A Comprehensive Survey on Image Labeling in Low-Label Environments: From Clustering to Vision-Language Models (2018–2025)

1. Introduction

In computer vision, the availability of labeled datasets is often a bottleneck for progress. Despite the abundance of raw image data, most of it remains unlabeled. Manually annotating millions of images is prohibitively expensive and time-consuming. For instance, the ImageNet project required over 2.5 years to label 14 million images; labeling 400 million images at that rate would take nearly 19 years of continuous labor. Moreover, domains like healthcare or finance pose legal and ethical challenges that make free data labeling impossible. These factors necessitate the exploration of alternative methods that reduce dependency on manual annotations, particularly for training high-performing deep learning models in image classification, segmentation, and detection.

2. Challenges in Learning from Unlabeled Image Data

- **Loss function design:** Without labels, common supervised losses like cross-entropy are inapplicable. Alternative strategies such as contrastive loss, clustering objectives, or consistency regularization must be employed. These are often less stable and harder to converge.

- **Performance gap:** Unsupervised models typically underperform supervised counterparts. For example, DeepCluster and MoCo reach only 40–60% Top-1 accuracy on ImageNet, compared to over 75% for supervised models.

- **Evaluation difficulties:** Without ground truth, internal metrics such as Silhouette Score or NMI are used but have weak correlation with real accuracy.

- **Noise propagation in pseudo-labeling:** Generating training targets from model predictions can lead to error amplification and overfitting if initial pseudo-labels are incorrect.

These limitations have spurred the emergence of new research lines: semi-supervised learning, self-supervised learning, clustering, active learning, and vision-language integration.

3. Major Directions in Image Labeling Techniques (2018–2025)

3.1. Clustering-Based Unsupervised Labeling

DeepCluster (Caron et al., 2018)

Overview: DeepCluster introduces an iterative approach to clustering image representations using K-Means and training convolutional neural networks (CNNs) using the resulting cluster

assignments as pseudo-labels. This method allows for feature learning without manual annotations by alternating between clustering features and training the CNN on those clusters.

Pipeline:

1. Extract image features using a CNN initialized randomly or pretrained.
2. Perform K-Means clustering on the extracted features to obtain cluster assignments.
3. Use these cluster assignments as pseudo-labels to train the CNN using supervised loss (e.g., cross-entropy).
4. Repeat the cycle: re-extract features with the updated CNN and re-cluster.

Applicability: DeepCluster is particularly suitable for large-scale datasets where no labels are available. However, it assumes the number of semantic clusters is roughly known and struggles with class imbalance or non-spherical cluster structures.

SCAN (Van Gansbeke et al., 2020)

Overview: SCAN builds upon self-supervised learning by enforcing semantic consistency among nearest neighbors in a learned feature space. Unlike DeepCluster, it separates the embedding learning stage from the clustering stage and avoids using K-Means, focusing instead on stability and consistency across local neighborhoods.

Pipeline:

1. Pretrain a CNN using self-supervised contrastive learning (e.g., SimCLR) to learn high-quality embeddings.
2. For each image, retrieve its k-nearest neighbors (kNN) in the feature space.
3. Add a clustering head and train it with a consistency loss that encourages the model to assign the same labels to each image and its neighbors across augmentations.

Applicability: SCAN is highly effective on small- to medium-scale datasets like CIFAR-10. It works well when reliable embeddings are available and is useful in tasks where semantic grouping is crucial, though its performance depends heavily on the quality of initial feature representations.

IIC (Invariant Information Clustering, Ji et al., 2019)

Overview: IIC formulates image clustering as a mutual information maximization problem between model predictions on original and augmented versions of the same image. It removes the need for an explicit clustering step by learning representations and cluster assignments in a fully differentiable, end-to-end manner.

Pipeline:

1. Generate paired inputs by applying data augmentations to each image.

2. Pass both original and augmented images through a shared CNN classifier to obtain soft cluster assignments.
3. Compute the joint probability matrix of class assignments and maximize the mutual information between the two distributions.

Applicability: IIC is ideal for unsupervised image classification and segmentation tasks. It excels in scenarios requiring end-to-end training and flexible class discovery. However, it demands well-crafted augmentations and may face optimization difficulties with complex or large label spaces.

3.2. Semi-Supervised Learning and Pseudo-Labeling

FixMatch (Sohn et al., 2020)

Overview: FixMatch is a simple yet powerful semi-supervised learning framework that combines pseudo-labeling with consistency regularization. It utilizes confident predictions on weakly augmented images to guide learning on strongly augmented counterparts, allowing the model to generalize from a small labeled set and a large pool of unlabeled images.

Pipeline:

1. Apply a weak augmentation to an unlabeled image and pass it through the model to obtain a predicted label.
2. If the model's confidence exceeds a threshold , retain the pseudo-label.
3. Apply a strong augmentation to the same image and train the model to match its prediction to the retained pseudo-label using cross-entropy loss.

Applicability: FixMatch is particularly suitable for low-label regimes such as medical imaging or small-scale educational datasets. It achieves near-supervised performance on datasets like CIFAR-10 with as few as 250 labeled samples. However, it is sensitive to the choice of the confidence threshold and the augmentation strategy.

Noisy Student (Xie et al., 2020)

Overview: Noisy Student is a two-stage semi-supervised learning framework that leverages a teacher-student paradigm. The teacher is trained on labeled data and used to generate pseudo-labels for unlabeled images, which are then used to train a larger, noise-robust student model.

Pipeline:

1. Train a teacher model on the available labeled dataset.
2. Use the teacher to generate pseudo-labels for a large unlabeled dataset.
3. Train a larger student model on both labeled and pseudo-labeled data using strong augmentations and noise injection.
4. Optionally repeat the process, using the student as the new teacher.

Applicability: Noisy Student is highly effective in scenarios with access to large unlabeled datasets and the computational resources to support multi-stage training. It set a new state-of-the-art on ImageNet by leveraging 300M unlabeled images. However, it requires a strong initial teacher model and significant training time.

3.3. Self-Supervised and Contrastive Learning

SimCLR (Chen et al., 2020)

Overview: SimCLR introduces a simple yet effective framework for contrastive learning of visual representations. It leverages augmentations and contrastive loss to push apart representations of different images while pulling together representations of augmented versions of the same image.

Pipeline:

1. Apply two different augmentations to each image to create a positive pair.
2. Pass each augmented image through a shared CNN encoder followed by a projection head (MLP).
3. Use a contrastive loss (NT-Xent) to maximize agreement between the positive pair while minimizing similarity with all other images in the batch.

Applicability: SimCLR is best suited for large-scale datasets with significant computing resources. It achieves state-of-the-art performance in representation learning but requires large batch sizes (≥4096) and substantial training time, making it more feasible in well-resourced research environments.

MoCo (He et al., 2020)

Overview: MoCo (Momentum Contrast) builds a dynamic dictionary of image features and uses momentum updates to maintain a consistent set of negative samples, addressing SimCLR's need for large batch sizes.

Pipeline:

1. Maintain two encoders: a query encoder (updated per batch) and a key encoder (updated via momentum).
2. Enqueue key features into a memory bank and sample negatives from it.
3. Train the model using contrastive loss between query-key pairs while leveraging the memory queue for stable and diverse negatives.

Applicability: MoCo is ideal for representation learning on datasets where compute is limited. Its memory-efficient design allows training with small batches, making it more practical for many use cases.

BYOL (Grill et al., 2020)

Overview: BYOL (Bootstrap Your Own Latent) is a self-supervised learning method that learns image representations without requiring negative samples. It uses two networks that interact in a bootstrapping manner to learn consistent representations.

Pipeline:

1. Use an online encoder to process an augmented view of an image.
2. Use a target encoder (updated via momentum) to process another augmented view.
3. Train the online network to predict the representation of the target encoder.

Applicability: BYOL avoids the collapse problem despite the absence of negative pairs. It's especially effective in limited-batch-size scenarios and achieves competitive results with minimal tuning.

SwAV (Caron et al., 2020)

Overview: SwAV (Swapped Assignments between Views) combines contrastive learning with clustering. It performs online clustering and swaps cluster assignments between views of the same image, enforcing consistency.

Pipeline:

1. Apply multiple augmentations to each image.
2. Assign soft cluster prototypes using online Sinkhorn-Knopp optimization.
3. Train the model to predict the cluster assignment of one view from another.

Applicability: SwAV integrates the benefits of contrastive and clustering-based learning. It scales well and performs robustly on various vision tasks with fewer augmentations and batch constraints.

3.4. Active Learning

Core-set (Sener & Savarese, 2018)

Overview: Core-set approaches active learning as a geometric coverage problem. It selects samples that best represent the full dataset by maximizing diversity and coverage.

Pipeline:

1. Represent data points in feature space using a pretrained model.
2. Solve a k-center problem to find the minimal set of points that best covers the entire distribution.
3. Query labels for these selected points and retrain the model.

Applicability: Core-set selection is effective when a limited labeling budget must be spent wisely. It is especially applicable in domains with high sample redundancy like facial recognition or surveillance.

VAAL (Sinha et al., 2019)

Overview: VAAL (Variational Adversarial Active Learning) combines a VAE with an adversarial discriminator to identify samples on the boundary between labeled and unlabeled distributions.

Pipeline:

1. Train a VAE on both labeled and unlabeled data to learn latent representations.
2. Use a discriminator to distinguish labeled vs. unlabeled representations.
3. Select samples that confuse the discriminator the most for labeling.

Applicability: VAAL performs well in tasks where latent uncertainty and representativeness are both critical, such as rare object detection or biomedical imaging.

3.5. Vision-Language Models

CLIP (Radford et al., 2021)

Overview: CLIP (Contrastive Language–Image Pretraining) learns to align image and text modalities by training on 400 million image–caption pairs. It enables zero-shot classification and performs well without fine-tuning.

Pipeline:

1. Encode images using a vision transformer or CNN.
2. Encode text using a transformer-based language model.
3. Train the model to bring corresponding image–text pairs closer in the embedding space via contrastive loss.

Applicability: CLIP is suitable for zero-shot and few-shot learning tasks, especially in open-world and low-resource environments. Its performance rivals fully supervised models on many benchmarks without task-specific data. However, training such models requires substantial compute, and outputs may reflect biases from internet-scale data.

4. Conclusion

Reducing the dependency on human-labeled data for image understanding has seen rapid advancements. Clustering, self-supervision, active learning, and vision-language modeling each offer powerful alternatives, with increasing convergence among them. While no single method is perfect, hybrid approaches that integrate multiple paradigms are likely to define the next generation of label-efficient computer vision systems. As models improve in generalization and multimodal reasoning, the role of manual labeling is expected to decline significantly.

References

[1] Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). Deep Clustering for Unsupervised Learning of Visual Features. *ECCV*. https://arxiv.org/pdf/1807.05520

[2] Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Gool, L.V., & Gool, L. V. (2020). SCAN: Learning to Classify Images without Labels. *ECCV*. https://arxiv.org/pdf/2005.12320

[3] Ji, X., Henriques, J.F., & Vedaldi, A. (2019). Invariant Information Clustering for Unsupervised Image Classification and Segmentation. *ICCV*. https://openaccess.thecvf.com/content_ICCV_2019/papers/Ji_Invariant_Information_Clustering_for_Unsupervised_Image_Classification_and_Segmentation_ICCV_2019_paper.pdf

[4] Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., & Raffel, C. (2020). FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *NeurIPS*. https://arxiv.org/pdf/2001.07685

[5] Xie, Q., Luong, M.T., Hovy, E., & Le, Q.V. (2020). Self-training with Noisy Student improves ImageNet classification. *CVPR*. https://arxiv.org/pdf/1911.04252

[6] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations. *ICML*. https://arxiv.org/pdf/2002.05709

[7] He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum Contrast for Unsupervised Visual Representation Learning. *CVPR*. https://arxiv.org/pdf/1911.05722

[8] Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., & Valko, M. (2020). Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. *NeurIPS*. https://arxiv.org/pdf/2006.07733

[9] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *NeurIPS*. https://arxiv.org/pdf/2006.09882

[10] Sener, O., & Savarese, S. (2018). Active Learning for Convolutional Neural Networks: A Core-Set Approach. *ICLR*. https://arxiv.org/pdf/1708.00489

[11] Sinha, S., Ebrahimi, S., & Darrell, T. (2019). Variational Adversarial Active Learning. *ICCV*. https://arxiv.org/pdf/1904.00370

[12] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *ICML*. https://arxiv.org/pdf/2103.00020