a straightforward training-free solution that effectively bridges this performance gap, serving as a strong baseline. The primary contribution of our work lies in the discovery of four key technologies that can enhance performance. Specifically, we suggest employing a superpixel algorithm to generate more precise initial point prompts, utilizing an image encoder with richer semantic knowledge to replace the SAM encoder for representing candidate objects, and adopting a multiscale mechanism and a transductive prototype scheme to update the representation of reference examples. By combining these four technologies, our approach achieves significant improvements over existing training-free methods and delivers performance on par with training-based ones.

## 4.2 Introduction

Class-Agnostic Counting (CAC) [73] is increasingly gaining recognition in the field of computer vision. In contrast to traditional class-specific object counting [89, 101], which relies on training models for specific predefined categories, CAC adopts a more flexible approach. It expects the model to adapt to a diverse array of objects across various classes using only a minimal number of examples [73]. This adaptability makes CAC a compelling area of research, offering potential for broad application in scenarios where rapid or dynamic object recognition is required.

To attain the ability for Class-Agnostic Counting, a common approach is to utilize a training dataset containing various objects along with their corresponding density maps. This training data is used to train a model that compares visual features be-

tween reference examples and the query image [24, 73, 101]. This methodology enables class-independent adaptation, which is essential for scenarios requiring dynamic counting. However, a significant drawback of these methods is their reliance on large-scale annotated training data, which presents challenges in terms of labor intensity and scalability across diverse visual categories [76]. Recently, the Segment Anything Model (SAM) [52] has garnered substantial attention across various research domains due to its remarkable ability to generate high-quality object masks for any given image. This versatility has led to its application in numerous research areas, as evidenced by multiple studies [75, 100, 118, 129]. In the context of CAC, researchers have started adopting SAM for producing instance-level segmentation masks corresponding to referenced objects. By counting the number of instance masks, they can determine the final count of objects.

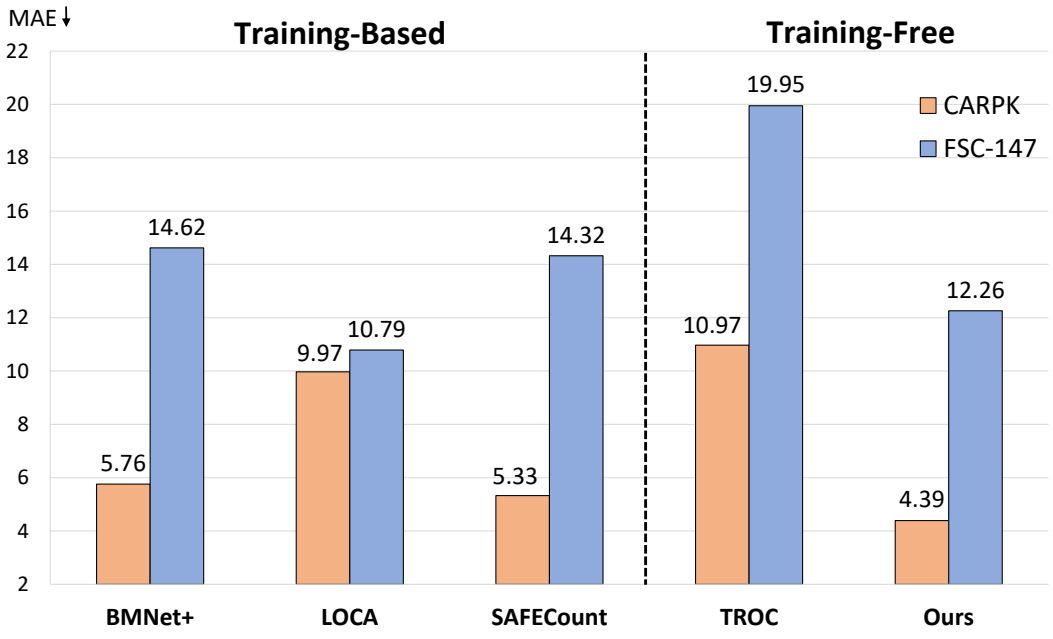



Figure 4.1: Visualizing the performance, our method notably narrows the gap in counting accuracy between training-free and training-based approaches. Impressively, it even surpasses the top-performing training-based methods on the CARPK dataset, underscoring the potential of training-free approaches.

Thanks to SAM's robust instance segmentation capabilities, this counting-via-instance-segmentation approach can function effectively even when utilizing the off-the-shelf SAM model without any additional training. Training-free methods hold appeal for CAC since they eliminate the need for a dedicated training dataset and alleviate the challenges associated with collecting extra training data. Nevertheless, despite their promise, current training-free CAC methods still lag behind their training-based counterparts in terms of performance, e.g., as demonstrated in Figure 5.1, the current top training-free method falls short of the best training-based approaches by 84.89% and 105.82% on the FSC-147 [89] and CARPK [42] datasets, respectively.

In our research, we introduce an improved approach to achieve training-free Class-Agnostic Counting using SAM. Our method provides a straightforward yet robust means to significantly narrow the performance gap often observed in training-free CAC methods. This positions our approach as a valuable benchmark for future research in this domain. Our approach to CAC relies on four pivotal technological components, each playing a crucial role: First, we emphasize the paramount importance of precise point prompt placement within SAM to achieve accurate object masking. To this end, we advocate the utilization of superpixels, which enables the creation of initial point prompts, resulting in high object recall without imposing significant computational overhead. Moreover, our thorough investigations have exposed limitations in SAM's current image encoder concerning its ability to effectively capture semantic-level similarity. In response, we propose the adoption of image encoders equipped to capture richer semantic information. This proposition holds the potential to enhance the representation of objects during the counting process. Additionally, we recognize the significance of implementing a multiscale mechanism, particularly for the precise counting of small objects. By incorporating this mechanism, our approach accommodates variations in object size across diverse instances, thus yielding more accurate counting results. Finally, we introduce a transductive prototype update strategy for reference examples. Through the augmentation of reference examples with candidate objects that are likely to share the same object category, we substantially enhance the reference-candidate similarity metric. This enhancement empowers us to more effectively determine whether an object proposal qualifies as the target of interest, ul-

timately contributing to the overall accuracy of our Class-Agnostic Counting methodology. Collectively, these four technological advancements serve as the cornerstone of our approach.

Our rigorous experiments demonstrated the remarkable performance of our proposed methods and the effectiveness of the four key components. Significantly, we show for the first time that a Class-Agnostic Counting (CAC) model without any training can achieve performance on par with that of training-based CAC approaches.

## 4.3 Related Work

### 4.3.1 Class-Agnostic Counting Methods

Lu et al. [73] first address CAC, employed convolutional neural networks to extract and concatenate features from both query images and exemplars to estimate object counts. However, this method faced challenges related to overfitting when directly regressing from concatenated features. Addressing this, CFOCNet [133] introduced improvements by explicitly modeling similarity, using a Siamese network to track to improve localization and count accuracy.

Further enhancements in similarity modeling were proposed by Shi et al. [101] and You et al. [133], who integrated advanced techniques such as self-attention and learnable similarity metrics to reduce appearance variability and guide feature fusion, respectively. Moreover, the incorporation of a vision transformer for feature extraction by Liu et al. [67] marked a significant step towards leveraging the strengths of transformer models in the field, demonstrating the use of cross-attention to merge image and exemplar features effectively. Most recently, LOCA [24] introduces an object prototype extraction module that iteratively merges exemplar shape and appearance with image features, enabling adaptation from low-shot to zero-shot object counting scenarios.

However, these methods all need numerous training data, and the exemplar annotation is labor-consuming. In specific, the most commonly used dataset in CAC task, FSC-147 [89], includes 3659 training images with more than 10,000 exemplar annotations needed.

Ma et al. [76] first address CAC with Segment Anything Model (SAM) [52] in a training-free manner. After that Shi et al. [102] proposed a novel training-free object counter that introduces prior-guided mask generation to improve accuracy, along with a two-stage text-based object counting method, yielding competitive results on standard benchmarks. However, the performance of these training-free methods heavily lags behind the training-based CAC methods. In this work, we propose a simple but effective training-free approach to significantly reduce the performance gap.

### 4.3.2 Superpixel Methods

Superpixel is an important concept in computer vision, primarily used to group pixels into perceptually meaningful regions, which can significantly reduce the complexity of image processing tasks. As a way to effectively reduce the number of image primitives for subsequent processing, superpixel algorithms have been widely adopted in vision problems such as saliency detection [127], object detection [32], and semantic segmentation [106].

Superpixel was popularized by Ren and Malik [92] through their pioneering work on the Normalized Cuts algorithm, and has been significantly advanced by the introduction of the Simple Linear Iterative Clustering (SLIC) [1] algorithm. SLIC uses K-means clustering in a five-dimensional space of color and pixel coordinates to create compact and uniform superpixels, which has made it a benchmark in the field. Several variants(LSC [63], Manifold SLIC [69], SNIC [2]) have been developed upon SLIC.

### 4.3.3 Vision Foundation Models

For pure vision area, DINO (self-DIstillation with NO labels) [10] and its more advanced iteration, DINOv2 [83], represent significant strides in self-supervised learning within computer vision. Through a teacher-student distillation process that does not rely on labeled data, DINO masters semantic-rich visual representation. DINOv2 builds on this by refining the architecture and training mechanisms, further improving the model's efficiency and the accuracy of learned representations.

Most recently, SAM [52] specializes in image segmentation, adeptly generating accurate object masks from prompts such as points and boxes. SAM's proficiency is evidenced by its remarkable performance on diverse segmentation benchmarks and its zero-shot transfer abilities, which extend across a range of datasets [18, 22].

These foundation models collectively redefine the capabilities of computer vision systems. They provide a versatile framework for tackling new tasks and adapting to unfamiliar data distributions without the necessity of bespoke training [75, 118, 129].

## 4.4 Method

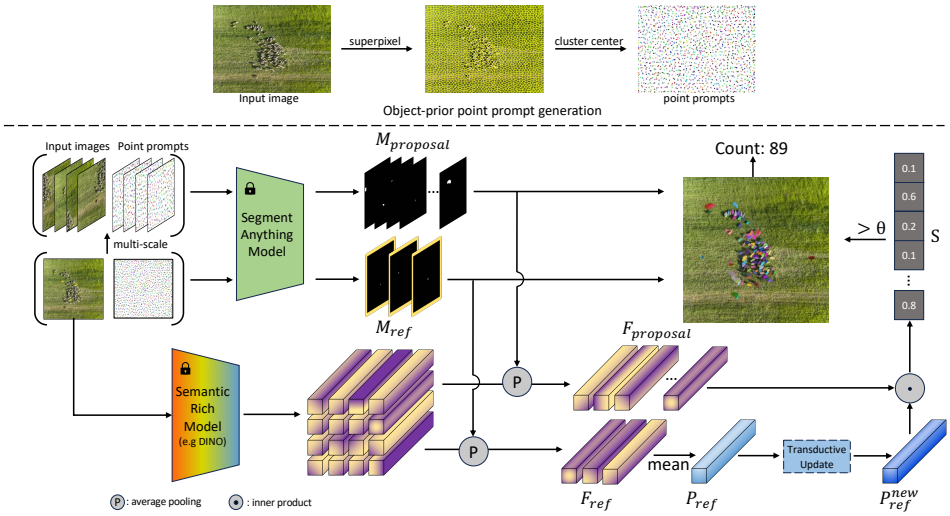



Figure 4.2: **Top** row illustrates the creation process of object-prior point prompts. **Bottom** row depicts a pipeline overview of our proposed training-free object counting approach. Details of the transductive update module are provided in Eq. 4.4. Reference objects are marked with yellow boxes in the input image (zoom in for clarity).

In this section, we introduce four innovative modules that enhance the effectiveness of employing SAM in training-free CAC, i.e., object-prior point prompt (Sec. 4.4.1) and semantic-rich feature (Sec. 4.4.2) are matters for the training-free model to deliver satisfactory counting outcomes. Additionally, the multi-scale strategy (Sec. 4.4.3) and transductive prototype updating (Sec. 4.4.4) help substantially elevate counting accuracy. The comprehensive framework of our method is depicted in Figure 4.2.

### 4.4.1 Object-prior Point Prompt Matters

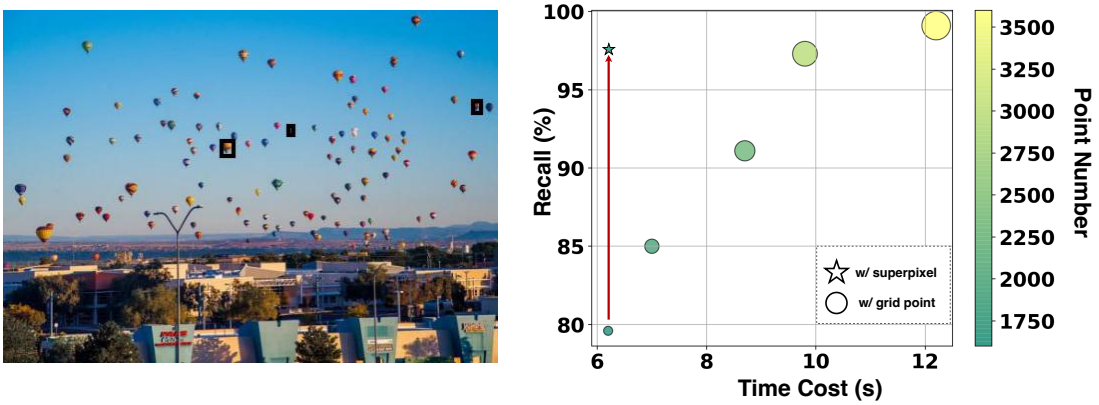



Figure 4.3: Effectiveness demonstration of the use of superpixel on the quality of mask proposals generated by SAM. **Left:** an image with reference exemplars (hot-air balloon) in black boxes. **Right:** a bubble chart illustrates the trade-off between the recall rate of the interested object and the time cost. Without increasing the number of point prompts, SAM with superpixel[1]can significantly improve the recall rate of hot-air balloons in the output mask proposals, thus avoiding the potential computational burden caused by the demand for denser point grids. Please find visualizations of the generated mask proposal in the appendix.

In the use of SAM for zero-shot object proposal generation [52], a regular grid of point prompts is utilized to automatically generate class-agnostic masks. In the scenario of object counting, counted objects are usually tiny or crowded. In order to increase the recall of the object proposals, fine-grained grid point prompts became indispensable. As presented in Figure 4.3, naively increasing the number of points grid from 1600 to 3600 increases recall by 20%. However, an increase in the number of point grids leads to a doubling of computational time overhead. In order to ensure the recall rate without increasing time complexity too much, we propose to utilize the superpixel segmentation algorithm [92] to firstly perceptual group pixels into clusters and then take cluster centers as object-prior point prompts to feed into SAM for object proposal generation.

[1]Performing superpixel on this example image takes only 10ms with AMD Ryzen 9 3900X CPU, a negligible computational expense compared to the SAM computation cost.

In this work, a superpixel algorithm named Simple Linear Iterative Clustering (SLIC) is used to generates compact superpixels with a low computational overhead [1].

Given the initial cluster number $K$, SLIC can directly output superpixel centers $\mathcal{C} = \{(x_i, y_i)\}_{i=1}^{K}$. These centers, along with the provided reference boxes (or points) $\mathcal{R}$, can then be fed into SAM batch by batch for mask proposal generation

$$\mathcal{M}_{\text{ref}} = \mathbb{D}(\mathbb{E}(I), \mathcal{R}), \quad \mathcal{M}_{\text{proposal}} = \mathbb{D}(\mathbb{E}(I), \mathcal{C}) \tag{4.1}$$

where $\mathbb{E}$ and $\mathbb{D}$ are encoder and decoder of SAM respectively. $I$ is the input image. $\mathcal{M}_{\text{ref}}$ denotes the generated masks for corresponding referenced prompts (reference boxes or points) and $\mathcal{M}_{\text{proposal}}$ means all potential mask proposals [1].

### 4.4.2 Semantic-rich Feature Matters

When object masks are ready, the next step is to recognize all object masks similar to the given reference objects from $\mathcal{M}_{\text{proposal}}$. One straightforward way is to compare the similarity of the region-of-mask features (RoMF) of referenced objects to that of the RoMF of others

$$\begin{aligned} \mathcal{S}^i &= \text{sim}(\mathsf{P}_{\text{ref}},\ F^i_{\text{proposal}}), \quad \text{where} \\ \mathsf{P}_{\text{ref}} &= \frac{1}{n_{\text{ref}}} \sum_{j=1}^{n_{\text{ref}}} \text{AvgPool}\left(\mathbb{E}(I),\ \mathcal{M}^j_{\text{ref}}\right) \\ F^i_{\text{proposal}} &= \text{AvgPool}\left(\mathbb{E}(I),\ \mathcal{M}^i_{\text{proposal}}\right) \end{aligned} \tag{4.2}$$

where AvgPool is the average pooling operation. $n_{\text{ref}}$ is the number of given reference objects and is set to 3 by default in the literature [89]. $\mathsf{P}_{\text{ref}}$ denotes the average of referenced object features, a.k.a prototypes. $\text{sim}(\cdot, \cdot)$ is the similarity function and the cosine function is used in this study. $\mathcal{S}^i$ is the similarity score for each object proposal $i$.

[1] The largest mask generated by SAM, which typically corresponds to the background, and masks for referenced objects are filtered from $\mathcal{M}_{\text{proposal}}$ for simplicity, i.e., $\mathcal{M}_{\text{proposal}} \cap \mathcal{M}_{\text{ref}} = \emptyset$.

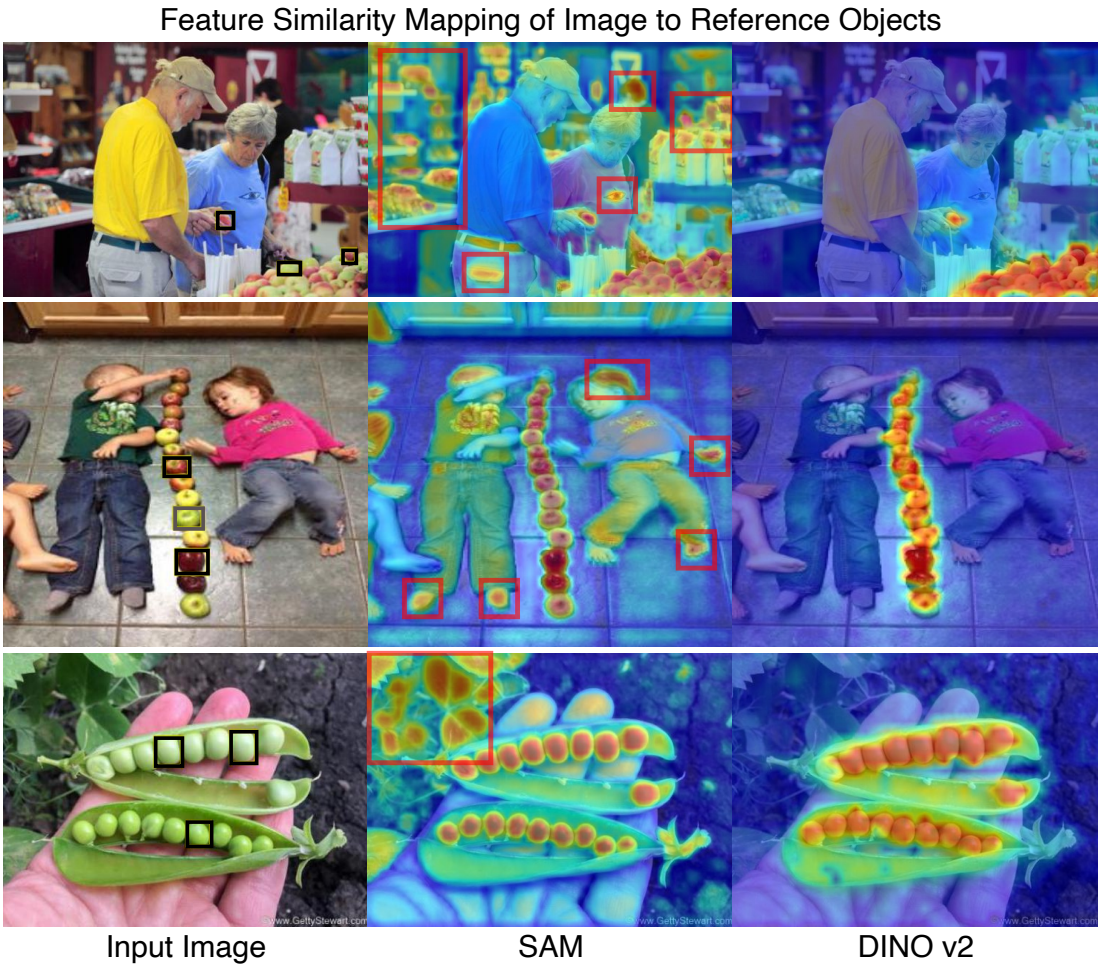



Figure 4.4: Visualization of the similarity mappings of image features to reference object features (marked in black boxes) using SAM and DINOv2. It distinctly shows that SAM's similarity mapping erroneously highlights numerous areas unrelated to the target object, whereas DINOv2's mapping accurately encompasses the objects of interest. This demonstrates that DINOv2's features possess more semantically relevant knowledge.

Next, an overall segmentation for interested objects can be obtained by selecting mask proposals whose similarity score is higher than a predefined threshold $\theta$. However, searching for an optimal $\theta$ can be hard due to the feature representations of SAM lacking clear semantic-related knowledge [58, 76]. As empirical analysis presented in

Figure 4.4, many objects of different categories from the reference object may have a quite close similarity score to those objects of the same category, thus it is inevitable to output poor counting results with noisy segment generation. For instance, the SAM feature struggles to differentiate between apples and other messy objects in row 1, as well as children's hair and feet in row 2 of Figure 4.4. Moreover, our empirical findings (the first two rows in Table 4.3) indicate that the enhanced recall of referenced objects achieved through the proposed superpixel method does not improve the final counting performance when the SAM feature representation is directly employed for object discrimination. Therefore, to obtain a plausible counting accuracy, using semantic-rich features for mask proposal selection becomes indispensable.

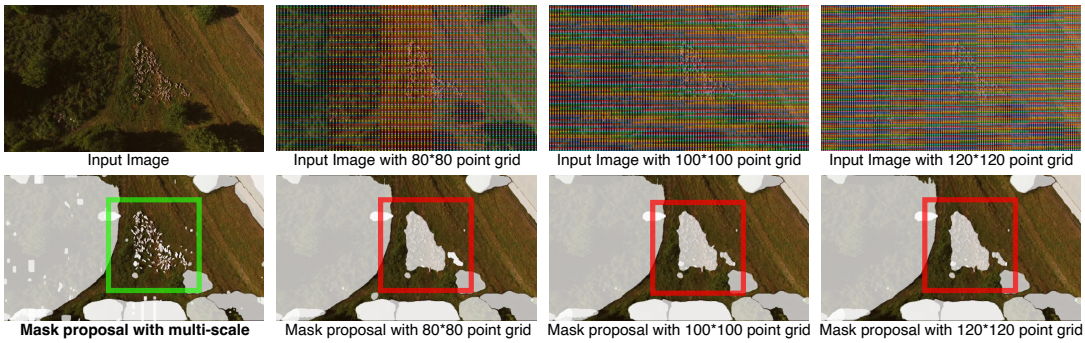



Figure 4.5: Effectiveness demonstration of our multi-scale mechanism in scenarios involving extremely tiny counting objects. Merely increasing the quantity of point prompts[1]in SAM fails to yield accurate instance-level mask proposals. However, integrating SAM with our multi-scale mechanism (using 32*32 points for aerial photography of sheep in this demonstration) effectively achieves better mask quality.

Recently, many vision foundation models with rich semantic knowledge have been made accessible to the public, such as DINO [10, 83].

[1]Multiple colors are used for the point grid in the figures of columns 2-4 for better visualization.

Combining with the semantic-rich feature representation, the final count can be obtained

$$
\begin{aligned}
&\text{count} = n_{\text{ref}} + \sum_{i=1}^{N} 1(\mathcal{S}^i > \theta), \quad \text{where} \\
&\mathcal{S}^i = \text{sim}(\mathsf{P}_{\text{ref}},\ F^i_{\text{proposal}}), \\
&\mathsf{P}_{\text{ref}} = \frac{1}{n_{\text{ref}}} \sum_{j=1}^{n_{\text{ref}}} \text{AvgPool}\left(\mathbb{E}_{\text{sem}}(I),\ h(\mathcal{M}^j_{\text{ref}})\right) \\
&F^i_{\text{proposal}} = \text{AvgPool}\left(\mathbb{E}_{\text{sem}}(I),\ h(\mathcal{M}^i_{\text{proposal}})\right)
\end{aligned}
\tag{4.3}
$$

where $\mathbb{E}_{\text{sem}}$ is the encoder with rich semantic knowledge and $h(\cdot)$ is the interpolation function to accommodate the resolution of SAM generated mask to that of the encoder output. $N$ is the number of object proposals. $\theta$ is the similarity threshold and is empirically set to 0.4 in this study.

### 4.4.3 Multi-scale Segmentation Helps

In the task of object counting, there exist some scenes where objects that need to be counted are extremely tiny and crowded, e.g., aerial photography of sheep. It becomes challenging for SAM to generate reasonable instance-level masks. As Figure 4.5 shows, naive increasing the number of points from 6400 to even 14400 still fails to make SAM generate precise mask proposals for the sheep instance. In order to tackle this problem, we propose a simple multi-scale mechanism. Specifically, we first cut the input image into $n_{\text{p}} \times n_{\text{p}}$ patches of the same size[1]. Then we resize these patches to the original size of the input image and feed them together with the original input image into SAM for mask proposal generation. Please see Figure 4.2 for details. As the bottom left picture in Figure 4.5 shows, using this simple multi-scale mechanism can successfully obtain precise instance-level mask proposals.

[1]Empirically, $n_{\text{p}} = 2$ is found to perform well in this study.

### 4.4.4 Transductive Prototype Updating Helps

In Sec. 4.4.2, the prototype of the interested object $\mathsf{P}_{\text{ref}}$ is simply the average of given few-shot reference object features. Since objects in the same class may have various appearances in the same image, this simple prototype may not be representative enough to match all interested objects. We argue that potentially interested-object features from mask proposals can be helpful for matching generalization and propose a novel transductive prototype updating strategy to improve the quality of the prototype

$$\mathsf{P}_{\text{ref}}^{\text{new}} = \frac{n_{\text{ref}} \cdot \mathsf{P}_{\text{ref}} + \sum_i^N 1(\mathcal{S}^i > \delta) \cdot F^i_{\text{proposal}}}{n_{\text{ref}} + \sum_i^N 1(\mathcal{S}^i > \delta)} \tag{4.4}$$

where $\delta$ is a similarity threshold and is set to 0.5 in this study for selecting plausible reference candidates.

## 4.5 Experiment

In this section, we first provide details of our experimental setup. Next, we assess our proposed method on two widely-used CAC benchmarks, comparing it against both training-based and training-free CAC methods. Lastly, we conduct comprehensive ablation studies to further explore the effectiveness of our approach.

### 4.5.1 Experimental Setup

**Dataset:** In our study, we assess our method using two popular datasets for counting tasks: FSC-147 [89] and CARPK [42]. FSC-147, the first extensive CAC dataset, includes 6,135 images across 147 categories, divided into 3,659 training, 1,286 validation, and 1,190 test images. Each image is annotated with bounding boxes and central points of three exemplars. Since our approach doesn't require training, we focus solely on the test set, which features 1,190 images from 29 categories. The CARPK dataset comprises 1,448 images of car parks captured from drones. Following the protocol in [24, 101, 102], we select twelve exemplars from the training images of CARPK for

use in all test images. Similar to FSC-147, our evaluation is confined to the test set, containing 459 images.

**Evaluation Metrics:** Follow [24, 101], we report the Mean Absolute Error (MAE), Root Mean Square Error (RMSE). In particular, MAE $= \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}|$, RMSE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2}$, where n is the number of test images, $y$ is the ground truth and $\hat{y}$ is the prediction.

**Comparing Methods:** Our approach is benchmarked against the latest state-of-the-art methods in both trained and training-free scenarios. For trained scenarios, we compare with several CAC techniques: GMN (General Matching Network [73]), FamNet (Few-shot adaptation and matching Network [89]), CFONet+ (Class-agnostic Fewshot Object Counting Network [131]), the advanced BMNet+ (Bilinear Matching Network [101]), SAFECount (similarity-aware feature enhancement [133]), and LOCA (Low-shot Object Counting network with iterative prototype Adaptation [24]). In the training-free setting, our models are evaluated alongside two existing methods: Count-Anything [76] and TFOC (training-free object counting [102]).

All experiments in this work are conducted on one NVIDIA GeForce RTX 3090 GPU and all other experimental settings (including the text prompt) are identical to the TFOC [102] approach for a fair comparison.

### 4.5.2 Comparison with State-of-the-Arts

**Quantitative Results:**

Table 4.1 reports the comparison results on two popular CAC benchmarks. First, our approach achieves consistent performance improvements over the compared training-free methods on various evaluation metrics. Specifically, On the FSC-147 dataset, our approach lowers the Mean Absolute Error (MAE) to 12.26, outperforming the naive SAM Baseline, which has an MAE of 42.48, and the Count-Anything method with an MAE of 27.97. Compared to the state-of-the-art training-free method TFOC [102], the performance gain of our approach reaches **+7.69** MAE on the FSC-147 dataset and **+6.58** MAE on CARPK dataset.

Meanwhile, the RMSE metric of our approach is also significantly lower than the TFOC method which demonstrates that our method can perform better in scenarios with more crowded objects.

| | Training | Reference Format | FSC-147 | | CARPK | |
|---|---|---|---|---|---|---|
| | | | MAE ↓ | RMSE ↓ | MAE ↓ | RMSE ↓ |
| GMN [73] [ACCV 19'] | Yes | Box | 26.52 | 124.57 | 9.90 | - |
| FamNet [89] [CVPR 21'] | Yes | Box | 22.08 | 99.54 | 18.19 | 33.66 |
| CFOCNet+ [131] [WACV 21'] | Yes | Box | 22.10 | 112.71 | - | - |
| BMNet+ [101] [CVPR 22'] | Yes | Box | 14.62 | 91.83 | 5.76 | 7.83 |
| SAFECount [133] [WACV 23'] | Yes | Box | 14.32 | 85.54 | 5.33 | 7.04 |
| PseCo [25] [CVPR 24'] | Yes | Box | 13.05 | 112.86 | - | - |
| LOCA [24] [ICCV 23'] | Yes | Box | 10.79 | 56.97 | 9.97 | 12.51 |
| SAM Baseline [102] | No | N.A. | 42.48 | 137.50 | 16.97 | 20.57 |
| Count-Anything [76] | No | Box | 27.97 | 131.24 | - | - |
| TFOC [102] [WACV 24'] | No | Box | 19.95 | 132.16 | 10.97 | 14.24 |
| **Ours** | No | Box | **12.26** | **56.33** | **4.39** | **5.70** |
| TFOC [102][WACV 24'] | No | Point | 20.10 | 132.83 | 11.01 | 14.34 |
| **Ours** | No | Point | **12.47** | **49.97** | **4.39** | **5.70** |
| TFOC [102][WACV 24'] | No | Text | 24.79 | 137.15 | - | - |
| **Ours** | No | Text | **23.59** | **113.60** | - | - |

Table 4.1: Quantitative comparisons between our methods and others using point and box reference formats on the FSC-147 and CARPK benchmarks are presented. Our approach surpasses other training-free methods and shows competitive results against training-based ones. The best counting outcomes among training-free methods are emphasized in bold font.

Second, our approach greatly reduces the performance gap between the training-based and training-free methods. For example, on FSC-147, the gap has decreased from 9.16 to 1.47 on MAE and even outperformed 0.64 on RMSE. On the CARPK dataset, our approach excels in all the compared algorithms and creates a state-of-the-art performance.

Finally, we also evaluate the proposed method with point references and the consistent superior performance demonstrates the robustness of our approach.

| | Training | FSC-147 | |
|---|---|---|---|
| | | **MAE ↓** | **RMSE ↓** |
| FamNet [89] [CVPR 21'] | Yes | 26.76 | 110.95 |
| BMNet+ [101] [CVPR 22'] | Yes | 16.89 | 96.65 |
| LOCA [24] [ICCV 23'] | Yes | **12.53** | 75.32 |
| TFOC [102] [WACV 24'] | No | 22.96 | 137.27 |
| **Ours** | No | 15.15 | **69.16** |

Table 4.2: Evaluation on one-shot setting on FSC-147 test set.

We also evaluate our pipeline on FSC-147 test set with a one-shot counting scenario. Table 4.2 indicates that our approach is robust on a one-shot counting scenario, which is consistent with the results of the three-shot above.

And for text prompts, we followed TFOC [102] and employ the CLIP-Surgery model to derive coarse bounding boxes for the target object using text references. Then our approach uses these coarse bounding boxes for object counting. As presented in last rows of the Table 4.1, our approach is sightly superior to the compared methods. The main reason is that the bounding boxes generated from TFOC [102] are not accurate. Technically, great improvement will be observed if high quality bounding boxes are given.

**Qualitative Analysis:**

Figure 4.6 displayed the visualization results of various counting methods in different scenarios and our approach can always more accurately identify objects of interest in various scenes. Specifically, for scenes in the first three rows, both of the compared training-based and training-free methods mistakenly identified areas that do not belong to the interested object, such as the teacup handle and the spoon in row 1, the crowded leaves in row 2 and the table legs and the top edge of the chair back in row 3. While our approach can successfully and accurately segment out the interested object and produce a more accurate counting result. This will be attributed to the utilization of semantic-rich feature representations for SAM's mask proposal selection in our approach. Furthermore, for more challenging scenes (the last two rows), the

density maps generated by the training-based BMNet+ method are vague and exhibit substantial deviations in counting outcomes. The TFOC method either erroneously includes extensive background areas in its object count or overlooks numerous objects in crowded scenes. Yet, our method yields accurate segmentation masks for individual objects due to its multi-scale mechanism, resulting in output counting numbers that more closely align with the ground truth.

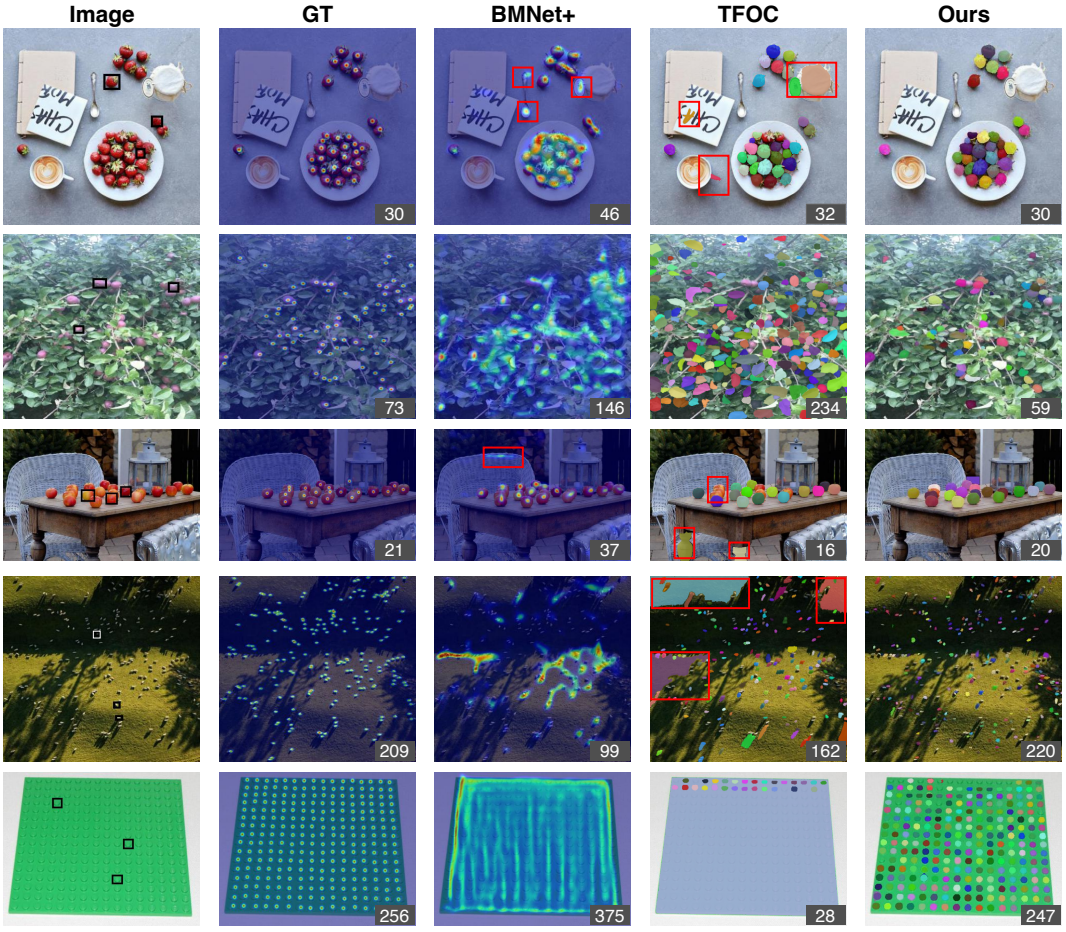



Figure 4.6: Qualitative results on the FSC-147 dataset are displayed. Exemplars are marked with black (white) boxes, and counting values are noted in the bottom-right corner. Error predictions are indicated by red boxes. Our method accurately predicts masks in both densely and sparsely populated scenes. Best viewed by zooming in.

## 4.6 Ablation study

| Components | | | | FSC-147 | |
|---|---|---|---|---|---|
| SP | DINOv2 | TPU | MS | MAE ↓ | RMSE ↓ |
| - | - | - | - | 30.91 | 140.10 |
| ✓ | - | - | - | 30.89 | 140.06 |
| - | ✓ | - | - | 20.19 | 134.22 |
| ✓ | ✓ | - | - | 17.32 | 132.64 |
| ✓ | ✓ | ✓ | - | 16.04 | 132.40 |
| ✓ | ✓ | - | ✓ | 13.77 | 87.40 |
| ✓ | ✓ | ✓ | ✓ | 12.26 | 56.33 |

Table 4.3: Performance with different components on FSC-147. SP: superpixel, TPU: transductive prototype updating, MS: multi-scale.

In this section, we are interested in ablating our approach from the following perspective views on FSC-147 dataset:

**Effectiveness of different components:** Table 4.3 showcases the impact of each component in our method in both dataset. Specifically, although SAM with superpixel can achieve higher interested-object recall (shown in Figure 4.3), due to the lack of semantic differentiation ability in SAM features (shown in Figure 4.4), simply replacing default point grid with superpixel does not necessarily bring performance gains, as the results shown in the first two rows of Table 4.3. The next two rows show that using DINOv2 feature representation for SAM's mask proposal selection significantly reduces the final counting error and the combing of superpixel can further bring performance gains. The subsequent two rows of Table 4.3 highlight the individual benefits of transductive prototype updating and the multi-scale mechanism on our model's effectiveness. When these four components are combined, they lead to the most impressive counting results, especially in the RMSE metric..

| number of K | MAE ↓ | RMSE ↓ | avg. time/image |
|---|---|---|---|
| 50*50 | 5.09 | 6.66 | 5.26s |
| 60*60 | 4.46 | 5.82 | 6.55s |
| 64*64 | 4.39 | 5.70 | 7.04s |

Table 4.4: Performance of our method with various K on CARPK.

**Different number of initialized points:** Table 4.4 shows that with 50*50 points, our method surpasses the previous SOTAs (see Table 4.1) in reasonable time. Increasing to 64*64 points further improves the MAE by 18%, while the computational time remains within seconds, demonstrating our method's ability to achieve significant performance gains with manageable time costs as the point density increases.

| | SAM Backbone | MAE ↓ | RMSE ↓ |
|---|---|---|---|
| TFOC [102] | ViT-B | 19.95 | 132.16 |
| | ViT-H | 30.41 | 138.28 |
| Ours | ViT-B | 14.43 | 87.54 |
| | ViT-H | 12.26 | 56.33 |

Table 4.5: Performance with different SAM backbone.

**Backbone size of SAM:** Table 4.5 compares TFOC and our method across different ViT backbones, revealing that our approach improves as the backbone scales from base to huge. In contrast, TFOC's performance declines with ViT-H, possibly due to an increase in mask proposals leading to poorer selection. Our method, leveraging semantically rich features, benefits from the higher number of generated object proposals.

**Various semantic-rich vision models:** Table 4.6 displays our method's performance across different SOTA foundation vision models, indicating that superior vision models yield improved counting accuracy. For CLIP, only the vision encoder is involved.

**Multi-round transductive prototype updating:** Our study typically employs one round of transductive prototype updating (TPU). Table 4.7 demonstrates that additional rounds do not enhance performance.

| Vision Model | Patch Grid | MAE ↓ | RMSE ↓ |
|---|---|---|---|
| CLIP [88] | 16*16 | 17.83 | 53.81 |
| DINO [10] | 28*28 | 14.28 | 59.33 |
| DINO v2 [83] | 37*37 | 12.26 | 56.33 |

Table 4.6: Performance with different semantic-rich model.

| Update Round | MAE ↓ | RMSE ↓ |
|---|---|---|
| 0 | 13.77 | 87.40 |
| 1 | 12.26 | 56.33 |
| 2 | 12.30 | 57.50 |

Table 4.7: Performance with various rounds of TPU.

## 4.7 Conclusion

This chapter presents a novel, training-free approach for Class-Agnostic Counting (CAC), significantly closing the performance gap with trained models. Key contributions include the use of superpixels, a richer semantic image encoder, a multiscale mechanism, and a transductive prototype updating strategy. Through rigorous experimental evaluation, the effectiveness of these four key components are demonstrated. The research shows that a CAC model without any training can achieve performance on par with training-based CAC approaches. This is a significant advancement in the field, indicating the potential of training-free approaches and setting a new benchmark for future research in CAC.

# Chapter 5

# Categorical Keypoint Positional Embedding for Robust Animal Re-Identification

We further explore the potential of powerful pretrained models to generate high-quality supplementary annotations, such as keypoints, using pretrained diffusion models. Keypoint annotations play a crucial role in fine-grained recognition tasks like wildlife animal re-identification (ReID), where distinguishing individuals based on subtle visual features is essential. However, manually annotating keypoints is time-consuming, labor-intensive, and often infeasible for large-scale datasets collected in diverse and uncontrolled environments. By leveraging pretrained diffusion models, we aim to automate and enhance the annotation process, reducing reliance on manual labeling while maintaining high-quality supervision.

# Statement of Authorship

| | |
|---|---|
| Title of Paper | Categorical Keypoint Positional Embedding for Robust Animal Re-Identification |
| Publication Status | ☐ Published ☐ Accepted for Publication<br>☑ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style |
| Publication Details | Lin, Yuhao, Lingqiao Liu, and Javen Shi. "Categorical Keypoint Positional Embedding for Robust Animal Re-Identification." arXiv preprint arXiv:2412.00818 (2024). |

## Principal Author

| | | | |
|---|---|---|---|
| Name of Principal Author (Candidate) | Yuhao Lin | | |
| Contribution to the Paper | Proposed the initial idea, developed the research concept, led results collection and analysis, and drafted and revised the manuscript. | | |
| Overall percentage (%) | 80% | | |
| Certification: | This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper. | | |
| Signature | | Date | 27 Dec 2024 |

## Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

| | | | |
|---|---|---|---|
| Name of Co-Author | Lingqiao Liu | | |
| Contribution to the Paper | Oversaw the project development and helped organize the experiments. | | |
| Signature | | Date | 6/1/2025 |

Please cut and paste additional co-author panels here as required.

| | | | |
|---|---|---|---|
| Name of Co-Author | Javen Qinfeng Shi | | |
| Contribution to the Paper | Oversaw the project development and revised the manuscript. | | |
| Signature | | Date | 13/01/2025 |

## 5.1 Abstract

Animal re-identification (ReID) has become an indispensable tool in ecological research, playing a critical role in tracking population dynamics, analyzing behavioral patterns, and assessing ecological impacts, all of which are vital for informed conservation strategies. Unlike human ReID, animal ReID faces significant challenges due to the high variability in animal poses, diverse environmental conditions, and the inability to directly apply pre-trained models to animal data, making the identification process across species more complex. This work introduces an innovative keypoint propagation mechanism, which utilizes a single annotated image and a pre-trained diffusion model to propagate keypoints across an entire dataset, significantly reducing the cost of manual annotation. Additionally, we enhance the Vision Transformer (ViT) by implementing Keypoint Positional Encoding (KPE) and Categorical Keypoint Positional Embedding (CKPE), enabling the ViT to learn more robust and semantically-aware representations. This provides more comprehensive and detailed keypoint representations, leading to more accurate and efficient re-identification. Our extensive experimental evaluations demonstrate that this approach significantly outperforms existing state-of-the-art methods across four wildlife datasets. The code will be publicly released.

## 5.2 Introduction

Animal re-identification (ReID) has emerged as a critical tool in ecological research, providing indispensable insights into the monitoring of wildlife populations, behavioral analysis, and ecological impact assessments. These insights are foundational to developing effective conservation strategies [97, 98, 114]. Unlike pedestrian ReID, animal ReID in natural settings presents formidable challenges that stem from the limited availability of annotated data, highly variable environmental conditions, and inability to directly apply pre-trained models to animal data [54] that obscure crucial visual information. These challenges are compounded by intrinsic variations in animal shapes, postures, and patterns, which significantly complicate the ReID process [91, 60].

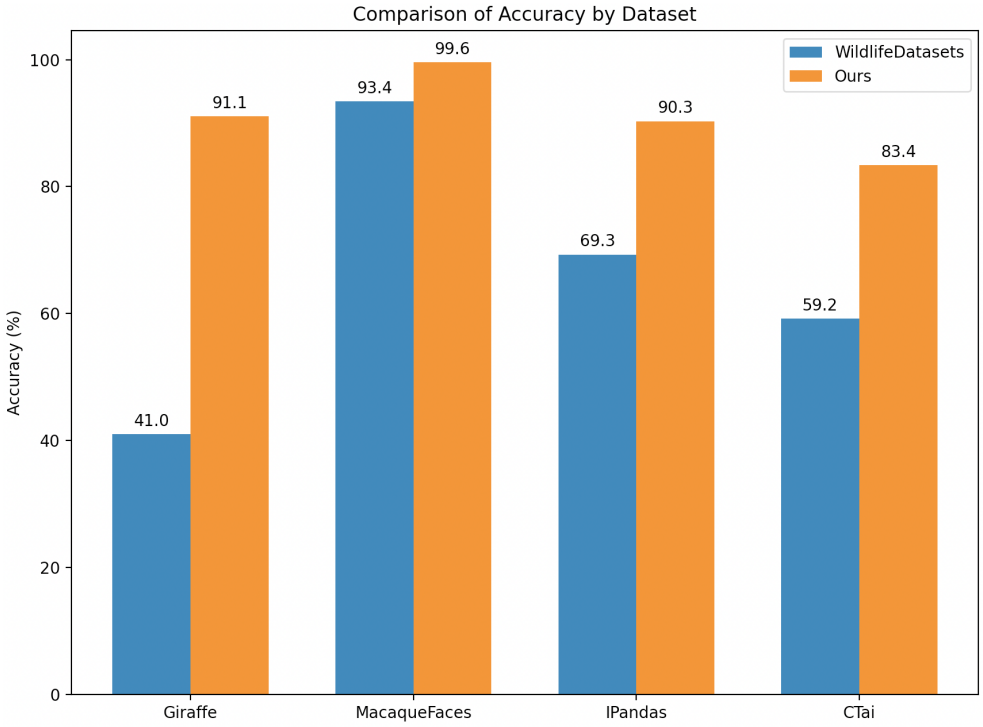



Figure 5.1: Comparison of model performance between the previous state-of-the-art (SOTA) and our proposed method across four datasets. Our approach significantly outperforms the previous SOTA, demonstrating notable improvements in accuracy for all datasets.

Advancements in animal ReID methodologies, though diverse, share a unified goal of accurately and efficiently recognizing individual animals within a species based on unique phenotypic traits such as markings, patterns, or other distinctive features [21, 60]. Traditional approaches frequently require manual selection of features at specific anatomical locations—such as ears or fins—which is labor-intensive and requires extensive expert knowledge, rendering these methods impractical for processing large wildlife datasets [55]. Recent studies have underscored the importance of keypoints in ReID tasks, as demonstrated through the introduction of the Keypoint-Promptable Transformer (KPR) for human ReID, highlighting the pivotal role of keypoints in enhancing identification accuracy [108]. However, these methods are primarily effective on human-centric datasets and face significant challenges when applied to the anatomical and behavioral diversity of animals. The emergency of corresponding diffusion features

has significantly enhanced the capability for effective keypoint propagation across images within a dataset, thereby facilitating this process.[109], e.g., as demonstrated in Figure 5.2.

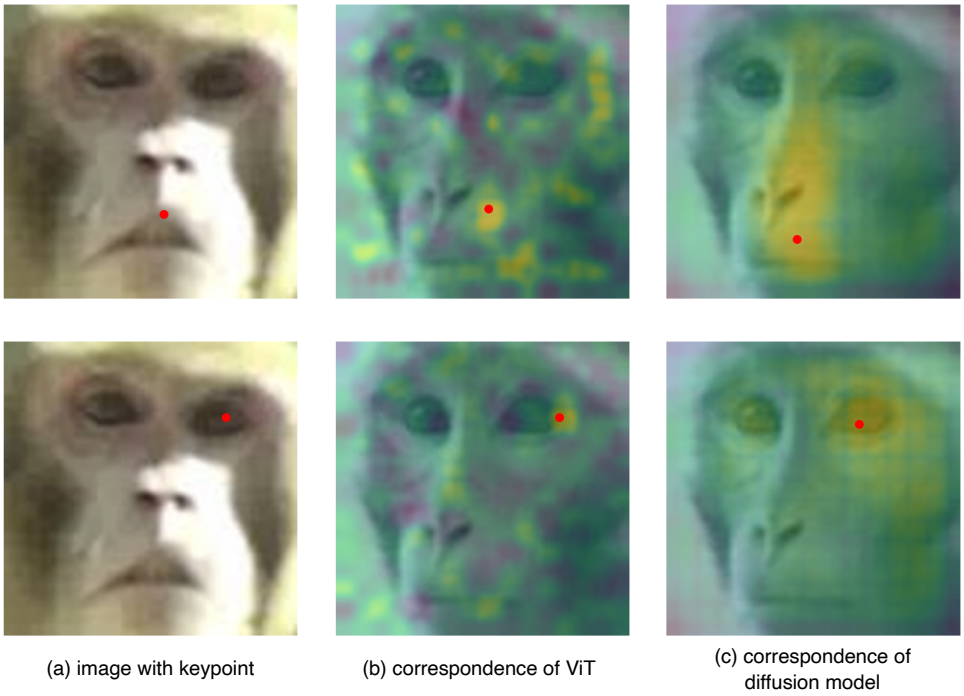



Figure 5.2: Comparison of keypoint similarity heatmaps between a pre-trained Vision Transformer (ViT) model and a pre-trained diffusion model. The most salient feature in the heatmap, indicating the keypoint, is highlighted. (a) is the image and the keypoint, above is the mouth and bottom is the right eye. (b) is the similarity heatmaps from ViT and (c) is the similarity heatmaps from Stable Diffusion. It is obvious that the pre-trained diffusion model exhibits stronger semantic correspondence compared to the pre-trained ViT model.

Motivated by these insights, our study introduces Keypoint Propagation, a novel mechanism where keypoints annotated on a single image by GPT-4 can be automatically propagated throughout the entire dataset, utilizing a pre-trained diffusion model that adapts to the significant variability inherent in wildlife imagery. We further propose

Keypoint Positional Embedding (KPE) and Categorical Keypoint Positional Embedding (CKPE). These two novel Embedding strategies enhance the capabilities of the Vision Transformer (ViT), enabling it to recognize and interpret the semantic relevance of keypoints. In specific, The KPE method encodes the spatial information of keypoints into a high-dimensional space, aiding the ViT in focusing on these positions during the learning process. Extending this, our CKPE approach incorporates categorical data about the keypoints, providing the model with semantic awareness necessary for distinguishing closely related species or individuals within species. Our primary contributions are as follows:

- We introduce a diffusion-based keypoint propagation mechanism that minimizes the need for extensive annotations.
- We propose Keypoint Positional Embedding (KPE) and Categorical Keypoint Positional Embedding (CKPE) to enhance the Vision Transformer's (ViT) ability to recognize and contextualize keypoints effectively.
- Our methods outperform existing state-of-the-art approaches in occluded ReID tasks, as evidenced by significant accuracy improvements ranging from +5.9% to +50.1% on four publicly available benchmarks (shown in Fig 5.1), respectively.

To facilitate further research and practical applications in this field, we will make our codebase, datasets, and detailed annotations publicly available, providing a comprehensive framework for the evaluation of promptable wildlife ReID methods.

## 5.3 Related Work

### 5.3.1 Animal ReID

Recent advances in deep learning, particularly in convolutional neural networks (CNNs), have significantly advanced wildlife re-identification (ReID), with notable contributions in feature extraction [3, 61] and metric learning [64, 124]. While person ReID has reached a certain level of maturity, wildlife ReID is still in its nascent stages, with many existing methods being species-specific, thereby limiting their broader applicability [35, 77, 82]. Current approaches can be broadly categorized into three main strategies: (1) *Global Feature Learning*, which utilizes whole animal images to extract global features. These methods, such as the multi-stream feature fusion network for giant pandas [116] and the specialized loss for manta ray ReID [81], are adaptable across species but often rely on techniques originally developed for person ReID. (2) *Species-specific Feature Extraction*, which targets distinct anatomical features, such as dolphin fins [9, 53, 121], whale tails [12], and elephant ears [120]. These methods, while effective within a specific species, are limited by challenges such as occlusions and viewpoint variations, which hinder their generalizability across species. (3) *Auxiliary Information Integration*, which incorporates pose keypoint estimation to enhance feature learning, as demonstrated in tiger [60] and yak [139] ReID. While these approaches can improve accuracy, they rely on auxiliary data, making them difficult to standardize across diverse species. Recently, the WildlifeDatasets toolkit [113] was introduced to establish a standard baseline for animal ReID, alongside several general approaches[46, 104] and datasets aimed at developing a universal and robust solution for wildlife ReID. Despite these advancements, there remains a critical need for a clearly defined task and high-performance solutions that can generalize across a wide range of species.

### 5.3.2 ViT for ReID

While convolutional neural network (CNN)-based methods have long dominated person and vehicle re-identification (ReID), the introduction of the Vision Transformer (ViT) has significantly advanced the field [132]. In general object ReID, models such

as TransReID [38], which is entirely ViT-based, have shown improved cross-camera ReID accuracy by emphasizing local feature learning and viewpoint information. Additionally, several hybrid models [59, 62, 137] combine ViT with CNNs, with techniques like dual cross-attention learning [142] enhancing both global and local feature learning through more refined attention mechanisms. Some methods [117] also leverage auxiliary information, such as pose estimation, to improve the extraction of human body-related features. [108] introduced keypoints from pre-trained human keypoint detection models, demonstrating the importance of auxiliary prompts. However, these prompts are only effective on human datasets and do not generalize well to wildlife images due to the increased complexity of animal scenes. In this chapter, we propose a novel mechanism for detecting and utilizing keypoints specifically in wildlife images.

### 5.3.3 Visual foundation model with correspondences

Recent advancements in visual foundation models, such as Vision Transformers (ViTs) and Stable Diffusion (SD), have demonstrated that pretrained features—learned through classification tasks or generative tasks [4, 5, 37, 87]—serve as powerful descriptors for semantic matching, often outperforming previous methods specifically designed for this task. Stable Diffusion, trained on billions of text-image pairs [99], has led to the development of widely adopted open-source text-to-image models. Studies [39, 109] have shown that correspondences can emerge in image diffusion models without explicit supervision, thereby suggesting the potential for the keypoint propagation within these models.

## 5.4 Method

In this section, we start from task definition (Sec. 5.4.1) and then introduce Keypoint Detection and Propagation (Sec. 5.4.2), Keypoint Positional Embedding (Sec. 5.4.3), Categorical Keypoint Positional Embedding (Sec. 5.4.4), and entire training testing process (Sec. 5.4.5). The comprehensive framework of our method is depicted in Figure 5.3.

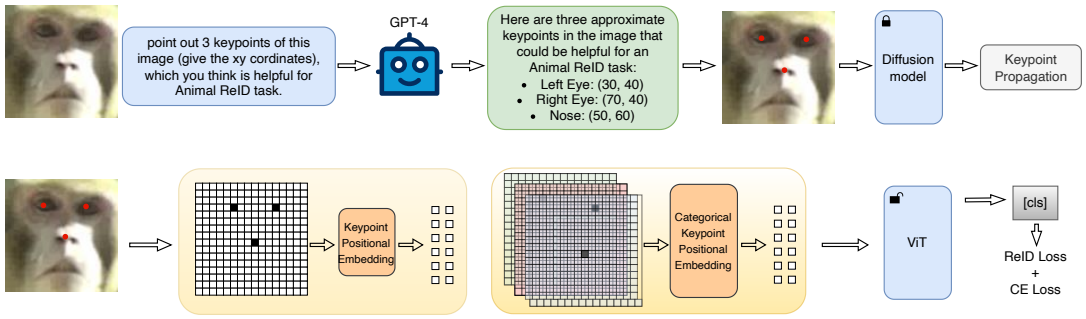



Figure 5.3: Architecture Overview of Our Proposed Categorical Keypoint Positional Embedding for Wild Animal ReID: Initially, GPT-4 identifies critical keypoints for the ReID task from a single image. Subsequently, a diffusion model is employed to propagate these keypoints across the entire dataset. The identified keypoints then inform our ViT-based ReID module through Keypoint Positional Embedding or Categorical Keypoint Positional Embedding, focusing on the content at these specific locations and semantic information to enhance feature discrimination and accuracy.

### 5.4.1 Task forming

Animal Re-Identification (ReID) is a computer vision task aimed at identifying and matching individual animals across different non-overlapping camera captures. We formally define this task as follows:

**Given:**

- A **training set** $\mathcal{T}$ used to train the model, which includes images and their ID labels.

- A **query set** $\mathcal{Q}_{\text{test}} = \{q_1, q_2, \ldots, q_n\}$ and a **gallery set** $\mathcal{G}_{\text{test}} = \{g_1, g_2, \ldots, g_m\}$ for testing. The training and testing sets are distinct, ensuring no overlap to simulate real-world scenarios where the model encounters completely unseen samples. This separation tests the model's generalization capabilities and effectiveness in practical applications.

**Objective:**

- Learn a mapping function $f : \mathcal{I} \to \mathbb{R}^d$ that maps an image $I$ to a point in a $d$-dimensional Euclidean space. This function is optimized so that images of the same animal are closer together, and those of different animals are farther apart.
- For each query image $q_i \in \mathcal{Q}_{\text{test}}$, find the set $\{g_j \in \mathcal{G}_{\text{test}} : \text{identity}(q_i) = \text{identity}(g_j)\}$, using the trained model to predict and match identities based on learned representations.

### 5.4.2 Keypoint Detection and Propagration

Given an input image, we begin by leveraging GPT-4 to identify keypoints with high discriminative power, which are critical for the re-identification (ReID) task. These keypoints, identified by GPT-4, capture critical spatial locations that distinguish different instances of the same species. Once identified, these keypoints are mapped to corresponding locations across the entire dataset using a pre-trained stable diffusion model. Specifically, let $p_i$ represent the location of the $i$-th keypoint in an image $I$, where $i \in [1, N]$ and $N$ denotes the total number of keypoints. Using the diffusion model, matching keypoints across all images in the dataset are retrieved, creating a consistent set of keypoints for each image, denoted as $\{p_1, p_2, \ldots, p_N\}$.

The process of keypoint detection and propagation is crucial in mitigating the data sparsity issue commonly faced in wildlife ReID, where manual annotation is expensive and often impractical. To compute the feature similarity map between two images $I^{ref}$ (the reference image) and $I^{target}$ (the target image), we first extract their respective diffusion features, $\mathbf{F}^{ref}$ and $\mathbf{F}^{target}$. For each keypoint location $p_i$, defined by the

coordinates $(x_i, y_i)$, we extract the corresponding feature vector $v_{p_i} = F^{ref}_{x_i,y_i}$ from the reference image. These feature vectors are normalized to unit length, resulting in normalized feature vectors $\hat{\mathbf{v}}_{p_i}$.

To compute the similarity map between keypoints and target image, we define the cosine similarity map between the normalized feature vector at the keypoint $p_n$ in the reference image and the feature map of the target image. The cosine similarity is expressed as:

$$S_n = \hat{\mathbf{v}}_{p_n} \cdot \mathbf{F}^{target}$$

where $S_n$ represents the similarity map between the feature vector at $p_n$ and the target image. This computation yields a similarity map $\mathbf{S}$ for each keypoint location in the reference image, where the point with the highest similarity score in $S_n$ corresponds to the matching keypoint location in the target image.

$$p_n = argmax\ S_n$$

By this, we can propagate the keypoints across the dataset effectively, ensuring consistent keypoint localization even when faced with varying poses, occlusions, and environmental conditions. By utilizing both GPT-4 for initial keypoint detection and stable diffusion models for propagation, we significantly reduce the need for manual annotations while enhancing the ReID model's ability to handle the variability present in real-world wildlife imagery.

### 5.4.3 Keypoint Positional Embedding (KPE)

We introduce the **Keypoint Positional Embedding (KPE)** algorithm, designed to enhance the Vision Transformer (ViT) by integrating both visual and spatial information derived from keypoints in an image. Given an image $I$ with a set of keypoints $\{p_1, p_2, \ldots, p_N\}$, the KPE algorithm works by mapping the identified keypoints $p$ into a one-hot vector mask representation. This mask representation is designed to encode

the spatial presence of each keypoint relative to the patches of the image as used by the ViT model. Specifically, for image $I$, we create the mask vector $\mathbf{M}_I$, where each element of the vector corresponds to a patch in the image. We define the mask as follows:

$$\mathbf{M}_I = \text{OneHot}(p) \in \mathbb{R}^{K+1}$$

where $K$ is the total number of image patches (corresponding to the patch configuration of ViT), and $+1$ accounts for the inclusion of the [CLS] token in the ViT architecture. For each patch in the image, if it contains the keypoint $p$, the corresponding value in the mask is set to 1; otherwise, it is set to 0. This one-hot encoding efficiently captures the spatial distribution of the keypoints across the image and ensures that the ViT can learn from both the content of the patches and the locations of the keypoints.

In the KPE approach, we introduce a learnable weight matrix $W_{\text{kp}} \in \mathbb{R}^{(K+1)\times d}$, where $d$ is the embedding dimension. The matrix $W_{\text{kp}}$ serves to project the one-hot vector mask into a high-dimensional embedding space, allowing the model to capture the spatial relationships between the keypoints and the image patches. For each keypoint mask $\mathbf{M}_i$, the corresponding embedding vector is computed by the following matrix multiplication:

$$\mathbf{M}_i \cdot W_{\text{kp}} \in \mathbb{R}^d$$

This embedding vector is then added to the image patch features extracted by the ViT model. By incorporating the spatial information of the keypoints into the ViT's attention mechanism, the KPE algorithm enables the model to focus on regions of the image that are most relevant for the ReID task, while also leveraging the semantic understanding provided by the keypoints. The integration of this keypoint embedding allows the ViT to learn both the visual patterns and the spatial positioning of keypoints, improving the model's ability to handle complex real-world scenarios where precise localization and recognition are crucial.

### 5.4.4 Categorical Keypoint Positional Embedding (CKPE)

Building upon the Keypoint Positional Embedding (KPE), we propose the **Categorical Keypoint Positional Embedding (CKPE)**, which enhances the model by incorporating category-specific information for each keypoint. The inclusion of this categorical information is particularly useful for tasks where different keypoints may represent distinct anatomical or semantic features, such as facial landmarks or body parts. For instance, in the case of a monkey face dataset, let $C_i$ denote the category of the $i$-th keypoint $p_i$, where $C_i \in \{\text{left eye}, \text{right eye}, \text{nose}\}$. This categorization allows the model to differentiate between various types of keypoints within the same image, providing it with more granular semantic information.

To encode this categorical information, we construct a categorical mask for the keypoints in the image. The categorical mask $\mathbf{CM}_I$ is a one-hot vector representation that maps each keypoint $p_i$ to its corresponding category. Specifically, we define the categorical mask as follows:

$$\mathbf{CM}_I = \text{OneHot}(P_I) \in \mathbb{R}^{(K+1)\times N_c}$$

where $K$ is the total number of image patches, which corresponds to the patch configuration in the ViT model, and $N_c$ is the number of categories associated with the keypoints (e.g., for the monkey face dataset, $N_c = 3$, corresponding to "left eye", "right eye", and "nose"). For each patch in the image, if it contains the keypoint $p_i$ of category $C_i$, the corresponding value in the mask is set to 1; otherwise, it is set to 0. This one-hot encoding allows the model to learn not only the spatial distribution of keypoints but also their specific categories, enriching the contextual understanding of each keypoint's role in the image.

To further enhance the model's ability to learn category-specific features, we introduce a learnable category-specific embedding matrix $W_{\text{ckp}} \in \mathbb{R}^{N_c \times d}$, where $N_c$ is the number of categories and $d$ is the embedding dimension. This matrix projects the one-hot vector representation of the categorical mask into a higher-dimensional space, allowing the model to capture the relationships between keypoints and their respective categories.

The categorical embedding for keypoints in image $I$ is then computed as:

$$\mathbf{CM}_I \cdot W_{\text{ckp}} \in \mathbb{R}^d$$

This embedding vector, representing the categorical information of the keypoints, is subsequently added to the image patch features extracted by the ViT model. The integration of categorical embeddings into the ViT's attention mechanism allows the model to not only focus on the spatial location of keypoints but also learn their semantic significance, such as distinguishing between different facial features in the case of monkey faces or differentiating body parts in animal datasets.

The advantage of CKPE lies in its ability to provide both spatial and semantic context for each keypoint, which is particularly beneficial for tasks such as wildlife ReID or human pose estimation, where the model must differentiate between various types of keypoints (e.g., eyes, ears, nose) that belong to the same individual or species. By embedding categorical data alongside positional information, CKPE enables the ViT to learn more robust and semantic-aware representations, providing a more comprehensive and nuanced representation of keypoints, ultimately leading to more accurate and efficient re-identification.

### 5.4.5 Training and Testing

Once the keypoint and categorical embeddings have been integrated into the Vision Transformer (ViT), we proceed with fine-tuning the model for the re-identification task. Specifically, we use the [CLS] token as a global feature representation, allowing the model to aggregate visual, positional and semantic information from the keypoints, including their spatial and categorical relationships. This enables the ViT to simultaneously learn discriminative features from the keypoints and their positional contexts within the image. The objective is for the model to learn a rich representation that not only captures the individual characteristics of each keypoint but also the semantic importance of its category, thereby improving the overall performance of the re-identification task.

For training, we define a composite loss function that consists of two primary components: the classification loss and the re-identification loss. The total loss function is given by:

$$\mathcal{L} = \mathcal{L}_{\text{reid}} + \lambda \mathcal{L}_{\text{ce}}$$

where the re-identification loss, $\mathcal{L}_{\text{reid}}$, is ArcFace loss based on a triplet loss function for re-identification task. $\mathcal{L}_{\text{ce}}$ represents the classification loss, which is cross entropy loss, weighted by a scalar $\lambda$ to balance the two objectives.

During the training phase, the model simultaneously learns to classify keypoints correctly, predict their categories, and distinguish between different identities in the dataset. The [CLS] token serves as the aggregated feature representation for each image, which encapsulates both the visual and semantic content, including the keypoint features and their respective categories. This joint optimization ensures that the learned representation effectively supports both the spatial and categorical aspects of the task.

For testing, we use the [CLS] embedding extracted from the trained ViT model. During the testing phase, the model evaluates the similarity between the query image and the gallery of test images, which consist of individuals never seen during the training process. To perform the matching, we compute the cosine similarity between the [CLS] embedding of the query image and the embeddings of the gallery images. Specifically, for a given query image $I^q$ and a set of gallery images $\{I_1^g, I_2^g, \ldots, I_N^g\}$, we compute the cosine similarity $S$ as:

$$S = \frac{\mathbf{z}^q \cdot \mathbf{z}^{\mathcal{G}}}{\|\mathbf{z}^q\|_2 \|\mathbf{z}^{\mathcal{G}}\|_2}$$

where $\mathbf{z}^q$ is the [CLS] embedding of the query image and $\mathbf{z}^{\mathcal{G}}$ is the [CLS] embeddings of gallery images. The gallery image with the highest similarity score to the query image is selected as the most likely match.

## 5.5 Experiment

In this section, we first provide details of our experimental setup. Next, we assess our proposed method on four widely used Animal ReID benchmarks, comparing it against previous SOTA methods. Lastly, we conduct comprehensive ablation studies to further explore the effectiveness of our approach.

| | MacaqueFaces | Giraffe | IPandas | CTai |
|---|---|---|---|---|
| WildLife Dataset[113] | 93.4% | 41.0% | 69.3% | 59.2% |
| ours (KPE) | 99.1% | 89.6% | 88.2% | 81.6% |
| ours (CKPE) | 99.6% | 91.1% | 90.3% | 83.4% |

Table 5.1: Quantitative comparisons between our methods and previous SOTA on the 4 public wildlife ReID benchmarks are presented. Our approach surpasses wildlife dataset and shows effectiveness of our KPE and CKPE.

### 5.5.1 Experimental Setup

**Dataset:** In our study, we assess our method using four popular public datasets for Animal ReID tasks: MacaqueFaces [123] and Giraffe [78]. "MacaqueFaces" dataset comprises 6,280 photographs of 81 individual rhesus macaques, designing for facial recognition and animal ReID. Similarly, the "Giraffe" dataset includes 1,393 images of 178 individual giraffes. The CTai [29] dataset contains 4,662 images of chimpanzee faces from 71 individuals, while the IPandas [116] dataset includes 6,874 images of 50 pandas. All the data split is 80% of individuals for training, and 20% for testing.

**Evaluation Metrics:** Follow [103, 113], we report the accuracy. In particular, Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$, where TP (True Positives) is the count of correctly identified positive cases. TN (True Negatives) is the count of correctly identified negative cases. FP (False Positives) is the count of negative cases incorrectly identified as positive. FN (False Negatives) is the count of positive cases incorrectly identified as negative.

All experiments in this work are conducted on one NVIDIA GeForce RTX 3090 GPU and all other experimental settings (except the data division) are identical to the WildlifeDatasets [113] approach for a fair comparison. $\lambda$ in $\mathcal{L}$ is set to 0.2 empirically.

### 5.5.2 Comparison with State-of-the-Arts

Table 5.1 presents a comprehensive comparison of our method against the state-of-the-art (SOTA) on four popular Wildlife ReID benchmarks. Our approach demonstrates consistent and substantial performance improvements across all datasets. Specifically, on the Giraffe dataset, our method achieves an accuracy of 91.1%, representing a significant improvement over the previous SOTA, which achieved only 41.0%. On the other three publicly available datasets—MacaqueFaces, IPandas, and CTai—our method outperforms the previous SOTA by 6.2%, 21.0%, and 24.2%, respectively. These results highlight the robustness and generalizability of our approach across different species and environments. Furthermore, the performance gain observed when using our **Categorical Keypoint Positional Embedding (CKPE)** over **Keypoint Positional Embedding (KPE)** underscores the effectiveness of incorporating both spatial information and semantic categories of keypoints, which plays a crucial role in improving model accuracy and robustness in wildlife ReID tasks.

## 5.6 Ablation study

In this section, we are interested in ablating our approach from the following perspective views on MacaqueFaces dataset (default size of ViT is large):

### 5.6.1 Different approaches to find keypoints

In this ablation study, we systematically compare several keypoint detection approaches to evaluate their performance in identifying discriminative regions in single images. We investigate four different methods: **Grad-CAM (ResNet)**, **Grad-CAM (ViT)**, **Experts (from the School of Biological Sciences)**, and **GPT-4**. Each method

brings unique characteristics and strengths, which we examine in the context of wildlife re-identification. The goal is to assess the accuracy and consistency of keypoint localization across various detection algorithms and identify the most effective approach for this task.

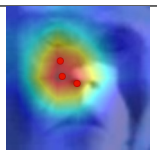
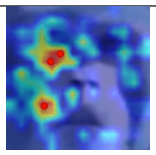
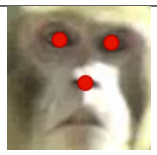
| Methods | Grad-CAM (ResNet) | Grad-CAM (ViT) | Expert | GPT-4 |
|---|---|---|---|---|
| kp |  |  |  |  |
| Acc | 96.9% | 97.7% | 99.20% | 99.20% |

Table 5.2: Effectiveness of different approaches of defining 3 keypoints for single image.

- **Grad-CAM (ResNet)**: Grad-CAM (Gradient-weighted Class Activation Mapping) applied to the ResNet architecture generates heatmaps by using the gradients of the target class with respect to the final convolutional layer. Grad-CAM highlights areas of the image that contribute most to the classification decision. In our case, Grad-CAM applied to ResNet tends to focus on the general area around the nose, capturing some key features of the face but with limited precision.

- **Grad-CAM (ViT)**: In contrast to ResNet, Grad-CAM applied to the Vision Transformer (ViT) visualizes attention maps from transformer-based models. ViT leverages self-attention mechanisms that enable the model to focus on different parts of the image. However, when applied to keypoint detection, Grad-CAM on ViT appears to focus on scattered and random points across the image, making it less effective for precise keypoint localization. While ViT excels in capturing global relationships within an image, its application to keypoint localization in wildlife images shows some limitations in pinpointing accurate anatomical features.

- **Experts (from the School of Biological Sciences)**: To establish a benchmark for keypoint detection, we incorporate expert annotations from professionals in biological sciences. These experts, with extensive domain-specific knowledge, manually identify keypoints based on anatomical landmarks. For example, in the case of monkey face images, experts can pinpoint key features such as the eyes, nose, and mouth with high accuracy. This provides a reliable and accurate ground truth for keypoints comparison.

- **GPT-4**: Lastly, we explore the use of GPT-4, a state-of-the-art language model that has been shown to perform well in multi-modal tasks involving both vision and language. By leveraging GPT-4's advanced understanding of contextual relationships, we apply it to identify keypoints in wildlife images. GPT-4 shows remarkable accuracy in detecting keypoints' locations, focusing on the most distinctive areas of the image, such as the eyes and nose. In this study, GPT-4 performs on par with expert annotations, demonstrating its ability to effectively localize keypoints without requiring explicit image labeling or manual intervention.

Table 5.2 summarizes the keypoint positions and the corresponding accuracy achieved by each method. The heatmaps generated by Grad-CAM on ResNet predominantly highlight the nose region, though it lacks the precision required for reliable keypoint localization. On the other hand, Grad-CAM on ViT appears to focus on random and scattered regions of the face, resulting in a less focused and less accurate keypoint detection. In contrast, both expert annotations and GPT-4 successfully localize keypoints in more distinctive regions such as the eyes and nose, providing more precise and relevant keypoint positions.

The accuracy results further support the effectiveness of GPT-4 in keypoint detection. Specifically, the performance of GPT-4 matches that of the expert annotations, demonstrating its high accuracy in identifying keypoints. Furthermore, GPT-4 outperforms both ResNet and ViT-based Grad-CAM methods, surpassing their accuracy significantly. These findings highlight the potential of GPT-4 as a powerful tool for keypoint localization, especially in tasks where manual annotations are unavailable.

| Time step | 0 | 50 | 100 | 150 | 200 | 250 | 300 | 500 | 1000 |
|---|---|---|---|---|---|---|---|---|---|
| Acc | 98.9% | 98.9% | 98.2% | 98.9% | 99.2% | 99.2% | 99.2% | 97.2% | 96.9% |

Table 5.3: Comparison between performance of keypoint propagation from different timesteps diffusion feature.

### 5.6.2 Different Timesteps to Propagate the Keypoints

In Stable Diffusion, the process of keypoint propagation is influenced by the timestep at which the diffusion model operates. Each timestep in the diffusion process represents a stage in the image generation or feature transformation, with earlier timesteps capturing more coarse, global features and later timesteps refining those features into more detailed representations. To investigate the effect of different timesteps on keypoint propagation, we compare the performance of keypoints extracted from multiple timesteps during the diffusion process.

Table 5.3 presents the performance of keypoint propagation at different diffusion timesteps. Specifically, we evaluate the accuracy of keypoint localization when propagated from the 0-300 timestep range, compared to the performance at later stages such as timestep 500 and timestep 1000. The results indicate that keypoints propagated from timesteps in the range of 0-300 yield stable and competitive performance, aligning with similar findings reported in [109]. These early timesteps appear to capture the most significant features related to keypoint locations, offering a balanced trade-off between feature accuracy and computational efficiency.

On the other hand, timesteps 500 and 1000, which correspond to later stages in the diffusion process, show diminishing returns in terms of performance when compared to the 0-300 timestep range. This suggests that while later timesteps may provide increasingly detailed features, they do not significantly improve the localization of keypoints in this particular task, and may even introduce noise or irrelevant details that detract from the model's focus on the keypoint positions.

| ViT Scale | W/O keypoints | KPE | CKPE |
|---|---|---|---|
| base | 92.5% | 98.5% | 97.3% |
| large | 96.9% | 98.9% | 99.2% |
| huge | 98.9% | 99.1% | 99.6% |

Table 5.4: Effectiveness of our KPE and CKPE on different scales of ViTs.

### 5.6.3 Different ViT Scales

We also evaluate the effectiveness of Keypoint Positional Embedding (KPE) and Categorical Keypoint Positional Embedding (CKPE) across different scales of Vision Transformers (ViTs), ranging from ViT-Base to ViT-Huge. As shown in Table 5.4, both KPE and CKPE consistently improve performance across all ViT scales. The improvements are particularly noticeable in the smaller ViT-Base models, where the embeddings help overcome the model's limited capacity to model complex relationships. As the scale increases to ViT-Large and ViT-Huge, the performance gains remain significant, indicating that KPE and CKPE continue to enhance the model's ability to refine spatial and categorical keypoint information, benefiting from the increased capacity of the larger models. These results demonstrate that KPE and CKPE offer scalable performance improvements, complementing the strengths of ViTs regardless of model size, making them effective across a range of ViT configurations, from base to huge.

### 5.6.4 Different numbers of keypoints

| NO keypoint | 1 | 3 | 6 | 10 |
|---|---|---|---|---|
| base | 97.3% | 97.3% | 97.8% | 96.9% |
| large | 97.6% | 99.2% | 98.9% | 97.3% |
| huge | 98.8% | 99.6% | 99.2% | 98.3% |

Table 5.5: Effectiveness of Keypoint Numbers (1, 3, 6, 10) for Animal Re-Identification

Table 5.5 presents the performance of Categorical Keypoint Positional Embedding (CKPE) with varying numbers of keypoints, ranging from 1 to 10. The results demonstrate that our method achieves consistent and solid improvements across different

numbers of keypoints, with performance gradually stabilizing as the number of keypoints increases. Notably, the addition of more than a few keypoints does not result in significant gains, suggesting that the model has reached an optimal balance between keypoint quantity and performance. This indicates that while a small number of well-chosen keypoints can substantially improve the model's accuracy, further increasing the number of keypoints does not provide additional benefits, and may even introduce redundancy. These findings highlight that the most critical factor in keypoint selection is the quality and relevance of the chosen keypoints, rather than their sheer quantity.

### 5.6.5 Keypoints vs Random points

| ViT | random | keypoint |
|---|---|---|
| base | 95.6% | 97.3% |
| large | 96.2% | 97.6% |
| huge | 98.8% | 98.8% |

Table 5.6: Comparison between performance of 1 keypoint vs 1 random point

| ViT | random | keypoint |
|---|---|---|
| base | 96.9% | 98.5% |
| large | 98.8% | 99.2% |
| huge | 99.2% | 99.6% |

Table 5.7: Comparison between performance of 3 keypoints vs 3 random points

Tables 5.6 and 5.7 compare the effectiveness of keypoints against randomly selected points, using 1 and 3 points, respectively. The results clearly demonstrate that keypoints provide significant improvements over randomly chosen points, highlighting their crucial role in enhancing the performance of animal re-identification (ReID). Specifically, keypoints enable the model to focus on discriminative regions of the image, which are essential for accurately identifying and distinguishing between individuals. In contrast, random points do not provide the same level of semantic or spatial relevance, leading to suboptimal performance. These findings underscore the importance of select-

ing informative keypoints, as they provide valuable spatial context and semantic meaning that random points cannot replicate, ultimately playing a critical role in boosting the model's ability to perform robustly in wildlife ReID tasks.

## 5.7 Conclusion

In this work, we introduce a novel keypoint propagation mechanism, providing a potential solution for keypoint detection in wildlife datasets. We also propose a new approach for wildlife re-identification (ReID) that combines keypoint-based feature extraction with Keypoint Positional Embedding (KPE) and Categorical Keypoint Positional Embedding (CKPE). Our method achieves significant performance improvements across multiple wildlife ReID benchmarks. We demonstrate that keypoint-based embeddings, particularly when enhanced with categorical information, play a crucial role in improving model accuracy. We believe that our approach, along with the insights gained from ablation studies, will contribute to the advancement of wildlife re-identification and related ecological monitoring tasks.

# Chapter 6

# Conclusion and Future Work

## 6.1 Conclusion

In this thesis, we have strategically explored how leveraging semi-supervised learning and pretrained models can significantly reduce the need for extensive training and manual annotation in computer vision. By focusing on semantic segmentation, class-agnostic object counting, and animal re-identification, our work presents practical and progressive approaches to enhancing learning efficiency while preserving—and often improving—model performance.

Regarding semi-supervised semantic segmentation in Chapter 3, we successfully revitalized and adapted classical autoencoding techniques within a contemporary framework. This approach significantly leveraged unlabeled data to enhance segmentation accuracy, thereby reducing the dependency on extensive labeled datasets. Our primary contributions in this domain include: 1) the introduction of an innovative modification to the image reconstruction task that better disentangles object features from background noise, and 2) demonstrating substantial improvements in segmentation accuracy across various datasets, establishing a robust methodology for future research in semi-supervised learning.

Building on semi-supervised learning, in chapter 4, we presented a novel approach to class-agnostic object counting that operates without the need for training on specific object classes. By utilizing pre-training models and integrating semi-supervised prototypes selection, we established a new baseline for training-free, class-agnostic count-

ing. This method not only simplifies the deployment of counting technologies but also matches and sometimes exceeds the performance of traditional training-based methods. Our results show that this approach is highly effective and practical for real-world applications, where the types of objects to be counted can vary significantly.

Finally, Chapter 5 extends these principles to animal re-identification, a crucial task for ecological monitoring and conservation. We introduced a keypoint propagation mechanism that utilizes a pre-trained diffusion model to significantly reduce the need for manual annotation. This was further enhanced by novel implementations within the Vision Transformer architecture, such as Keypoint Positional Encoding and Categorical Keypoint Positional Embedding. These enhancements led to substantial improvements in the accuracy and robustness of re-identification systems across challenging wildlife datasets.

## 6.2 Future Work

We have presented some solutions to the issues mentioned earlier, however, these methods are not without limitations and there are still some unresolved issues.

### 6.2.1 Semi-Supervised Semantic Segmentation

We may improve our methods by:

- Enhanced Reconstruction Focus: Current methods primarily reconstruct foreground objects, performing well on datasets with a defined background class but less so on datasets without one. Future research could extend the reconstruction process to encompass all given classes, not just the foreground. This approach would likely improve model generalization across diverse datasets, including those where background elements are critical to accurate segmentation.

- Advanced Generative Models: Investigating the use of more sophisticated generative models could further refine the segmentation quality. These models could be trained to better understand the nuances between different classes, potentially using variational autoencoders or generative adversarial networks tailored to handle complex inter-class variations.

- Adaptive Segmentation Frameworks: Developing adaptive frameworks that can dynamically adjust their focus between foreground and background elements based on the dataset characteristics could provide another layer of refinement, making the segmentation process more flexible and broadly applicable.

### 6.2.2 Class-Agnostic Object Counting

- Integration with Real-time Learning: Implementing real-time learning mechanisms that can continuously update and refine counting algorithms as new data becomes available would be particularly useful in dynamic environments where object types and quantities frequently change.

- Utilization of SEEM Over SAM: Exploring the use of SEEM (Segment Everything Everywhere All at Once) instead of SAM (Segment Anything Model) could potentially yield richer semantic features, enhancing the system's ability to discern and count diverse objects more accurately.

- Integrating pretrained multi-modal models to enhance class-agnostic counting by leveraging cross-modal alignment, enabling more robust object localization and counting across diverse or ambiguous categories through semantic-aware visual-textual feature fusion.

### 6.2.3 Categorical Keypoint Positional Embedding

- Automated Keypoint Discovery: Our current method relies on manually annotated keypoints, which is labor-intensive and may not scale well. A future direction could focus on automated keypoint discovery using unsupervised or semi-supervised methods. One approach might involve sampling numerous random points across images and filtering these points based on stability and similarity of diffusion features, possibly using thresholding techniques or clustering algorithms to identify stable, distinctive keypoints without human intervention.

- Application to Larger, More Varied Datasets: Extending the current methodologies to larger and more varied datasets could test the scalability and robustness of the proposed solutions. Additionally, adapting these methods for real-time video processing could facilitate continuous monitoring and tracking of wildlife, a significant advancement for ecological studies and conservation efforts.

# Bibliography


[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.

[2] Radhakrishna Achanta and Sabine Susstrunk. Superpixels and polygons using simple non-iterative clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4651–4660, 2017.

[3] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3908–3916, 2015.

[4] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021.

[5] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. On the effectiveness of vit features as local semantic descriptors. In *European Conference on Computer Vision*, pages 39–55. Springer, 2022.

[6] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[7] Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 37–49. JMLR Workshop and Conference Proceedings, 2012.

[8] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019.

[9] Soren Bouma, Matthew DM Pawley, Krista Hupman, and Andrew Gilman. Individual common dolphin identification via metric embedding learning. In *2018 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE, 2018.

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[11] O. Chapelle, B. Scholkopf, and A. Zien, Eds. Semi-supervised learning (chapelle, o. et al., eds.; 2006) [book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

[12] Ted Cheeseman, Ken Southerland, Jinmo Park, Marilia Olio, Kiirsten Flynn, John Calambokidis, Lindsey Jones, Claire Garrigue, Astrid Frisch Jordán, Addison Howard, et al. Advanced image recognition: a fully automated, high-accuracy photo-identification matching system for humpback whales. *Mammalian Biology*, 102(3):915–929, 2022.

[13] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.

[14] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[15] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of*

*the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2016.

[16] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 801–818, 2018.

[17] Shuai Chen, Gerda Bortsova, Antonio García-Uceda Juárez, Gijs van Tulder, and Marleen de Bruijne. Multi-task attention-based semi-supervised learning for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 457–465. Springer, 2019.

[18] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023.

[19] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021.

[20] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[21] Jonathan P Crall, Charles V Stewart, Tanya Y Berger-Wolf, Daniel I Rubenstein, and Siva R Sundaresan. Hotspotter—patterned species instance recognition. In *2013 IEEE workshop on Applications of Computer Vision*, pages 230–237. IEEE, 2013.

[22] Ruining Deng, Can Cui, Quan Liu, Tianyuan Yao, Lucas W Remedios, Shunxing Bao, Bennett A Landman, Lee E Wheless, Lori A Coburn, Keith T Wilson,

et al. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging. *arXiv preprint arXiv:2304.04155*, 2023.

[23] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[24] Nikola Djukic, Alan Lukezic, Vitjan Zavrtanik, and Matej Kristan. A low-shot object counting network with iterative prototype adaptation. *arXiv preprint arXiv:2211.08217*, 2022.

[25] Huang et al. Point segment and count: A generalized framework for object counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[26] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

[27] Jiashuo Fan, Bin Gao, Huan Jin, and Lihui Jiang. Ucc: Uncertainty guided cross-head co-training for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9947–9956, 2022.

[28] L-CCGP Florian and Schroff Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 6, 2017.

[29] Alexander Freytag, Erik Rodner, Marcel Simon, Alexander Loos, Hjalmar S Kühl, and Joachim Denzler. Chimpanzee faces in the wild: Log-euclidean cnns for predicting identities and attributes of primates. In *Pattern Recognition: 38th German Conference, GCPR 2016, Hannover, Germany, September 12-15, 2016, Proceedings 38*, pages 51–63. Springer, 2016.

[30] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of*

*the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[31] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016.

[32] Stephen Gould, Jim Rodgers, David Cohen, Gal Elidan, and Daphne Koller. Multi-class segmentation with relative location prior. *International Journal of Computer Vision*, 80:300–316, 2008.

[33] Dayan Guan, Jiaxing Huang, Aoran Xiao, and Shijian Lu. Unbiased subclass regularization for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9968–9978, 2022.

[34] Xifeng Guo, Xinwang Liu, En Zhu, and Jianping Yin. Deep clustering with convolutional autoencoders. In *International Conference on Neural Information Processing*, pages 373–382. Springer, 2017.

[35] Kelly M Halloran, James D Murdoch, and Matthew S Becker. Applying computer-aided photo-identification to messy datasets: a case study of t hornicroft's giraffe (g iraffa camelopardalis thornicrofti). *African Journal of Ecology*, 53(2):147–155, 2015.

[36] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, pages 991–998. IEEE, 2011.

[37] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[38] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15013–15022, 2021.

[39] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. *Advances in Neural Information Processing Systems*, 36, 2024.

[40] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[41] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[42] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4145–4153, 2017.

[43] Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. *Advances in Neural Information Processing Systems*, 34:22106–22118, 2021.

[44] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 603–612, 2019.

[45] Minsung Hyun, Jisoo Jeong, and Nojun Kwak. Class-imbalanced semi-supervised learning. *arXiv preprint arXiv:2002.06815*, 2020.

[46] Bingliang Jiao, Lingqiao Liu, Liying Gao, Ruiqi Wu, Guosheng Lin, Peng Wang, and Yanning Zhang. Toward re-identifying any animal. *Advances in Neural Information Processing Systems*, 36, 2024.

[47] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep

neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4037–4058, 2020.

[48] Mingeun Kang, Kiwon Lee, Yong H Lee, and Changho Suh. Autoencoder-based graph construction for semi-supervised learning. In *European Conference on Computer Vision*, pages 500–517. Springer, 2020.

[49] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *European Conference on Computer Vision*, pages 429–445. Springer, 2020.

[50] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018.

[51] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[52] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[53] Dmitry A Konovalov, Suzanne Hillcoat, Genevieve Williams, R Alastair Birtles, Naomi Gardiner, and Matthew I Curnock. Individual minke whale recognition using deep learning convolutional neural networks. *Journal of Geoscience and Environment Protection*, 6:25–36, 2018.

[54] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11977–11986, 2019.

[55] Mayank Lahiri, Chayant Tantipathananandh, Rosemary Warungu, Daniel I Rubenstein, and Tanya Y Berger-Wolf. Biometric animal databases from field photographs: identification of individual zebra in the wild. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, pages 1–8, 2011.

[56] Lei Le, Andrew Patterson, and Martha White. Supervised autoencoders: Improving generalization performance with unsupervised regularizers. *Advances in Neural Information Processing Systems*, 31, 2018.

[57] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013.

[58] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*, 2023.

[59] He Li, Mang Ye, Cong Wang, and Bo Du. Pyramidal transformer with conv-patchify for person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7317–7326, 2022.

[60] Shuyuan Li, Jianguo Li, Hanlin Tang, Rui Qian, and Weiyao Lin. Atrw: a benchmark for amur tiger re-identification in the wild. *arXiv preprint arXiv:1906.05586*, 2019.

[61] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.

[62] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2898–2907, 2021.

[63] Zhengqin Li and Jiansheng Chen. Superpixel segmentation using linear spectral clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1356–1363, 2015.

[64] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of*

*the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2197–2206, 2015.

[65] Di Lin, Yuanfeng Ji, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Multi-scale context intertwining for semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 603–619, 2018.

[66] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3194–3203, 2016.

[67] Chang Liu, Yujie Zhong, Andrew Zisserman, and Weidi Xie. Countr: Transformer-based generalised visual counting. *arXiv preprint arXiv:2208.13721*, 2022.

[68] Yifan Liu, Bohan Zhuang, Chunhua Shen, Hao Chen, and Wei Yin. Auxiliary learning for deep multi-task learning. *arXiv preprint arXiv:1909.02214*, 2019.

[69] Yong-Jin Liu, Cheng-Chi Yu, Min-Jing Yu, and Ying He. Manifold slic: A fast method to compute content-sensitive superpixels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 651–659, 2016.

[70] Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. *arXiv preprint arXiv:2111.12903*, 2021.

[71] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[72] Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. Data-free knowledge distillation for deep neural networks. *arXiv preprint arXiv:1710.07535*, 2017.

[73] Erika Lu, Weidi Xie, and Andrew Zisserman. Class-agnostic counting. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth,*

*Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 669–684. Springer, 2019.

[74] Wei Luo, Jun Li, Jian Yang, Wei Xu, and Jian Zhang. Convolutional sparse autoencoders for image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 29(7):3289–3294, 2017.

[75] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023.

[76] Zhiheng Ma, Xiaopeng Hong, and Qinnan Shangguan. Can sam count anything? an empirical study on sam counting. *arXiv preprint arXiv:2304.10817*, 2023.

[77] Maximilian Matthé, Marco Sannolo, Kristopher Winiarski, Annemarieke Spitzen-van der Sluijs, Daniel Goedbloed, Sebastian Steinfartz, and Ulrich Stachow. Comparison of photo-matching algorithms commonly used for photographic capture–recapture studies. *Ecology and Evolution*, 7(15):5861–5872, 2017.

[78] Vincent Miele, Gaspard Dussert, Bruno Spataro, Simon Chamaillé-Jammes, Dominique Allainé, and Christophe Bonenfant. Revisiting animal photo-identification using deep metric learning and network analysis. *Methods in Ecology and Evolution*, 12(5):863–873, 2021.

[79] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2016.

[80] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2018.

[81] Olga Moskvyak, Frederic Maire, Feras Dayoub, Asia O Armstrong, and Mahsa Baktashmotlagh. Robust re-identification of manta rays from natural markings by learning pose invariant embeddings. In *2021 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2021.

[82] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25):E5716–E5725, 2018.

[83] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[84] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020.

[85] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2009.

[86] George Papandreou, Liang-Chieh Chen, Kevin Murphy, and Alan L Yuille. Weakly- and semi-supervised learning of a dcnn for semantic image segmentation. *arxiv:1502.02734*, 2015.

[87] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.

[88] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[89] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3394–3403, 2021.

[90] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. *Advances in Neural Information Processing Systems*, 28, 2015.

[91] Prashanth C Ravoor and TSB Sudarshan. Deep learning methods for multi-species animal re-identification and tracking–a survey. *Computer Science Review*, 38:100289, 2020.

[92] Ren and Malik. Learning a classification model for segmentation. In *Proceedings ninth IEEE International Conference on Computer Vision*, pages 10–17. IEEE, 2003.

[93] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[94] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[95] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[96] Olga Russakovsky, Li-Jia Li, and Li Fei-Fei. Best of both worlds: human-machine collaboration for object annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2121–2131, 2015.

[97] Stefan Schneider, Graham W Taylor, and Stefan C Kremer. Similarity learning networks for animal individual re-identification: an ecological perspective. *Mammalian Biology*, 102(3):899–914, 2022.

[98] Gail Schofield, Kostas Papafitsoros, Chloe Chapman, Akanksha Shah, Lucy Westover, Liam CD Dickson, and Kostas A Katselidis. More aggressive sea turtles win

fights over foraging resources independent of body size and years of presence. *Animal Behaviour*, 190:209–219, 2022.

[99] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[100] Qiuhong Shen, Xingyi Yang, and Xinchao Wang. Anything-3d: Towards single-view anything reconstruction in the wild. *arXiv preprint arXiv:2304.10261*, 2023.

[101] Min Shi, Hao Lu, Chen Feng, Chengxin Liu, and Zhiguo Cao. Represent, compare, and learn: A similarity-aware framework for class-agnostic counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9529–9538, 2022.

[102] Zenglin Shi, Ying Sun, and Mengmi Zhang. Training-free object counting with prompts. *arXiv preprint arXiv:2307.00038*, 2023.

[103] Risa Shinoda and Kaede Shiohara. Petface: A large-scale dataset and benchmark for animal identification. *arXiv preprint arXiv:2407.13555*, 2024.

[104] Risa Shinoda and Kaede Shiohara. Petface: A large-scale dataset and benchmark for animal identification. In *European Conference on Computer Vision*, pages 19–36. Springer, 2025.

[105] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

[106] Guang Shu, Afshin Dehghan, and Mubarak Shah. Improving an object detector and extracting regions using superpixels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3721–3727, 2013.

[107] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simpli-

fying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[108] Vladimir Somers, Alexandre Alahi, and Christophe De Vleeschouwer. Keypoint promptable re-identification. In *European Conference on Computer Vision*, pages 216–233. Springer, 2025.

[109] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[110] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30, 2017.

[111] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

[112] Joost E. van Engelen and Holger H. Hoos. A survey on semi-supervised learning. *Machine Learning*, 109:373–440, 2020.

[113] Vojtěch Čermák, Lukáš Picek, Lukáš Adam, and Kostas Papafitsoros. WildlifeDatasets: An Open-Source Toolkit for Animal Re-Identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5953–5963, January 2024.

[114] Maxime Vidal, Nathan Wolf, Beth Rosenberg, Bradley P Harris, and Alexander Mathis. Perspectives on individual animal identification from biology and computer vision. *Integrative and Comparative Biology*, 61(3):900–916, 2021.

[115] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(12), 2010.

[116] Le Wang, Rizhi Ding, Yuanhao Zhai, Qilin Zhang, Wei Tang, Nanning Zheng, and

Gang Hua. Giant panda identification. *IEEE Transactions on Image Processing*, 30:2837–2849, 2021.

[117] Tao Wang, Hong Liu, Pinhao Song, Tianyu Guo, and Wei Shi. Pose-guided feature disentangling for occluded person re-identification based on transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2540–2549, 2022.

[118] Teng Wang, Jinrui Zhang, Junjie Fei, Yixiao Ge, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, Shanshan Zhao, Ying Shan, et al. Caption anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677*, 2023.

[119] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4248–4257, 2022.

[120] Hendrik Weideman, Chuck Stewart, Jason Parham, Jason Holmberg, Kiirsten Flynn, John Calambokidis, D Barry Paul, Anka Bedetti, Michelle Henley, Frank Pope, et al. Extracting identifying contours for african elephants and humpback whales using a learned appearance model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1276–1285, 2020.

[121] Hendrik J Weideman, Zachary M Jablons, Jason Holmberg, Kiirsten Flynn, John Calambokidis, Reny B Tyson, Jason B Allen, Randall S Wells, Krista Hupman, Kim Urian, et al. Integral curvature representation and matching algorithms for identification of dolphins and whales. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2831–2839, 2017.

[122] Tiancheng Wen, Shenqi Lai, and Xueming Qian. Preparing lessons: Improve knowledge distillation with better supervision. *Neurocomputing*, 454:25–33, 2021.

[123] Claire L Witham. Automated face recognition of rhesus macaques. *Journal of Neuroscience Methods*, 300:157–165, 2018.

[124] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, pages 1–16. Springer, 2014.

[125] Hai-Ming Xu, Lingqiao Liu, Qiuchen Bian, and Zhen Yang. Semi-supervised semantic segmentation with prototype-based consistency regularization. *arXiv preprint arXiv:2210.04388*, 2022.

[126] Hai-Ming Xu, Lingqiao Liu, and Dong Gong. Semi-supervised learning via conditional rotation angle estimation. In *2021 Digital Image Computing: Techniques and Applications (DICTA)*, pages 01–08. IEEE, 2021.

[127] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3166–3173, 2013.

[128] Greg Yang, Zhilin Song, Michael King, and Wei Xu. Scaling laws for neural language models. *arXiv preprint arXiv:2301.12345*, 2023.

[129] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023.

[130] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[131] Shuo-Diao Yang, Hung-Ting Su, Winston H Hsu, and Wen-Chin Chen. Class-agnostic few-shot object counting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 870–878, 2021.

[132] Mang Ye, Shuoyi Chen, Chenyue Li, Wei-Shi Zheng, David Crandall, and Bo Du. Transformer for object re-identification: A survey. *arXiv preprint arXiv:2401.06960*, 2024.

[133] Zhiyuan You, Kai Yang, Wenhan Luo, Xin Lu, Lei Cui, and Xinyi Le. Few-shot object counting with similarity-aware feature enhancement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6315–6324, 2023.

[134] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[135] Jianlong Yuan, Yifan Liu, Chunhua Shen, Zhibin Wang, and Hao Li. A simple baseline for semi-supervised semantic segmentation with strong data augmentation. *arXiv preprint arXiv:2104.07256*, 2021.

[136] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.

[137] Guowen Zhang, Pingping Zhang, Jinqing Qi, and Huchuan Lu. Hat: Hierarchical aggregation transformers for person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 516–525, 2021.

[138] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2013.

[139] Tingting Zhang, Qijun Zhao, Cuo Da, Liyuan Zhou, Lei Li, and Suonan Jiancuo. Yakreid-103: A benchmark for yak re-identification. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021.

[140] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021.

[141] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017.

[142] Haowei Zhu, Wenjing Ke, Dong Li, Ji Liu, Lu Tian, and Yi Shan. Dual cross-attention learning for fine-grained visual categorization and object re-

identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4692–4702, 2022.

[143] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*, 2020.