# Appendix

# Implementation Details

## Hyperparameters

Table 2: Hyperparameter settings, which are inherited from the implementation by Oliver et al. (Oliver et al. 2018), where hyperparameters are tuned on a validation set. An Adam optimiser is deployed with the same learning rate decay schedule. The following hyperparameters are used for all experiments. Note: we adopt the *same* ramp-up function for all methods.

| **shared** | |
|---|---|
| training iterations | 500,000 |
| coefficient rampup from 0 util | 200,000 |
| learning decay factor | 0.2 |
| learning decay at iteration | 400,000 |
| **supervised baseline** | |
| initial learning rate | 0.003 |
| **pseudo-label** | |
| initial learning rate | 0.003 |
| max consistency coefficient | 1 |
| pseudo-label threshold | 0.95 |
| **VAT** | |
| initial learning rate | 0.003 |
| max consistency coefficient | 0.3 |
| VAT $\epsilon, \xi$ | 6.0, $10^{-6}$ |
| **$\Pi$-Model** | |
| initial learning rate | 0.0003 |
| **Mean-Teacher** | |
| initial learning rate | 0.0004 |
| max consistency coefficient | 8 |
| Exponential moving average decay | 0.95 |
| **SWA** | |
| initial learning rate | 0.001 |
| max consistency coefficient | 8 |
| weight averaging interval | 5,000 |
| **UASD** | |
| initial learning rate | 0.001 |
| max distillation coefficient | 1 |

## Data Augmentation

Table 3: Data augmentation. Note: ZCA image pre-processing is *only* applied on CIFAR10.

| dataset | gaussian noise $\sigma = 0.15$ | horizontal flip $p = 0.5$ | random translation $[-2, +2]$ |
|---|---|---|---|
| CIFAR10 | ✓ | ✓ | ✓ |
| CIFAR100 | × | ✓ | ✓ |
| TinyImageNet | × | ✓ | ✓ |

# Evaluation Protocols

Table 4: Evaluation protocols of SSL under class mismatch. $p\%$: Class distribution mismatch proportion among unlabelled data. $K$: number of known classes in labelled data. $L_{\text{num}}$: Labels per class.

| Dataset | $p\%$ | $K$ | $L_{\text{num}}$ | Labelled classes | Unlabelled classes |
|---|---|---|---|---|---|
| CIFAR10 | 0<br>25<br>50<br>75<br>100 | 400 | 6 | 2,3,4,5,6,7 | 4,5,6,7<br>0,5,6,7<br>0,1,6,7<br>0,1,8,7<br>0,1,8,9 |
| CIFAR100 | 50 | 50 | 100 | 0-50 | 25-75 |
| TinyImageNet | 50 | 100 | 100 | 0-100 | 50-150 |
| CIFAR100 + TinyImageNet | 86.5 | 100 | 100 | CIFAR100 | TinyImageNet |

# Tabular Results

## Evaluation on CIFAR10

Table 5: Evaluation under varying class distribution mismatch proportion on CIFAR10. Test error rates are reported at the point of lowest validation error. Results with reduction in error rate compared to supervised learning baseline are highlighted in **bold**. Best results are highlighted in **red**.

| | Class Distribution Mismatch Proportion | | | | |
|---|---|---|---|---|---|
| Method | 0% | 25% | 25% | 50% | 100% |
| baseline | | | 24.03 ± 0.75 | | |
| pseudo-label | **22.37 ± 0.66** | **24.02 ± 0.86** | 25.37 ± 0.86 | 26.11 ± 1.91 | 26.29 ± 0.71 |
| VAT | **20.63 ± 1.77** | **23.08 ± 0.49** | **23.78 ± 0.70** | 25.52 ± 0.84 | 26.23 ± 0.37 |
| Π-Model | **21.56 ± 1.29** | 24.80 ± 1.32 | 25.92 ± 1.61 | 26.43 ± 0.81 | 26.61 ± 0.79 |
| Temporal Ensembling | **21.93 ± 0.43** | 24.23 ± 0.96 | 25.66 ± 1.21 | 26.33 ± 0.56 | 27.00 ± 1.39 |
| Mean-Teacher | **21.68 ± 0.88** | 24.13 ± 1.22 | 24.79 ± 1.53 | 25.90 ± 1.00 | 26.78 ± 0.38 |
| SWA | **21.63 ± 0.38** | **23.31 ± 0.85** | **23.70 ± 0.61** | **23.90 ± 0.85** | 24.11 ± 0.65 |
| UASD (ours) | **20.59 ± 0.51** | **21.34 ± 0.52** | **21.88 ± 0.69** | **22.39 ± 0.48** | **22.49 ± 0.90** |

Table 6: Evaluation under varying class distribution mismatch proportion on CIFAR10. Test error rates are reported as the median of last 20 epochs. Results with reduction in error rate compared to supervised learning baseline are highlighted in **bold**. Best results are highlighted in **red**.

| | Class Distribution Mismatch Proportion | | | | |
|---|---|---|---|---|---|
| Method | 0% | 25% | 25% | 50% | 100% |
| baseline | | | 23.82 ± 0.61 | | |
| pseudo-label | **22.70 ± 0.42** | 24.42 ± 0.87 | 26.47 ± 1.01 | 27.82 ± 1.10 | 28.07 ± 1.01 |
| VAT | **23.07 ± 0.49** | 27.27 ± 1.36 | 27.45 ± 2.17 | 28.46 ± 2.62 | 28.79 ± 1.11 |
| Π-Model | **22.97 ± 0.46** | 26.48 ± 0.66 | 29.01 ± 2.67 | 28.19 ± 0.97 | 29.43 ± 1.88 |
| Temporal Ensembling | **22.45 ± 0.59** | 25.33 ± 0.81 | 26.94 ± 0.57 | 27.59 ± 0.62 | 28.16 ± 0.70 |
| Mean-Teacher | **22.09 ± 0.57** | 25.40 ± 0.41 | 26.46 ± 0.78 | 27.83 ± 1.43 | 29.09 ± 1.44 |
| SWA | **21.70 ± 0.34** | **23.36 ± 0.74** | 23.83 ± 0.61 | 24.15 ± 0.90 | 24.31 ± 0.55 |
| UASD (ours) | **20.55 ± 0.41** | **21.66 ± 0.71** | **22.01 ± 0.78** | **22.31 ± 0.65** | **22.63 ± 0.78** |

## Ablative Analysis

Table 7: Evaluation under varying class distribution mismatch proportion on CIFAR10. “size”: ensemble size. Test error rates are reported as the median of last 20 epochs. Results with reduction in error rate compared to supervised learning baseline are highlighted in **bold**. Best results are highlighted in **red**.

| | Class Distribution Mismatch Proportion | | | | |
|---|---|---|---|---|---|
| Method | 0% | 25% | 25% | 50% | 100% |
| baseline | | | 23.82 ± 0.61 | | |
| size 10+ | **21.12 ± 0.69** | **22.64 ± 0.77** | **23.41 ± 0.56** | 24.09 ± 0.72 | 24.68 ± 0.70 |
| size 100+ | **20.57 ± 0.39** | **21.40 ± 0.41** | **22.59 ± 0.49** | **22.44 ± 0.70** | **22.80 ± 0.55** |
| size 1000+ (ours) | **20.55 ± 0.41** | **21.66 ± 0.71** | **22.01 ± 0.78** | **22.31 ± 0.65** | **22.63 ± 0.78** |

Table 8: Evaluation under varying class distribution mismatch proportion on CIFAR10. “w/o both”: w/o soft distillation and w/o OOD filter. “w/o soft”: w/o soft distillation. “w/o OOD”: w/o OOD filter. Test error rates are reported as the median of last 20 epochs. Results with reduction in error rate compared to supervised learning baseline are highlighted in **bold**. Best results are highlighted in **red**.

| | Class Distribution Mismatch Proportion | | | | |
|---|---|---|---|---|---|
| Method | 0% | 25% | 25% | 50% | 100% |
| baseline | | | 23.82 ± 0.61 | | |
| w/o both | **23.50 ± 0.86** | 24.78 ± 0.64 | 25.78 ± 0.57 | 26.05 ± 0.79 | 27.43 ± 0.74 |
| w/o soft | **21.84 ± 0.53** | **23.27 ± 0.60** | 24.67 ± 0.60 | 24.67 ± 0.82 | 25.52 ± 0.92 |
| w/o OOD | **21.19 ± 0.31** | **22.34 ± 0.52** | **22.61 ± 0.99** | **22.68 ± 0.56** | **23.11 ± 0.70** |
| Full UASD (ours) | **20.55 ± 0.41** | **21.66 ± 0.71** | **22.01 ± 0.78** | **22.31 ± 0.65** | **22.63 ± 0.78** |

## Classes in CIFAR100 and TinyImageNet

Table 9: Classes in CIFAR100. Class labels which overlap with (i.e. same as or similar to) TinyImageNet are highlighted in **bold**. Number of same or similar labels: 19.

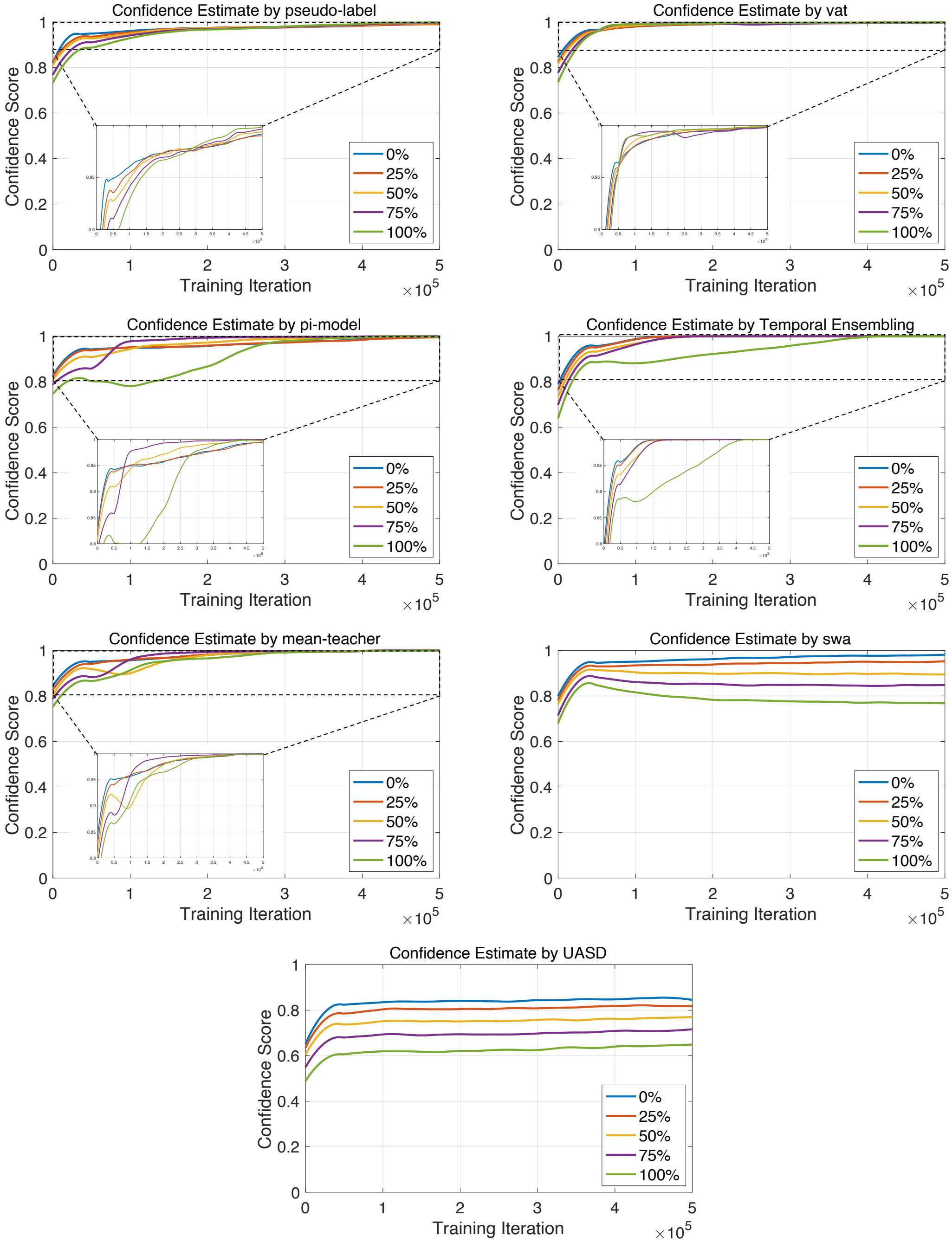
| Superclass | Class labels |
|---|---|
| aquatic mammals | beaver, dolphin, otter, seal, whale |
| fish | aquarium fish, flatfish, ray, shark, trout |
| flowers | orchids, poppies, roses, sunflowers, tulips |
| food containers | **bottles**, bowls, cans, cups, **plates** |
| fruit and vegetables | apples, **mushrooms**, **oranges**, pears, sweet peppers |
| household electrical devices | clock, **computer keyboard**, **lamp**, telephone, television |
| household furniture | bed, **chair**, couch, **table**, wardrobe |
| insects | **bee**, beetle, **butterfly**, caterpillar, **cockroach** |
| large carnivores | **bear**, leopard, **lion**, tiger, wolf |
| large man-made outdoor things | **bridge**, castle, house, road, skyscraper |
| large natural outdoor scenes | cloud, forest, mountain, plain, sea |
| large omnivores and herbivores | **camel**, cattle, **chimpanzee**, **elephant**, kangaroo |
| medium-sized mammals | fox, porcupine, possum, raccoon, skunk |
| non-insect invertebrates | crab, **lobster**, **snail**, spider, worm |
| people | baby, boy, girl, man, woman |
| reptiles | crocodile, dinosaur, lizard, snake, turtle |
| small mammals | hamster, mouse, rabbit, shrew, squirrel |
| trees | maple, oak, palm, pine, willow |
| vehicles 1 | bicycle, **bus**, motorcycle, pickup truck, **train** |
| vehicles 2 | lawn-mower, rocket, streetcar, tank, **tractor** |

Table 10: Classes in TinyImageNet. Class labels which overlap with (i.e. same as or similar to) CIFAR100 are highlighted in **bold**. Number of same or similar labels: 27. Class distribution mismatch proportion compared to CIFAR100: $(200 - 17)/200 = \mathbf{86.5}\%$.

| ID | Class labels |
|---|---|
| 1-5 | Egyptian cat, reel, volleyball, **rocking chair**, lemon |
| 6-10 | bullfrog, basketball, cliff, espresso, plunger |
| 11-15 | parking meter, German shepherd, **dining table**, monarch, **brown bear** |
| 16-20 | school bus, pizza, guinea pig, umbrella, organ |
| 21-25 | oboe, maypole, goldfish, potpie, hourglass |
| 26-30 | seashore, **computer keyboard**, **Arabian camel**, ice cream, nail |
| 31-35 | space heater, cardigan, baboon, **snail**, coral reef |
| 25-30 | albatross, **spider web**, sea cucumber, backpack, Labrador retriever |
| 36-40 | pretzel, king penguin, **sulphur butterfly**, tarantula, lesser panda |
| 46-50 | **pop bottle**, banana, sock, **cockroach**, projectile |
| 51-55 | **beer bottle**, mantis, freight car, guacamole, remote control |
| 56-60 | European fire salamander, lakeside, **chimpanzee**, pay-phone, fur coat |
| 61-65 | alp, **lampshade**, torch, abacus, moving van |
| 66-70 | barrel, tabby, goose, koala, **bullet train** |
| 71-75 | CD player, teapot, birdhouse, gazelle, academic gown |
| 76-80 | **tractor**, ladybug, miniskirt, golden retriever, triumphal arch |
| 81-85 | cannon, neck brace, sombrero, gasmask, candle |
| 86-90 | desk, frying pan, **bee**, dam, **spiny lobster** |
| 91-95 | police van, iPod, punching bag, beacon, jellyfish |
| 96-100 | wok, potter's wheel, sandal, **pill bottle**, butcher shop |
| 101-105 | slug, hog, cougar, crane, vestment |
| 106-110 | dragonfly, cash machine, **mushroom**, jinrikisha, water tower |
| 111-115 | chest, snorkel, sunglasses, fly, limousine |
| 116-120 | black stork, dugong, sports car, water jug, **suspension bridge** |
| 121-125 | ox, ice lolly, turnstile, Christmas stocking, broom |
| 126-130 | scorpion, wooden spoon, picket fence, rugby ball, sewing machine |
| 131-135 | **steel arch bridge**, Persian cat, refrigerator, barn, apron |
| 136-140 | Yorkshire terrier, swimming trunks, stopwatch, lawn mower, thatch |
| 141-145 | fountain, black widow, bikini, **plate**, teddy |
| 146-150 | barbershop, confectionery, beach wagon, scoreboard, **orange** |
| 151-155 | flagpole, **American lobster**, **trolleybus**, drumstick, dumbbell |
| 156-160 | brass, bow tie, convertible, bighorn, orangutan |
| 161-165 | American alligator, centipede, syringe, go-kart, brain coral |
| 166-170 | sea slug, cliff dwelling, mashed potato, viaduct, military uniform |
| 171-175 | pomegranate, chain, kimono, comic book, trilobite |
| 176-180 | bison, pole, boa constrictor, poncho, bathtub |
| 181-185 | grasshopper, walking stick, Chihuahua, tailed frog, **lion** |
| 186-190 | altar, obelisk, beaker, bell pepper, bannister |
| 191-195 | bucket, magnetic compass, meat loaf, gondola, standard poodle |
| 196-200 | acorn, lifeboat, binoculars, cauliflower, **African elephant** |

# Confidence Score Estimated by Different SSL Methods

Figure 8: Average confidence score (i.e. maximum class probability) on unlabelled data, estimated by different teaching signals *during training* under varying class distribution mismatch proportion (i.e. 0, 25, 50, 75, 100%) on CIFAR10. Most SSL methods are prone to produce ***overconfident*** teaching signals, regardless of the underlying unlabelled class distribution. This hinders the possibility to be aware of uncertainty, and blindly reinforces the overconfident wrong class assignments on those irrelevant unlabelled samples. In contrast, UASD produces soft teaching signals that encode higher uncertainty, and exhibits different levels of confidence score that are clearly stratified to reflect the underlying class distribution mismatch proportions.


Confidence Estimate by pseudo-label
Confidence Estimate by vat
Confidence Estimate by pi-model
Confidence Estimate by Temporal Ensembling
Confidence Estimate by mean-teacher
Confidence Estimate by swa
Confidence Estimate by UASD
Confidence Score
Training Iteration
×10^5