UHASSELT

KNOWLEDGE IN ACTION

Faculteit Bedrijfseconomische Wetenschappen

master handelsingenieur in de beleidsinformatica

Masterthesis

Omni-supervised learning

Sergej Jurev

Scriptie ingediend tot het behalen van de graad van master handelsingenieur in de beleidsinformatica

PROMOTOR :

dr. Frank VANHOENSHOVEN

BEGELEIDER :

Mevrouw Manal LAGHMOUCH

2021
2022

UHASSELT

KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT

KNOWLEDGE IN ACTION

Faculteit Bedrijfseconomische Wetenschappen

master handelsingenieur in de beleidsinformatica

Masterthesis

Omni-supervised learning

Sergej Jurev

Scriptie ingediend tot het behalen van de graad van master handelsingenieur in de beleidsinformatica

PROMOTOR :

dr. Frank VANHOENSHOVEN

BEGELEIDER :

Mevrouw Manal LAGHMOUCH

Omni-Supervised Learning

Sergej Jurev

Hasselt University, Agoralaan Building D, 3590 Diepenbeek, Belgium
sergej.jurev@student.uhasselt.be

Abstract. This paper reviews omni-supervised learning regimes in real-world settings and successfully constructs a data distillation pipeline for street segmentation on the Cityscapes dataset. During pipeline experimentation, a fully-supervised teacher sets a baseline performance of 0.8184 IoU for predicting pseudo-labels. Student models achieve a performance of 0.8378 IoU after training with pseudo-labels and 0.8474 IoU after implementing test-time data augmentations. A data transformation strategy consisting of random color adjustments was successfully applied, which was a largely untested methodology in omni-supervised learning literature. Model distillation and iterative student training was also attempted, but yielded no significant improvements in model performance.

Keywords: Omni-supervised · Distillation · Ensemble.

1 Introduction

Google's Open Images dataset [30] contains over 9 million annotated images and is one of the largest labeled datasets for machine learning. It can be used to train state-of-the-art supervised learning models for image classification, object detection, and visual relationship detection. Other large labeled datasets, such as Microsoft COCO [32] and ImageNet [9] have been used as labeled training data in numerous models. Kumar et al. claim that 5 exabytes of new raw data are produced every 24 hours worldwide, continually producing even larger sets for potential model training. However, these datasets are unlabeled and require either (manual) annotation or an adequate model designed to deal with absent labels during training. The growing quantities of organically generated data motivates research into specialized models that can compensate the absence of label annotations during parameter training. There have been many advancements on the front of so called semi-supervised learning algorithms [38,59], including self-training [42], co-training [1], multi-view learning [2] and graph-based models [58]. Semi-supervised models exploit both labeled and unlabeled datasets for training, allowing for a more versatile model training pipeline.

This paper investigates a special regime of semi-supervised learning proposed by Radosavovic et al. [39] in which the learning model exploits all available labeled data plus internet-scale sources of unlabeled data. Omni-supervised learning pipelines use distillation techniques to extract knowledge from models and from data in large-scale contexts where annotations are scarce or too costly to

produce. Medicine is a good example area where omni-supervised learning techniques can prove invaluable; the number of expert annotators who can annotate datasets such as brain scans is limited and the cost to hire them is high [21].

The literature review will analyze seminal papers on various omni-supervised learning pipelines, as well as real-world applications inspired by them. To discuss the implementation of omni-supervised learning pipelines in practice, a U-net teacher model will be trained for a challenging street segmentation task with a significant unlabeled set. This teacher model will generate pseudo-labels from derived unlabeled data, which the student will use in addition to the real labels during training. Thus, the student model trains on a larger set and attempts to outperform the teacher model. Important considerations for future research on omni-supervised learning are summarized in the results discussion.

2 Literature Review

2.1 Model Distillation

When high-performing neural networks are available for inference, it may be possible to harness the predictive power of multiple models in a single ensemble of their predictions. Dietterich et al. [11] argue that ensembles of multiple classification methods are able to perform better than a single classification method within the ensemble. They cite 3 reasons why this methodology can achieve good performance. First, individual classifiers with similar accuracy values may have opposing approximations to the ground truth. Averaging such approximations can compensate for errors and reduce the risk of selecting a single suboptimal classifier. Second, ensembles can help with local optimization traps. By averaging multiple classifiers stuck in different local optima, it may be possible to approximate a representation closer to the global optimum. Third, perfect parameters are often unattainable within the search space for hypotheses. Ensembles can expand the search space of representable functions and produce results closer to the individually unattainable optimum.

Caruana et al. [5] applied model ensemble techniques to thousands of models with robust ensemble selection criteria. The selection criteria included sampling models with replacement, initializing the highest performing models in the initial selection set, and bagging sampled models for the final ensemble. Caruana et al. were able to demonstrate the performance improvement of their selection criteria by outperforming other ensemble techniques on 10 different metrics with an accuracy of 0.956. Buciluǎ et al. [3] extended this research by using prediction ensembles to compress high-performance models that are too large and computationally expensive to deploy. Generated predictions are fed into faster and more compact student models as (additional) training examples to transfer knowledge from the teacher models and maintain high performance while reducing model complexity. This pipeline of ensembling strong teacher predictions can be applied to unlabeled or synthetic data, which are much easier to obtain than fully labeled examples.

Hinton et al. [20] further refine the ensembling of large-scale model predictions through knowledge distillation (also known as model distillation). They extend model compression methods with the following augmentations:

- The soft labels produced (from both labeled and unlabeled examples) are scaled to manipulate the distribution of generated class probabilities
- Confidences of teacher ensembles are not converted to hard labels and are instead treated as soft targets. The training loss of the student model is optimized by approximating the generated labels and, by extension, the knowledge of the teacher models
- For the loss function of the model, the weighted average of two different objective functions is used. The first target function reflects the soft label optimization of unlabeled examples, while the second target function implements real labels (if available) without scaling factors

Hinton et al. demonstrate the potential of model distillation by training 10 predictors for automatic speech recognition using a Hidden Markov Model [19] on a dataset of 2000 hours of spoken English data. The baseline performance of a single model is 58.9%. When the predictions of all 10 models are combined, an accuracy of 61.1% is achieved. After applying model distillation techniques, the final distilled single model performs with an accuracy of 60.8%, outperforming the base model by 2.01% and lagging behind the ensemble of teacher models by only 0.03%. According to the review by Gou et al. [15] the applications of model distillation have received much attention from the machine learning community in recent years.

2.2 Dataset Distillation

Wang et al. [52] propose dataset distillation as a variant of model distillation. Instead of models, large-scale , large datasets are compressed into a handful of synthetic images via dataset distillation. These images are derived based on the gradient descent steps of a network trained on real data. The goal is to generate images that provide similar weight updates to full batches of real data. The loss of a dataset distillation network takes into account the weight updates during backpropagation of the model and updates its own weights accordingly, generating more appropriate synthetic images in the process. Liu et al. [33] succeeded in applying dataset distillation techniques to unlabeled data in the omni-supervised context of facial expression recognition. The researchers trained a teacher model to structure unlabeled data into 7 groups corresponding to emotion classes and find the most reliable predictions for each class. These selected images are compressed into a single training sample using data distillation and used for training in the teacher model. Dataset distillation was investigated in a number of other problems [35,47,50] showing that it works in real-world scenarios.

2.3 Data Distillation

Radosavovic et al. [39] propose data distillation, a method that creates ensembles of teacher model predictions from multiple transformations (such as flipping

and scaling) of the unlabeled data, to automatically generate new training annotations for student model training. This method builds on the concept of model distillation, where ensembles are created from transformed input data rather than the output of model inference. Radosavovic et al. argue that the multi-transform inference steps in data distillation pipelines are faster than the model training steps in model distillation pipelines.

Radosavovic et al. claim that mainstream semi-supervised learning methods are likely to be upper bounded by fully supervised learning with all annotations, while omni-supervised learning is lower bounded by the accuracy of training on all annotated data. Radosavovic et al. succeeded in providing evidence of this claim by using omni-supervised learning to increase the performance of a model for detecting human keypoints by 2 points of average precision (AP) compared to a fully supervised baseline. These results show the potential of omni-supervised learning methods compared to the performance increase of approximately 3 AP by manually annotating a similar amount of examples [37]. Radosavovic et al. demonstrated the application of omni-supervised learning by training a Mask R-CNN model [16] for multi-person keypoint detection. ResNet-50/101 [18] and ResNeXt-101 [53] with Feature Pyramid Networks [31] were used as backbones for their experiments. Training data were used from the COCO dataset [32] (115k annotated human body symbols and 120k unlabeled images) and the Sports-1M dataset [27] (180k static frames). The following sections summarize the results of the pioneering omni-supervised learning experiments.

3 different cases were defined based on the training data used for data distillation:

- Small-scale: 35k labeled COCO images for supervised learning and 80k labeled COCO images for data distillation
- Large-scale with similar distribution: 115k labeled COCO images for supervised learning and 120k unlabeled COCO images for data distillation
- Large-scale with dissimilar distribution: 115k labeled COCO images for supervised learning and 180k Sports-1M frames for data distillation

The case with small-scale data simulates a setting where internet-scale data is not used; instead, a subset of the labeled dataset is used for data distillation to better understand the performance limits when unlabeled data is introduced into a model training pipeline. When trained on all 115k labeled COCO images (including the subset for data distillation), the ResNet-50 model achieved a performance of 65.1 AP, which Radosavovic et al. consider the upper bound for semi-supervised learning methods in this context. When trained only on the labeled 35k COCO split, the fully supervised ResNet-50 teacher model achieved an AP of 54.9 (-10.2 AP compared to the upper limit). After using the teacher model for data distillation on 80k unlabeled examples, the student model, trained on both the labeled and self-labeled data, reached an AP of 60.2 (-5.3 AP compared to the upper bound). With the first experiment, Radosavovic et al. succeeded in demonstrating an initial success of data distillation as a semi-supervised learning method in an artificial context.

Both large-scale cases cover real-world scenarios of omni-supervised learning where an annotated dataset is augmented with internet-scale quantities of unlabeled examples. These scenarios include the use of unlabeled datasets from both similar (COCO) and dissimilar (Sports-1M) distributions and the use of different backbone models with varying depth and capacity. A summary of the AP values for each large-scale experiment is provided in the following table:

Backbone	**Baseline AP**	**COCO AP**	**Sports-1M AP**
ResNet-50	65.1	67.1 (+2)	66.6 (+1.5)
ResNet-101	66.1	67.8 (+1.7)	67.5 (+1.4)
ResNeXt-101-32x4	66.8	68.7 (+1.9)	68.0 (+1.2)
ResNeXt-101-64x4	67.3	69.1(+1.8)	68.5 (+1.2)

The experiments of Radosavovic et al. show a consistent improvement between 1.2 and 1.9 AP from the use of omni-supervised learning compared to a fully supervised baseline. They claim that these results are non-trivial when contrasted with research that chooses to manually label similar amounts of additional data and that results in an improvement of 3 AP points [37]. This consistent increase in performance is also evidence of robustness in the use of omni-supervised learning; the large-scale experiments demonstrate the ability of producing real-world results with different data distributions and varying model complexity.

Consistent model improvements were also observed when omni-supervised learning was applied for COCO object detection using Faster R-CNN [40] with bounding box voting as the ensembling strategy. Specifically, an increase of nearly 1 AP point was realized in all experiments with omni-supervised learning compared to a fully supervised baseline, suggesting a more limited ability to improve object detection models with the proposed methods. However, consistent performance increase in different contexts shows that omni-supervised learning has the potential to improve model performance in multiple use cases.

Huang et al. [21] applied omni-supervised learning with 3D U-Net [6] architectures to detect the volumes, locations, and centers of 6 brain structures in 4044 3D brain scans of fetuses. 5% of the images (344) were annotated by medical experts; a time cost of 120 hours was attributed to this effort, with a theoretical time cost of 1251 hours for annotation of the entire dataset. By using omni-supervised learning on a partially labeled dataset, Huang et al. managed to achieve the following accuracy boost compared to a fully supervised baseline (mean performance of the baseline versus the omni-supervised learning model in parentheses):

- 7.2% lower error in predicting brain volume (22.8% vs. 15.6%)
- 4.9% higher 3D IoU in predicting brain structures (57.9% vs. 62.8%)
- 0.31 mm lower center deviation in predicting structure center coordinates (2.07mm vs. 1.76mm)

Huang et al. compare these results with their other publication on brain scan segmentation [22] in which they outperformed 3D U-Net architectures with new

fully supervised VP-Nets scoring an average 3D IoU of 62.0%. However, VP-Nets lag behind compared to the 62.8% 3D IoU score achieved by omni-supervised learning with ordinary 3D U-Nets trained on a similar dataset. The superior performance of omni-supervised learning compared to fully-supervised methods, combined with the motivation for low-cost alternatives to manual dataset annotation, provides another clear signal about the practical benefits of omni-supervised learning in medical applications.

The omni-supervised learning implementation of Huang et al. differed from the original authors in several important ways. First, the impact of different transformations was investigated more thoroughly. The omni-supervised learning models that used 7 rotations of 10 degrees (60.6% 3D IoU) performed better than models that used 7 translations of [-10, 0, 10] pixels in each orthogonal direction (60.9% 3D IoU). Huang et al. attribute the difference in performance to the orientational variations in brain scan imaging. These variations are also the reason why an implementation with R-CNN's sliding-window schemes was discouraged for this problem. Second, data distillation was used in conjunction with model distillation to generate unknown labels. Huang et al. trained 2 identical architectures with different loss functions to derive labels from the data transformations. These loss functions, binary cross-entropy and die similarity, focus on better optimization of volume estimation and IoU metrics respectively. Given the varied training conditions, the obtained models are expected to perform differently during the inference process. The performance of the retrained student model was improved to 61.8% 3D IoU when 2 teacher models were used in conjunction with data distillation. Third, the use of soft labels for automatic annotation was recommended over hard labels. Radosavovic et al. preferred the use of hard-labels to be consistent with manually labeled data. Huang et al. argued that soft-labels contain rich information that can be passed on to student models, agreeing with the findings of Hinton et al. [20].

Finally, it is important to note that Huang et al. draw similar conclusions when compared to the findings of Radosavovic et al. [39]. Both authors express a preference for retraining from scratch over fine-tuning teacher models due to the risk of local extrema traps. Both authors also note how the average model performance scales positively with the number of unlabeled examples introduced, with Huang et al. also seeing a decrease in the variance of the performance metrics with larger training sets.

Venturini et al. [49] criticize the lack of selection criteria for auto-labeled predictions in omni-supervised learning pipelines, arguing that uncertainty estimation could be an asset for improving student model performance. Previous omni-supervised learning publications either introduce auto-labeled data randomly [21] or with weak heuristics using detection thresholds and average detected instances [39]. Venturini et al. argue that omni-supervised learning requires data and model diversity, which are also methods to measure aleatoric and epistemic uncertainty [28] respectively.

The researchers propose the following hypotheses:

- Data volumes with lower aleatoric uncertainty correlate with clearer and more challenging data, leading to better segmentation performance
- Data volumes with lower epistemic uncertainty most closely resemble labeled training data and introduce bias based on fully supervised segmentations

The hypotheses were validated on ADNI MRI [25] (135 labeled volumes, 680 unlabeled volumes) and INTERGROWTH-21st ultrasound [36] (146 labeled volumes, 802 unlabeled volumes) datasets by measuring the Dice coefficient and the Hausdorff distance of student learning after random, aleatoric and epistemic selection of volume datasets. Omni-supervised learning with random data selection (0.700 ± 0.023 Dice) produced a statistically significant increase for the ultrasound dataset compared to a fully supervised baseline (0.673 ± 0.046 Dice). Lowest aleatoric uncertainty selection (0.727 ± 0.014 Dice) resulted in the highest significant performance increase for ultrasound images, while lowest epistemic uncertainty did not result in a statistically significant improvement (0.689 ± 0.040 Dice) over baseline.

The effects on the MRI dataset were smaller, and the only statistically significant increase from baseline (0.848 ± 0.015 Dice) was observed for aleatoric selection (0.851 ± 0.015 Dice). Venturi et al. attribute the limited performance gain in the MRI dataset to the high baseline performance and the amount of training examples. The researchers argue that the large contrast in performance gain in the ultrasound dataset may be related to the greater variation in ultrasound image quality; they believe that datasets with lower or more variable quality benefit most from aleatoric selection based on uncertainty. The study by Venturi et al. shows that low aleatoric uncertainty should be considered as a selection criterion for auto-labeling data with omni-supervised methods, especially when training on datasets with variable quality.

2.4 Multi-source learning

Ren et al. propose an alternative omni-supervised learning framework for object detection that simultaneously handles strong labels, various forms of partial annotations (tags, dots, and scribbles), and unlabeled data. Instead of using data distillation pipelines with unlabeled examples, the authors train models with mixed annotations by using region proposal refinement pipelines and pseudo-label generators with robust loss functions. Liu et al. [33] use a similar methodology to include weakly labeled and unlabeled data in their training pipeline by using focal loss to distill the soft pseudo-labels from unlabeled data. Ren et al.'s UFO2 managed to outperform Faster-RCNN [40] by an additional 4.5 AP points when training on the COCO-80 [32] data set. Their research shows that omni-supervised learning pipelines are not limited to data and model distillation and that weakly-supervised methods can be advantageous when deployed in an omni-supervised context. Ren et al. also argue the merit of budget-conscious labeling policies, where a fraction of the human labeling budget is spent on creating cost-efficient weak labels rather than devoting all available time on a limited amount of strong(er) labels. Because of the robustness of UFO2, it allows for

the distribution of the labeling budget across different labeling techniques. Ren et al. show that fully allocating labeling budget to strong annotation box labels yields lower accuracy (13.97 ± 0.98 AP) than reserving 20% of the budget to weaker class point labeling (14.11 ± 1.01 AP). This again points to the practical benefits of weak-supervised learning for omni-supervised learning pipelines.

Yang et al. [54,55] propose a multi-source omni-supervised learning framework that uses multiple datasets with varying label definitions. This scheme is achieved by adding multiple classification sub-models at the end of a network, which allow predictions to be passed to the corresponding label mapping. This approach allows the model to use a number of heterogeneous labels from different labeled data sources to learn robust feature representations. Individual datasets are further enriched by using unlabeled data via data distillation. Yang et al. [54] succeeded in improving the performance of the model with these robust techniques in a number of different problem areas, including panoramic street segmentation, traversable area detection, and nighttime street segmentation.

Duan et al. [13] propose webly-supervised learning pipelines to exploit the vast quantities of unlabeled data from the web. They achieve this by using teacher models to filter unlabeled examples and by transforming data from different sources into a similar data format. Duan et al. demonstrate the effectiveness of their OmniSource network on the Kinetics-400 dataset [4] by establishing a record performance of 83.6% Top-1 accuracy (previous state-of-the-art achieving 82.6% Top-1 accuracy.

3 Methods

3.1 Objective

The following experiments will measure the effectiveness of adding omni-supervised learning steps to fully supervised learning pipelines. Specifically, the impact of data distillation and model distillation will be analyzed during separate steps in the given pipeline. Best practices and suggestions from related publications are thoroughly considered during implementation.

A machine learning model will be trained to solve pixel-level instance segmentation tasks in urban street scenes. This application choice is justified by the following arguments:

- Understanding street labels does not require expert knowledge, especially when compared to medical applications such as brain segmentation where medical knowledge is paramount. An application context without major knowledge barriers allows focusing on machine learning methodology rather than understanding (and possibly exploiting) the application area
- The generation of accurate pixel-level segmentation maps is time consuming, which is a good real-world motivator for auto-labeling subsets of the training data
- Image segmentation masks can be treated as single color channel images. Thus, data enrichment techniques are consistent between training images and annotation masks

– Instance segmentation with a high number of classes is a difficult machine learning task, yielding fully supervised learning models with capped optimal performance. This leaves some margin for further performance improvements with techniques outside of a fully supervised regime

3.2 Training dataset

The Cityscapes Dataset for Semantic Urban Scene Understanding [7], a collection of 25000 German street scene frames, will be used in the machine learning pipeline. The authors state that this dataset is one of the largest and most diverse dataset of street scenes with 5000 high quality and 20000 coarse annotations spread across 30 different classes. Of the 5000 images with detailed annotations, a subset of 1525 annotations are kept private for competitions, leaving 3475 images for training and model validation.

For the purpose of the experiments, the 3475 high-quality annotations are treated as ground-truth labels, while the subset of 20000 coarse labels are ignored, with the corresponding images treated as unlabeled data from the same distribution. This setup can simulate a scenario with a label budget constraint where only 15% of all available data can be annotated. The labeled data is randomly shuffled and distributed among a training set of 2416, a validation set of 704, and a test set of 352. These are all multiples of a batch size of 16 and approximate a train-valid-test distribution of 70-20-10. To remain consistent with the labeled data, the unlabeled dataset will contain 19984 images, as 2 images were corrupted by the authors (preventing the final batch of 16 images from being used). The omni-supervised implementation of Wang et al. also [51]

3.3 Model architecture

The machine learning models will be based on the U-net [41] architecture widely used by the medical community for image segmentation [6,12,26,34,44,56]. In addition to the segmentation of biological structures, U-nets were successfully implemented in a number of different problem spaces [23,43,57]. A U-net consists of two intertwined architectures: an encoder network that attempts to encode input images into abstract feature representations and a decoder network that attempts to decode abstract representations into output segmentation maps. The largest layers of these architectures are connected with a bridge, while the shallow layers are connected with skip connections [1]. The double connections allow learning complex feature representations (thanks to the bridge) while avoiding vanishing or exploding gradients during backpropagation (thanks to the skip connections).

For the input layer, all Cityscape images (recorded at a resolution of 2048 by 1024 pixels) are reduced with nearest interpolation to a resolution of 512 by

[1] The bridge architecture consists of an encoder block, identical to a single block of the encoder network. Skipped connections are established by merging the block of the encoder and decoder network with corresponding filter sizes.

256 pixels. In the default setting of a U-net architecture, height and width of the inputs are equal and the resulting input image is square in shape. However, the Cityscapes images are recorded with an aspect ratio of 2:1, so a similar aspect ratio is used for the model input. U-net's ability to work with any given input size is used and it is assumed that doubling the number of input pixels will only benefit the model. In addition, the width and height of the input are defined as powers of 2, which fits well with the computer hardware. For the output layer, the number of output classes is kept at 20 to remain consistent with the Cityscapes competition rules. The labels of all omitted classes are set to a background label, leaving the following instance labels: road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky, person, rider, car, truck, bus, train, motorcycle, and bicycle.

Filters and batch sizes are defined based on available computing resources. Specifically, each encoder and decoder block contains 2 3x3 convolutions with filter sizes of 64, 128, 256, and 512. The filter size is increasing in the encoder network and decreasing in the decoder network. The 2 convolutions used in the bridge contain the largest filter with a size of 1024. This configuration results in 34,526,396 trainable parameters.

Following the implementation of Huang et al. [21] batch normalization [24] is applied after each convolutional layer to optimize the convergence rate and ReLU is applied afterwards for nonlinearity. Since ReLU is used as the preferred activation function, the filters of convolutional layers are initialized by sampling the following Gaussian distribution, as proposed by He et al. [17]:

$$\mu = 0, \sigma = \sqrt{n\_inputs}$$

(where n_inputs is the number of variables that pass through a given layer)

Datta et al. [8] points out that this filters initialization strategy works optimally with nonlinear activation functions in deep neural networks, which is applicable to a U-net architecture. For regularization, dropout [46] was applied at a rate of 0.50 after the largest convolutional block of the encoder network, decoder network, and bridge.

3.4 Loss and metrics

At the time of this publication, the Cityscape benchmark suite contained 281 model submissions. The Jaccard index (better known as IoU score) [14] was used as a measure of model accuracy. The highest scoring model achieved an average score of 86.5, while the top 100 and top 200 models had scores of 80.6 and 69.5, respectively. Unless otherwise noted, all index scores should be considered averages for a given training, validation, or testing series. For the optimization of teacher model loss, a generalization of the Jaccard index, the Tversky index [48], is used as the optimization metric. The Tversky index is defined as follows:

$$\frac{TP}{TP + \alpha * FN + \beta * FP}$$

where alpha and beta are scale parameters and TP, FN and FP are true positives, false positives and true negatives, respectively. The introduction of the alpha and beta parameters allows for the creation of different loss functions with minimal effort, facilitating model distillation pipelines. This generalization allows the importance of precision and recall to be verified during training. Setting alpha and beta equal to 1 yields the Jaccard index, while setting both to 1/2 yields the Sørensen-Dice [10,45] coefficient. All the above index scores are between 0 and 1, with higher values implying better accuracy of the model. The loss function can thus be defined as the inverse or negative Tversky index, where the model must maximize the index to minimize the loss. To obtain a score between 0 and 1 for both index and loss, the difference between 1 and the negative index is taken as the final loss function for training. The loss is optimized using Adam (Adaptive Moment Estimation) [29] with an initial learning rate of 0.0001.

3.5 Data and model distillation

Most previous omni-supervised learning applications use transformations such as flipping, translating, rotating, cropping and zooming. Radosavovic et al. [39] suggest that color augmentation is also a suitable approach for data distillation, which was chosen as the preferred technique for data distillation. In contrast to color augmentation, spatial transformations (other than flipping) come at the cost of losing information located at the edges of the image, such as corners being cut off after rotation or blurred images after zooming. In applications such as human key point annotation and segmentation of biological structures, the objects of interest are usually in the center of the image; thus, the risk of information loss in the image perimeter is minimal to non-existent. This is not the case for the Cityscapes dataset, where the labeled pixels are scattered throughout the entire image. Another argument for preferring color jittering to spatial transformations in this context has to do with the practical implications of this data distillation strategy. When transformations such as rotation and translation are applied to an input image, the same transformations must be applied to the corresponding labels and vice versa to the predictions. These steps are avoided when color jittering is used; the segmentation maps for all possible transformations with color jittering remain invariant and thus apply to the original image. Horizontal flipping, however, is a spatial transformation in which no pixels are lost due to padding.

The hypothesis is that data distillation with color jittering and flipping will result in improvements in omni-supervised learning pipelines. 5 types of data enrichment transformations are defined:

- Horizontal flipping with a probability of 50%
- Contrast adjustment by a factor
- Brightness adjustment
- Color saturation adjustment
- Sharpness adjustment (blurring and sharpening filter)

Each shift factor is sampled from a uniform distribution and each of the 4 color transformations is performed unless otherwise stated.

During the data distillation steps, 16 transform vectors of size 4 are sampled from a uniform distribution to determine the intensity of the 4 color jittering techniques, resulting in 16 magnified images for model inference. Model distillation is applied to each magnified image by averaging all predictions of 2 contrasting models on that image. The distilled prediction map is flipped according to the sampled flipping factor, then a single ensemble is created by averaging all 16 model ensembles.

As proposed by Venturini et al. [49] the sum of the variance per pixel between the transformations of each ensemble is calculated as a selection criterion. 13888 of the self-labeled images (equivalent to 400% of the labeled data) with the lowest aleatoric uncertainty are selected for student model training. During student training, 6 images from the auto-labeled selection are sampled and added to a batch of 10 images with real labels, according to the recommendations of Radosavovic et al. [39].

3.6 Pipeline summary

The complete experimental steps of the data and model distillation pipeline are defined as follows:

- Train a preliminary model for hyperparameter tuning
- Train 2 teacher models with varying Tversky loss parameters and continue to use the optimal parameters from the previous step
- Test 2 teacher models from the previous step with all possible combinations of model and data distillation.
- Select the best performing scenario from the previous step and ensemble (hard) data distillation prediction labels for the next step
- Train the first student model using the labeled dataset, unlabeled examples, and corresponding predictions from the previous step
- Test the first student model with and without data distillation
- Select the best performing scenario from the previous step and generate hard prediction labels for the next step
- Train a second student model using the labeled dataset, unlabeled examples, and corresponding predictions from the previous step
- Test the second student model with and without data distillation

4 Results

4.1 Preliminary testing

A preliminary model was trained to determine appropriate hyperparameters for teacher training and data distillation. The filter size of the model was limited to 25% for faster training feedback and more margin for improvement with distillation techniques. Tversky alpha and beta are both set to 0.5, so the model

optimizes for the Dice index. Data magnification is applied in accordance with the transformations used during data distillation, with a factor sampled evenly between 0.5 and 2. Table 1 summarizes the inference measurements on the test set under different regimes.

Distillation	Dice	IoU
None	0.8205 (0.0852)	0.7039 (0.1182)
D.D. [0.5, 2.0]	0.7808 (0.0807)	0.6468 (0.0977)
D.D. [1, 1.2]	0.8343 (0.8387)	0.7235 (0.1092)

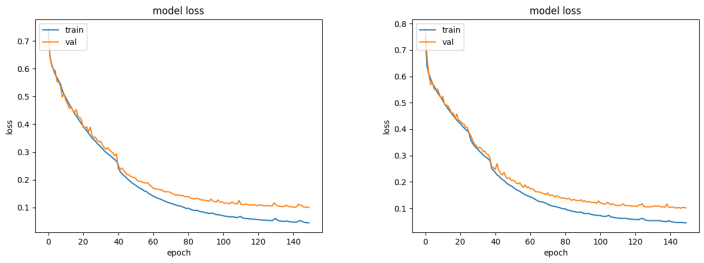
Table 1: Preliminary results yielded successful results when data distillation was performed with sampling interval [1, 1.2].

4.2 Teacher training

Following the insights from the preliminary tests, the following 2 teacher models are trained for model distillation:

- Model teacher$_{\text{precision}}$ with Tversky parameters alpha = 0.3 and beta = 0.7 (more weighting of false positives and optimization for better precision)
- Model teacher$_{\text{recall}}$ with Tversky parameters alpha = 0.7 and beta = 0.3 (false negatives weighted more heavily and optimized for better recall)

Figure 1 shows the teacher training and validation loss after 150 epochs. Table 2 summarizes model testing without data augmentation.



(a) Teacher$_{\text{Precision}}$ training (b) Teacher$_{\text{Recall}}$ training

Fig. 1: The loss of training and validation sets begins to diverge after 40 epochs.

4.3 Transformations and ensembles

5 multi-transform and/or multi-model inference scenarios are examined on the test set with the 2 trained teachers:

Teacher	Dice	IoU
$\text{Teacher}_{\text{precision}}$	0.8963 (0.069)	0.8184 (0.1001)
$\text{Teacher}_{\text{recall}}$	0.8932 (0.0729)	0.8138 (0.1034)

Table 2: $\text{Teacher}_{\text{precision}}$ achieves a superior score.

- Only model distillation
- Only data distillation using $\text{Teacher}_{\text{precision}}$
- Only data distillation using $\text{Teacher}_{\text{recall}}$
- Model distillation and data distillation and consistent color jitter sampling
- Model distillation and data distillation with inconsistent color jitter sampling

In the latter scenario, the color transformation is sampled from the uniform distribution [1,2, 1,5], as opposed to the uniform distribution [1, 1,2] that both teacher models use to supplement the data during training.

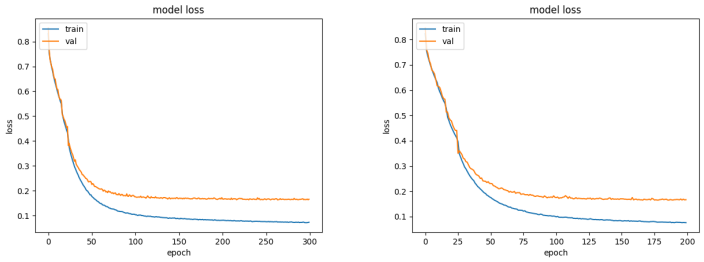
Table 3 summarizes the metric scores for inference during data and model enrichment with teacher predictors. The labeling strategy with the highest IoU score, namely data distillation with $\text{Teacher}_{\text{precision}}$, is used for annotating unlabeled examples. Using this strategy, 13888 ensembles with the lowest aleatoric uncertainty are selected for training student models.

Distillation	Dice	IoU
M.D	0.8948 (0.07)	0.8159 (0.1005)
D.D. $\text{Teacher}_{\text{precision}}$	0.9037 (0.0658)	0.8301 (0.0961)
D.D $\text{Teacher}_{\text{recall}}$	0.8997 (0.0708)	0.8241 (0.1009)
M.D + D.D. from [1, 1.2]	0.9017 (0.0677)	0.827 (0.0979)
M.D + D.D. from [1.2, 1.5]	0.8341 (0.0746)	0.7218 (0.1013)

Table 3: Data distillation with teacher precision and D.D. yields highest performance, model distillation yields inferior results

4.4 Student training

The first student_1 model is trained for 300 epochs (allowing the model enough time to converge and learn from pseudo-labeled examples as suggested by Radosavovic et al. [39]) with batches of 10 labeled examples and 6 pseudo-labeled examples. Labeled examples are passed through the model once per epoch, while pseudo-labeled examples are sampled each batch. After student_1 is trained, it becomes the teaching model for a second student_2 model by predicting a new set of pseudo-labeled examples. The second student model is trained on these new predictions for 200 epochs. Figure 2 gives an overview of the student training and table 4 gives an overview of the test metrics before and after the test-time data augmentation for both students.



(a) Student$_1$ training (b) Student$_2$ training

Fig. 2: Training and validation loss for student models begin to diverge after 25 epochs, despite additional regularization of variable batch samples

Model	Dice	IoU
Student$_1$	0.9081 (0.0687)	0.8378 (0.0994)
Student$_1$ D.D.	0.914 (0.0664)	0.8474 (0.0966)
Student$_2$	0.9062 (0.0707)	0.8349 (0.1016)
Student$_2$ D.D.	0.9124 (0.0668)	0.8449 (0.0978)

Table 4: Comparison of student performance. Data expansion during the test remains favorable for both student models. The difference in quality between the students is negligible.

5 Discussion

5.1 Data distillation performance

Performance improvements in the proposed pipeline sent clear signals regarding the potential of omni-supervised learning, specifically the success of data distillation with color jitter. A baseline segmentation performance of teacher$_{\text{Precision}}$ 0.8184 IoU (after test-time augmentations) corresponding to the best baseline teacher was outperformed by a student model trained on teacher predictions, with an IoU of 0.8474. Part of this performance improvement is attributed to the use of test-time augmentations rather than the exploitation of distilled pseudo-labels; the IoU of teacher and student models improved by 0.0117 and 0.0096, respectively, when transformational ensembles were exploited during label assignment. However, the student model consistently outperformed its teacher, even without test-time augmentations. These results are consistent with Radosavovic et al.'s [39] hypothesis that student accuracy is lower bounded by the performance of a teacher model.

5.2 Model distillation performance

Tests with model distillation did not yield conclusive results. Averaging model predictions reduced teacher$_{\text{Precision}}$ U-nets' performance by 0.0025, while teacher$_{\text{Recall}}$

U-nets' performance increased by 0.0021. Opposite minuscule shifts in score, despite the fact that teacher U-nets were trained with different loss functions and weight initializations, may motivate research toward more robust initialization strategies and parametric loss functions for model distillation. Nonetheless, training results indicate performance flattening when model complexity is increased to increase performance. Despite rigorous regularization techniques, such as dropout, noisy labels, and varying batches through sampling, validation and training loss diverge when performance reaches a level above 0.7 IoU. After the first 150 epochs, student validation loss remains almost twice as large as training loss, with decreasing movements in both losses during the last 150 epochs. These insights suggest that model optimization yields only small results when a sufficiently large performance is achieved. The contrasting success of data augmentation methods contributes to this hypothesis. When well-performing models are optimized, techniques such as data distillation have a lower risk of being ineffective because the model plateaus at high accuracy.

The first two reasons cited by Dietterich et al. [11] for why model ensembling can be beneficial do not seem to apply in this scenario:

- Ensembles of the two models do not appear to be contradictory, since the performance of the ensemble is between the individual performances of the models. This suggests that the ensemble is the average of similar predictions from a superior and inferior model
- Local optima traps are not expected with numerous trainable model weights and convergence around a consistent loss value

However, the plateau performance of models trained for multiple epochs suggests that an individual predictor cannot find the perfect parameters without some form of distillation. This reasoning is supported by the success of data distillation methods in exceeding the plateau performance of the teacher. The reasoning of Dietterich et al. for the success of model distillation in better approximating a space with perfect parameters can also be partially applied to data distillation.

5.3 Iterative student training performance

Training student_2 on the predictions of student_1 did not result in an IoU increase, despite the improvement in teacher quality. The superior accuracy of student_1 is likely attributed to the fact that he trained over 100 more epochs than student_2. Another possible explanation for the lack of IoU increase in the following student training iterations is the high performance of the teacher baseline. Student_1 achieves a stronger baseline due to data distillation, making it more difficult to further improve performance with similar techniques. It is also likely that the increase in baseline performance between teacher and student is not large enough to produce significantly better training examples for the next student. Radosavovic et al. [39] hypothesized that iterative student training may be beneficial for data distillation pipelines. In this context with strong baseline performance, the iterative nature of data distillation could not be confirmed.

5.4 Color jittering as a technique for data distillation

Preliminary increases in test time resulted in performance below baseline when color change factors between 0.5 and 2 were sampled, resulting in a performance drop of 0.0571 compared to an original baseline of 0.7039. Such large intervals were defined to produce more varied examples for data distillation, with the expectation that highly varied examples will produce superior ensembles. 0.5 and 2 were chosen as interval extrema for intuitive understanding of the sampling space; a factor of 0.5 results in a halving of a color feature, while a factor of 2 results in a doubling of that feature. The uniform sampling space would result in a tendency toward positive adjustments (sampling space between 1 and 2) rather than negative adjustments (sampling space between 0.5 and 1).

Despite the application of a similar strategy during training, the model did not appear to generalize well across highly diversified transformations. Applying a less extreme sampling interval does provide a performance increase during test-time augmentation. However, there is evidence that data distillation works better with subtle changes in the input image. This assumption is further validated during data distillation of teacher models, using a sampling interval of [1.2, 1.5] for test-time data augmentation. This strategy underperformed by 0.1052 IoU, despite being more conservative than the preliminary data augmentation approach. Future research on data distillation should take into account the choice of size of data transformation techniques. However, this paper confirms that color jittering can be used for both test-time augmentation and data distillation pipelines.

5.5 Further improvements to optimize performance on the Cityscapes dataset

Given the objective of this paper, which is the application of omni-supervised learning methods, the research time was mainly spent on deepening the knowledge in these methodologies rather than optimizing the prediction accuracy on the Cityscapes dataset. At the time of writing, the model would rank in the top 20 of the Cityscapes Benchmark Suite ranking with a prediction accuracy of 0.8474 IoU. However, the Cityscapes benchmarks differ from the design of the experiment by not calculating background labels (zero) in the IoU metric. For the purposes of this experiment, these classes were treated no differently than the 19 defined foreground labels. When optimizing for benchmark performance, it is recommended to train a model with 19 labels and ignore the background classification during loss optimization.

Another limitation of this research in producing optimal models for Cityscape benchmarks is the exclusion of coarse labels from the training pipeline. Introducing weakly labeled data pipelines in addition to distillation could increase the accuracy of the performance of pseudo-labelers. In this approach, teacher models convert coarse annotations to fine labels and student models use all 20000 coarsely labeled examples without filtering criteria or batch sampling.

In addition to using coarse labels, a test split is unnecessary when testing on the private Cityscapes test set. Thus, all available 3475 images can be used for training and validation. A thorough analysis of the model's performance on individual classes could indicate better loss definitions with measures of class imbalance.

5.6 Time and computation costs of omni-supervised learning

Radosavovic et al. [39] argue the advantage of data distillation over model distillation due to the lower time cost of inference compared to model training. In the proposed pipeline, a U-net architecture consisting of 34526396 trainable parameters was used. Both inference and training can be computationally intensive in this context and require important considerations before developing a model pipeline. To obtain the results in this paper, an NVIDIA RTX 3090 was used as the training GPU. This significantly accelerated the model training, but increased the complexity related to preparing training environments.

The model training time consists of approximately 800 milliseconds per training step. The teacher and student model epochs consist of 151 and 241 steps, respectively, resulting in a training time of roughly 2 to 3 minutes per epoch. A training of 200 epochs produces models after 6 to 10 hours of training. In contrast, data distillation pipelines produce annotations at an estimated rate of 2 annotations per second, requiring an estimated annotation time of up to 3 hours for data distillation on the unlabeled dataset Cityscapes and up to 5 minutes during test-time data enrichment. Results may vary significantly depending on the hardware used.

The implementation of soft labels for predictions of unlabeled data also proved infeasible. Storing labels on a hard disk takes 1 MB of space when the label consists of a logit and not a one-hot encoding. For 20000 labels, 20 GB must be reserved. For one-hot labels of size 20, that number increases to 400 GB. Implementations with soft labels are most advantageous when the number of classes is limited.

5.7 The state of omni-supervised learning

After reviewing the literature, a dichotomy of omni-supervised learning applications can be distinguished. One group of researchers of omni-supervised learning builds on the work of Radosavovic et al. and considers omni-supervised learning to be synonymous with (data) distillation in a large-scale context. The other group of researchers generalizes omni-supervised learning as exploiting internet-scale quantities of weakly labeled or unlabeled data, with or without distillation methodologies. The applications of the second group are more tied to their application context, while publications on general distillation techniques provide a good theoretical basis for other projects. Considering the number of publications available based on the research on data distillation by Radosavovic et al. [39] and model distillation of Hinton et al. [20], there is clear evidence that these methods can work well in real-world applications. There is a clear lack of

knowledge in the search for heuristics for the application of distillation methods. Future research should consider contributing to the development of guidelines for omni-supervised implementation, similar to the publication of Venturini et al. [49]

6 Conclusion

This research presents omni-supervised learning as a robust methodology for extracting knowledge from unlabeled datasets. Ensembling and distillation techniques have proven effective in real-world scenarios through numerous omni-supervised literature publications in medicine, object detection, speech recognition, and more. Some omni-supervised methodologies were also particularly effective in exploiting a mix of heterogeneous label types in their learning pipeline under a regime of multi-source learning. Models that employ data distillation often benefit from lower computational and time costs due to the speed of inference compared to training time.

This paper also demonstrated the effectiveness of color jittering as a data distillation strategy. A student model with a performance of 0.8184 IoU managed to outperform its teacher model with a performance of 0.8474 IoU by implementing multitransform inference and test-time data augmentation. Model distillation did not produce a performance increase in this context. Also, training a second learning model on the distilled labels of a previous learner did not yield any improvement. Extreme data distillation transformations are discouraged and may harm model performance when performed.

Further research related to omni-supervised learning should consider an expansion of heuristics that could constitute best practices for practitioners of omni-supervised learning. Data distillation allows for a number of different transformation augmentations, but due to a lack of selection criteria, they are often chosen at random. Researchers who explore omni-supervised learning in this capacity can provide a foundation for many other implementations.

References

1. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the eleventh annual conference on Computational learning theory. pp. 92–100 (1998)
2. Brefeld, U., Büscher, C., Scheffer, T.: Multi-view discriminative sequential learning. In: European Conference on Machine Learning. pp. 60–71. Springer (2005)
3. Buciluǎ, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 535–541 (2006)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
5. Caruana, R., Niculescu-Mizil, A., Crew, G., Ksikes, A.: Ensemble selection from libraries of models. In: Proceedings of the twenty-first international conference on Machine learning. p. 18 (2004)

6. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016)
7. Cordts, M., Omran, M., Ramos, S., Scharwächter, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset. In: CVPR Workshop on the Future of Datasets in Vision. vol. 2. sn (2015)
8. Datta, L.: A survey on activation functions and their relation with xavier and he normal initialization. arXiv preprint arXiv:2004.06632 (2020)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
10. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
11. Dietterich, T.G.: Ensemble methods in machine learning. In: International workshop on multiple classifier systems. pp. 1–15. Springer (2000)
12. Du, G., Cao, X., Liang, J., Chen, X., Zhan, Y.: Medical image segmentation based on u-net: A review. Journal of Imaging Science and Technology **64**, 1–12 (2020)
13. Duan, H., Zhao, Y., Xiong, Y., Liu, W., Lin, D.: Omni-sourced webly-supervised learning for video recognition. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. pp. 670–688. Springer (2020)
14. Everingham, M., Eslami, S., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**(1), 98–136 (2015)
15. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision **129**(6), 1789–1819 (2021)
16. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
17. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. pp. 1026–1034 (2015)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine **29**(6), 82–97 (2012)
20. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
21. Huang, R., Noble, J.A., Namburete, A.I.: Omni-supervised learning: scaling up to large unlabelled medical datasets. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 572–580. Springer (2018)
22. Huang, R., Xie, W., Noble, J.A.: Vp-nets: Efficient automatic localization of key brain structures in 3d fetal neurosonography. Medical image analysis **47**, 127–139 (2018)
23. Iglovikov, V., Shvets, A.: Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. arXiv preprint arXiv:1801.05746 (2018)

24. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456. PMLR (2015)
25. Jack Jr, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L. Whitwell, J., Ward, C., et al.: The alzheimer's disease neuroimaging initiative (adni): Mri methods. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine **27**(4), 685–691 (2008)
26. Kandel, M.E., He, Y.R., Lee, Y.J., Chen, T.H.Y., Sullivan, K.M., Aydin, O., Saif, M.T.A., Kong, H., Sobh, N., Popescu, G.: Phase imaging with computational specificity (pics) for measuring dry mass changes in sub-cellular compartments. Nature communications **11**(1), 1–10 (2020)
27. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (2014)
28. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? arXiv preprint arXiv:1703.04977 (2017)
29. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
30. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4. International Journal of Computer Vision **128**(7), 1956–1981 (2020)
31. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
32. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
33. Liu, P., Wei, Y., Meng, Z., Deng, W., Zhou, J.T., Yang, Y.: Omni-supervised facial expression recognition: A simple baseline. arXiv preprint arXiv:2005.08551 (2020)
34. Nazem, F., Ghasemi, F., Fassihi, A., Dehnavi, A.M.: 3d u-net: A voxel-based method in binding site prediction of protein structure. Journal of Bioinformatics and Computational Biology **19**(02), 2150006 (2021)
35. Nguyen, T., Novak, R., Xiao, L., Lee, J.: Dataset distillation with infinitely wide convolutional networks. Advances in Neural Information Processing Systems **34**, 5186–5198 (2021)
36. Papageorghiou, A.T., Ohuma, E.O., Altman, D.G., Todros, T., Ismail, L.C., Lambert, A., Jaffer, Y.A., Bertino, E., Gravett, M.G., Purwar, M., et al.: International standards for fetal growth based on serial ultrasound measurements: the fetal growth longitudinal study of the intergrowth-21st project. The Lancet **384**(9946), 869–879 (2014)
37. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multi-person pose estimation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4903–4911 (2017)
38. Prakash, V.J., Nithya, D.L.: A survey on semi-supervised learning techniques. arXiv preprint arXiv:1402.4645 (2014)
39. Radosavovic, I., Dollár, P., Girshick, R., Gkioxari, G., He, K.: Data distillation: Towards omni-supervised learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4119–4128 (2018)

40. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28**, 91–99 (2015)
41. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
42. Rosenberg, C., Hebert, M., Schneiderman, H.: Semi-supervised self-training of object detection models (2005)
43. Schonfeld, E., Schiele, B., Khoreva, A.: A u-net based discriminator for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8207–8216 (2020)
44. Siddique, N., Paheding, S., Elkin, C.P., Devabhaktuni, V.: U-net and its variants for medical image segmentation: A review of theory and applications. Ieee Access **9**, 82031–82057 (2021)
45. Sorensen, T.A.: A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. Biol. Skar. **5**, 1–34 (1948)
46. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)
47. Sucholutsky, I., Schonlau, M.: Soft-label dataset distillation and text dataset distillation. In: 2021 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2021)
48. Tversky, A.: Features of similarity. Psychological review **84**(4), 327 (1977)
49. Venturini, L., Papageorghiou, A.T., Noble, J.A., Namburete, A.I.: Uncertainty estimates as data selection criteria to boost omni-supervised learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 689–698. Springer (2020)
50. Wang, K., Zhao, B., Peng, X., Zhu, Z., Yang, S., Wang, S., Huang, G., Bilen, H., Wang, X., You, Y.: Cafe: Learning to condense dataset by aligning features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12196–12205 (2022)
51. Wang, P., Cai, Z., Yang, H., Swaminathan, G., Vasconcelos, N., Schiele, B., Soatto, S.: Omni-detr: Omni-supervised object detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9367–9376 (2022)
52. Wang, T., Zhu, J.Y., Torralba, A., Efros, A.A.: Dataset distillation. arXiv preprint arXiv:1811.10959 (2018)
53. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
54. Yang, K., Hu, X., Wang, K., Stiefelhagen, R.: In defense of multi-source omni-supervised efficient convnet for robust semantic segmentation in heterogeneous unseen domains. In: 2020 IEEE Intelligent Vehicles Symposium (IV). pp. 1386–1393. IEEE (2020)
55. Yang, K., Zhang, J., Reiß, S., Hu, X., Stiefelhagen, R.: Capturing omni-range context for omnidirectional segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1376–1386 (2021)
56. Yao, W., Zeng, Z., Lian, C., Tang, H.: Pixel-wise regression using u-net and its application on pansharpening. Neurocomputing **312**, 364–371 (2018)

57. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. IEEE Geoscience and Remote Sensing Letters **15**(5), 749–753 (2018)
58. Zhou, D., Huang, J., Schölkopf, B.: Learning from labeled and unlabeled data on a directed graph. In: Proceedings of the 22nd international conference on Machine learning. pp. 1036–1043 (2005)
59. Zhu, X.J.: Semi-supervised learning literature survey (2005)