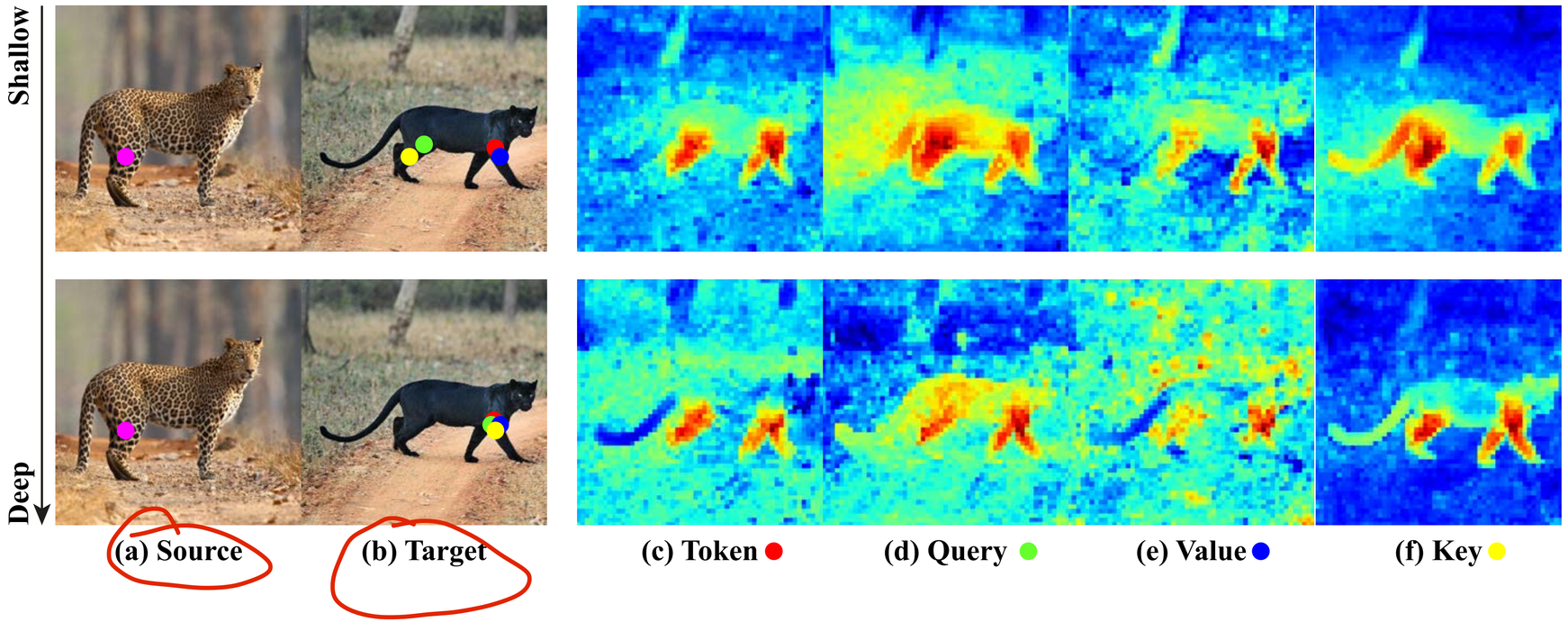
Unsupervised Learning: Downstream Applications

What do DINO features encode?

- DINO (and other SSL methods) provides semantic correspondence (almost) out-of-the-box.

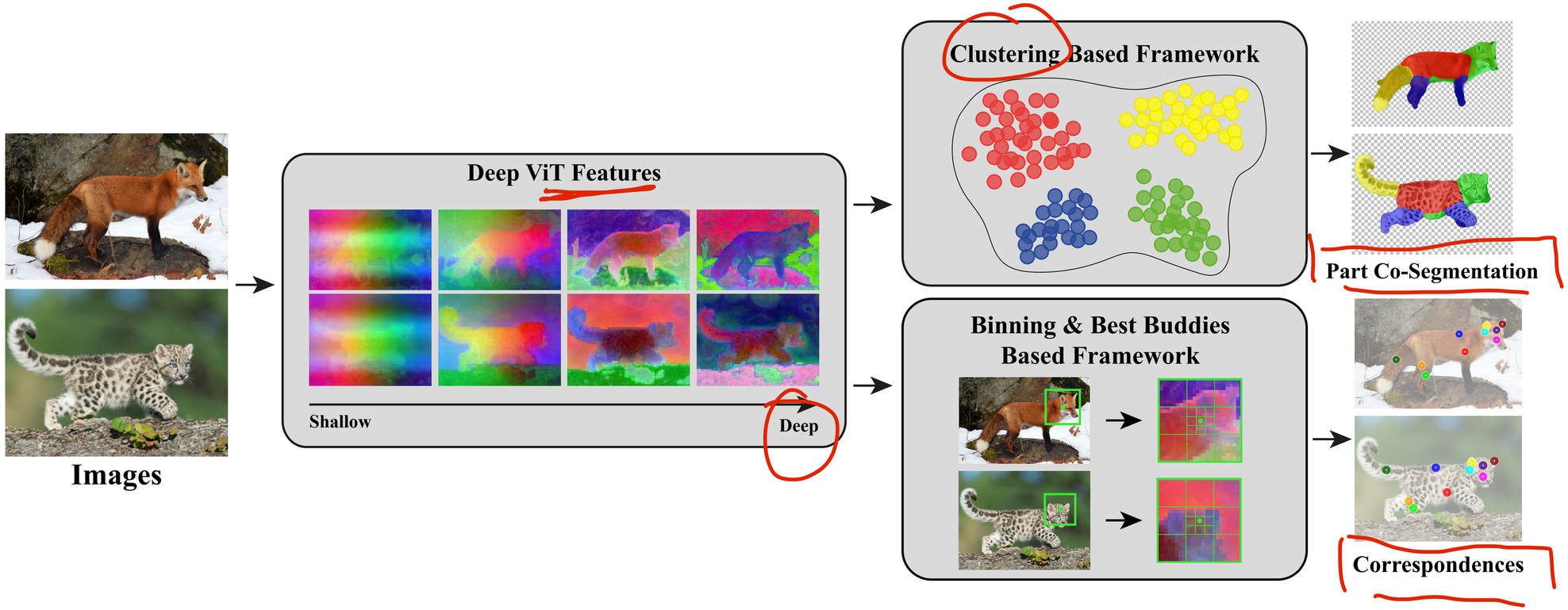
Matching source (magenta) to target image.



Amir et al., “Deep ViT Features as Dense Visual Descriptors ” (2021).

Part co-segmentation

- DINO (and other SSL methods) provide it (almost) out-of-the-box:



Amir et al., “Deep ViT Features as Dense Visual Descriptors ” (2021).

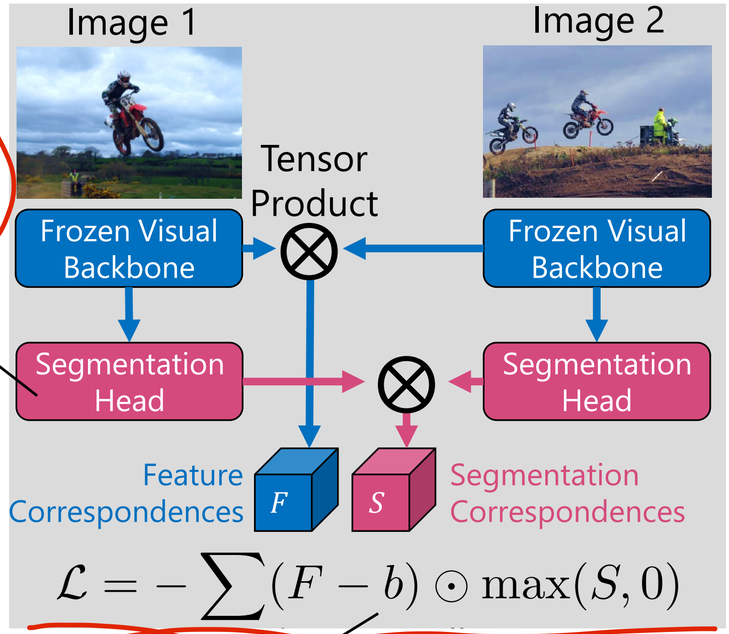
Semantic segmentation

We can also cluster “background” (stuff) areas, hence obtain semantic segmentation.

Hamilton et al., “Unsupervised Semantic Segmentation by Distilling Feature Correspondences” (ICLR 2022)

Semantic segmentation

High-level idea: Learn a lower-dimensional embedding, $S(f(i), f(j))$, such that clustering in this space yields semantic masks.



F is the original (e.g., DINO) pixel-level cosine similarity;

S is the pixel similarity of the learned (new) embedding.

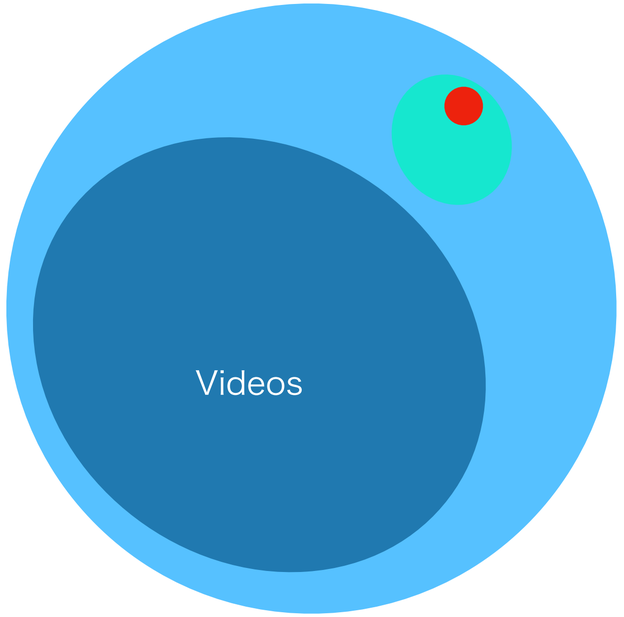
Learning S to mimic F (the original DINO embedding) amplifies the correlation patterns.

What is the effect of $b \neq 0$ (QUIZ)?

Hamilton et al., "Unsupervised Semantic Segmentation by Distilling Feature Correspondences" (ICLR 2022)

Self-supervision in videos

- What about videos?
- Two groups of problems:
  - Given a video, segment objects in the video (using motion cues).
  - Given a video dataset, learn to track objects.

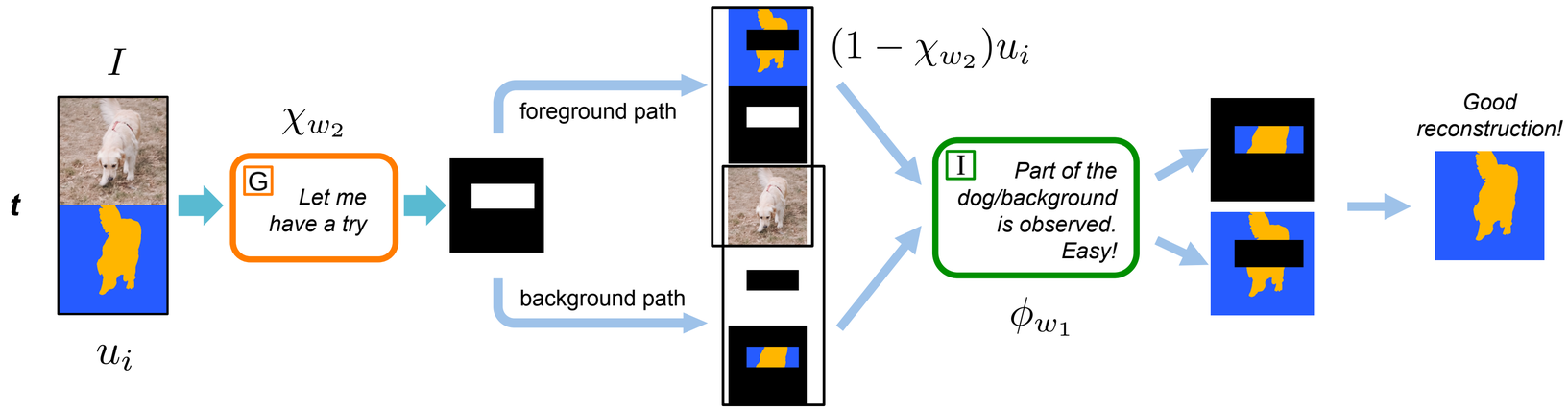


Segmentation from motion cues

Idea: If the object mask is correct, we cannot reconstruct object-related optical flow. We train two networks:

- Network G: Given an image and optical flow, predict object mask (foreground/background).
- Network I: Given a masked optical flow and the image (not masked), reconstruct the original optical flow.

Consider case A:



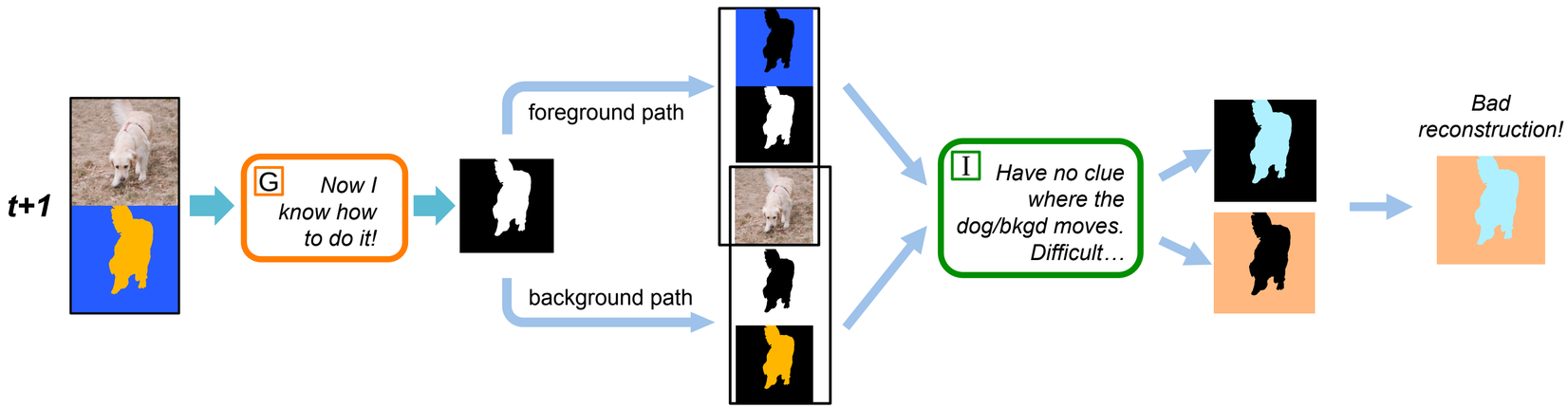
Yang et al., “Unsupervised Moving Object Detection via Contextual Information Separation” (2019).

Segmentation from motion cues

Idea: If the object mask is correct, we cannot reconstruct object-related optical flow. We train two networks:

- Network G: Given an image and optical flow, predict object mask (foreground/background).
- Network I: Given a masked optical flow and the image (not masked), reconstruct the original optical flow.

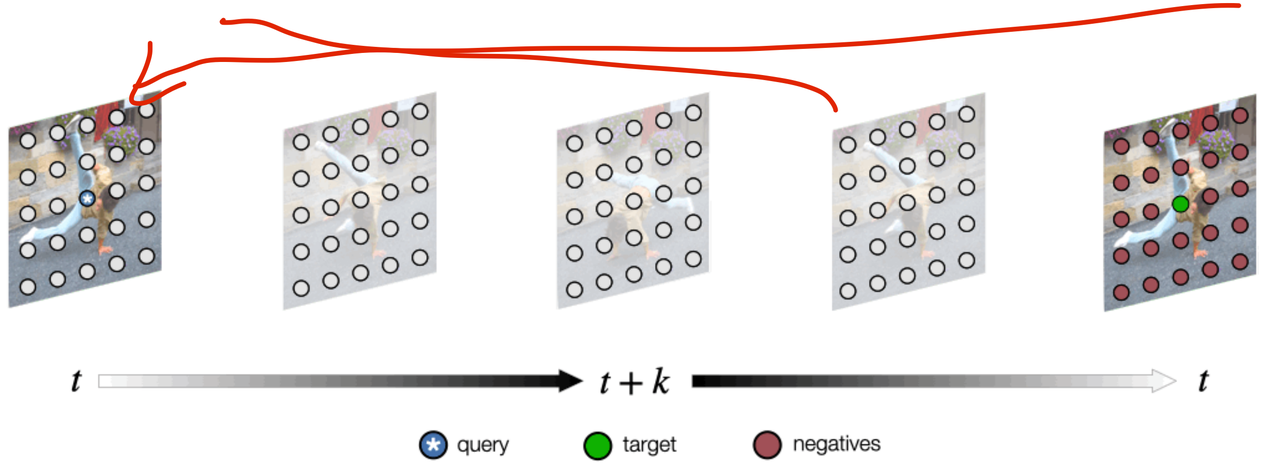
Consider case B:



Yang et al., “Unsupervised Moving Object Detection via Contextual Information Separation” (2019).

Contrastive random walk

- Forward-backward cycle consistency:
  - Given a video, construct a palindrome (i.e. $t_1, ..., t_{N-1}, t_N, t_{N-1}, ..., t_1$)
  - Label each patch in the image and propagate them through a video.
    - We compute affinity (using cosine similarity) between patches of subsequent frames.
  - Cycle consistency loss: Each label should arrive at its original location.

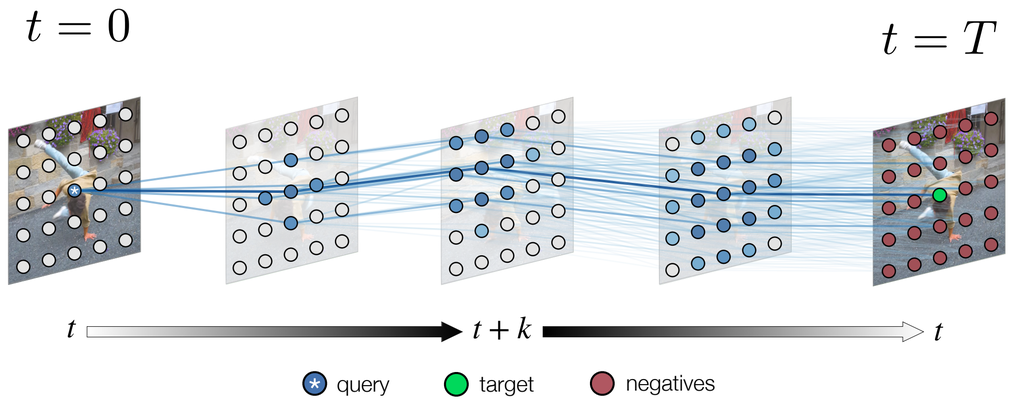


Jabri et al., “Space-Time Correspondence as a Contrastive Random Walk” (2021).

Contrastive random walk

- Let's formalise this:
  - Compute affinity between t and t+1:

$$A_{t:t+1} = F_t F_{t+1}^T, \quad F_t \in \mathbb{R}^{N \times d}$$



  - Here, F_t is our learned representation.
  - Let's assign each patch in t = 0 a unique one-hot label, $L \in \mathbb{R}^{N \times N}$.
  - We can propagate those labels forward, then backward:

$$L_{t+1} = \text{softmax}(A, 1) L_t$$

  - We can use cross-entropy on L_T, since we know the initial labels.

Contrastive random walk

- Dense tracking with the trained features:

Jabri et al., “Space-Time Correspondence as a Contrastive Random Walk” (2021).

References and additional reading

The list goes on...

- Chen et al., “A Simple Framework for Contrastive Learning of Visual Representations” (2020).
- Wei et al., “Masked Feature Prediction for Self-Supervised Visual Pre-Training” (2022).
- DINO: Caron et al., “Emerging Properties in Self-Supervised Vision Transformers” (2021).
- Caron et al., “Location-Aware Self-Supervised Transformers” (2023).
- Jabri et al., “Space-time correspondence as a contrastive random walk” (2020).
- Araslanov et al., “Dense unsupervised learning for video segmentation” (2021).

and many more...

Conclusion

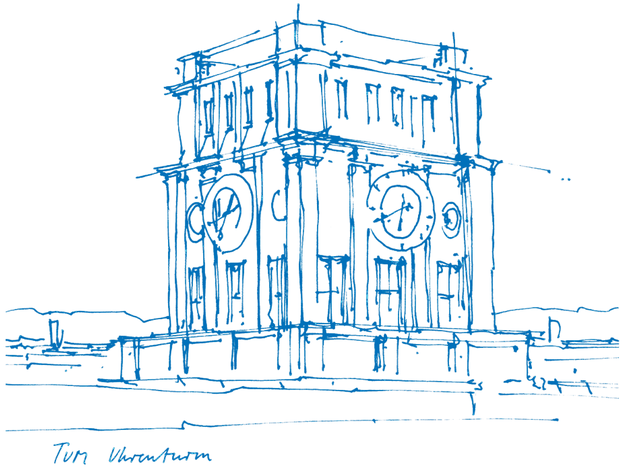
- Unsupervised learning dominates research landscape
  - We can train more accurate models with less supervision.
- Requires large computational resources (dozens of high-end GPUs).
- Yet do not scale (too) well with the amount of data (saturation).
- Many open questions:
  - What is a good proxy task?
  - How to make computational requirements manageable?
  - How (and/or why) does it work?

Computer Vision III:

Semi-supervised learning

Dr. Nikita Araslanov

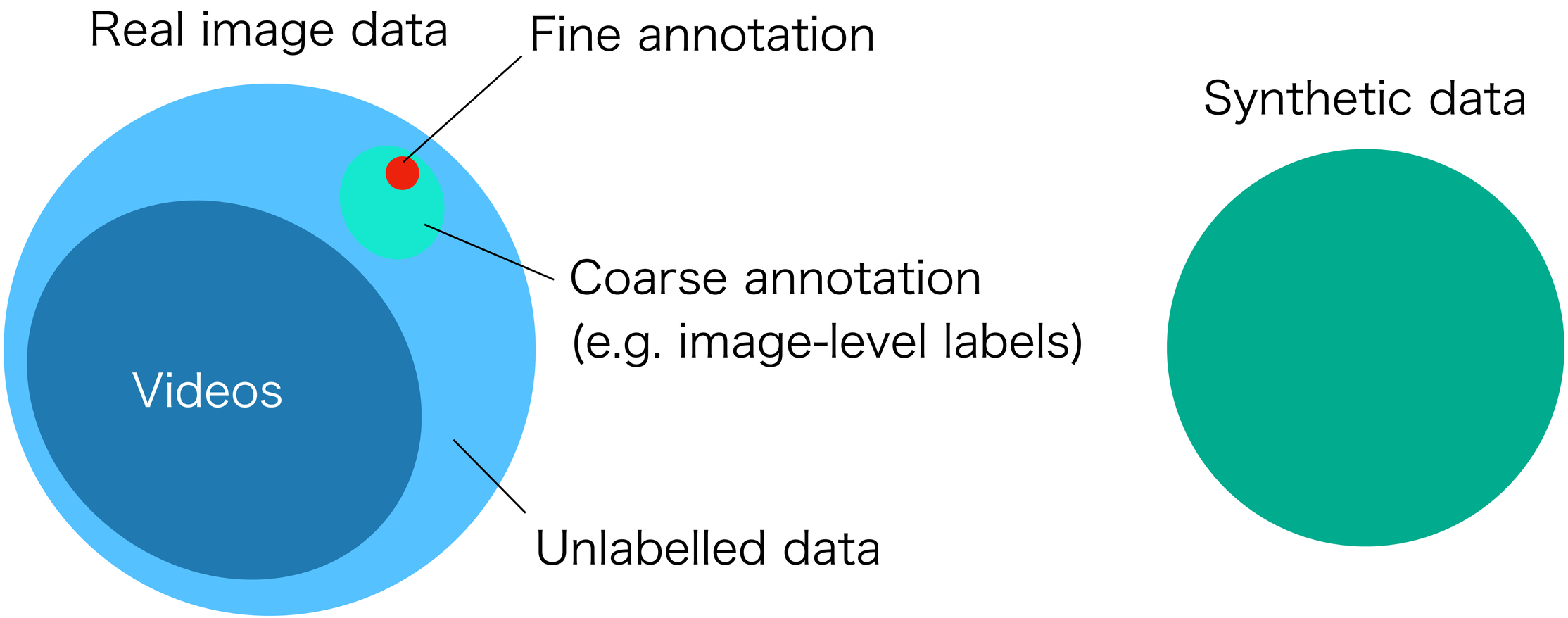
16.01.2024



Course progress

1. Introduction
2. Object detection 1
3. Object detection 2
4. Single object tracking
5. Multiple object tracking
6. Semantic segmentation
7. Instance & panoptic segmentation
8. Video object segmentation
9. Transformers
10. Unsupervised DST
11. **Semi-supervised DST** ← Today

Limited supervision



Semi-supervised learning: making use of both labelled and unlabelled data.

General remarks

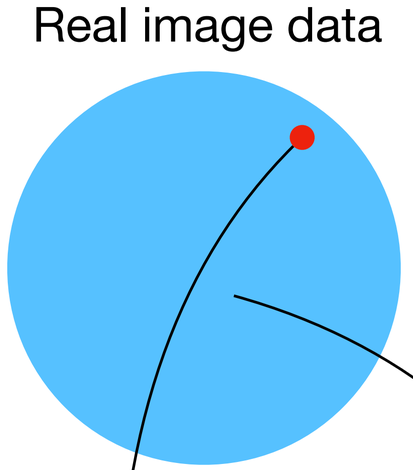
- Using both labelled and unlabelled data is a very practical scenario.
- If the goal is to get the best accuracy, semi-supervised learning is the way to go.
  - Current state-of-the-art frameworks take this approach (rather than full supervision).

Small print:

1. Improvement is not always guaranteed.
   - It depends on the model, the technique used and the unlabelled data.
2. Semi-supervised techniques are often complementary.
   - A combination of multiple techniques yield the best results (though make the framework more complex).

Semi-supervised loss

A practical perspective:



Notation

Labelled data: $\{(x_i, y_i)\}_i$

Unlabelled data: $\{(\hat{x}_i)\}_i$

$$\mathcal{L}(\{(x_i, y_i)\}_i, \{\hat{x}_i\}_i) = \sum_i \mathcal{L}_{\text{supervised}}(x_i, y_i) + \lambda \sum_i \mathcal{L}_{\text{unsupervised}}(\hat{x}_i)$$

Assumptions

Assumptions about semi-supervised learning:

1. Smoothness assumption
2. Low-density assumption
3. Manifold assumption

van Engelen and Hoos, “A survey on semi-supervised learning” (2020)

Assumptions

Assumptions about semi-supervised learning:

1. **Smoothness assumption**

- If two input points are close by, their labels should be the same.
- Transitivity:
  - We have a labelled $x_1 \in X_L$ and two unlabelled $x_2, x_3 \in X_U$ inputs.
  - Suppose x_1 is close to x_2, and x_2 is close to x_3, but x_1 is not close to x_3 .
  - Then we can still expect x_3 to have the same label as x_1 .

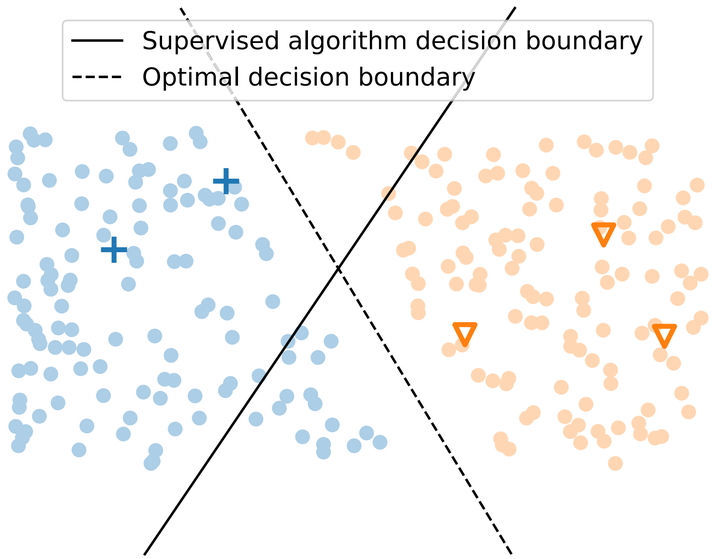
van Engelen and Hoos, “A survey on semi-supervised learning” (2020)

Assumptions

Assumptions about semi-supervised learning:

1. Smoothness assumption
2. **Low-density assumption**

   The decision boundary should pass through a region with low density $p(x)$.

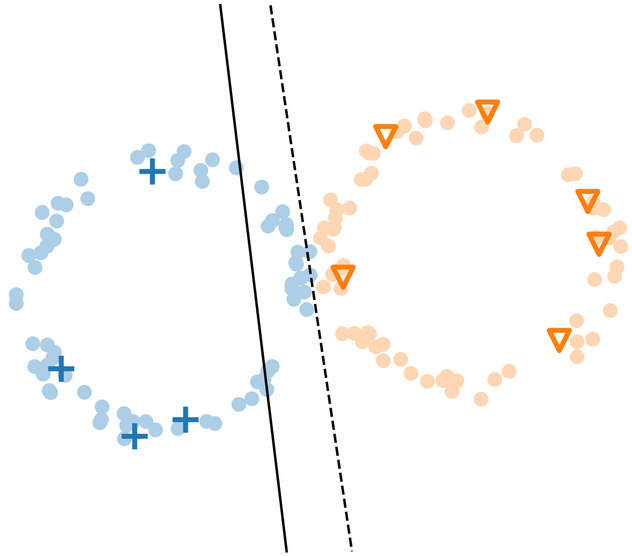


van Engelen and Hoos, “A survey on semi-supervised learning” (2020)

Assumptions

Assumptions about semi-supervised learning:

1. Smoothness assumption
2. Low-density assumption
3. **Manifold assumption**



- Data comes from multiple low-dimensional manifolds.
- Data points sharing the same manifold, share the same label.

van Engelen and Hoos, “A survey on semi-supervised learning” (2020)

Assumptions

Assumptions about semi-supervised learning:

1. Smoothness assumption
2. Low-density assumption
3. Manifold assumption

Remark: Which assumptions to make depends on what we know about how our data distribution $p(x)$ interacts with the class posterior $p(y \mid x)$.

van Engelen and Hoos, “A survey on semi-supervised learning” (2020)

Taxonomy

- Unsupervised pre-processing, e.g.:
  - pre-training, clustering, etc.;
- Wrapper methods, e.g.:
  - self-training;
- Intrinsically semi-supervised, e.g.:
  - entropy minimisation;
  - virtual adversarial networks.

- Learning from synthetic data:
  - domain alignment;
  - consistency regularisation.

Inductive: Given a labelled and an unlabelled dataset, produce a classifier (operating on any input point).

Transductive: Given a labelled and an unlabelled dataset, produce labels for the unlabelled dataset.

van Engelen and Hoos, “A survey on semi-supervised learning” (2020)

Taxonomy

- **Unsupervised pre-processing, e.g.**:
  - pre-training, clustering, etc.;
- Wrapper methods, e.g.:
  - self-training;
- Intrinsically semi-supervised, e.g.:
  - entropy minimisation;
  - virtual adversarial networks.

- Learning from synthetic data:
  - domain alignment;
  - consistency regularisation.

Inductive: Given a labelled and an unlabelled dataset, produce a classifier (operating on any input point).

Transductive: Given a labelled and an unlabelled dataset, produce labels for the unlabelled dataset.

van Engelen and Hoos, “A survey on semi-supervised learning” (2020)

Unsupervised pre-processing

- We actually already know this – from previous lecture.
- Two stages:
  - Unsupervised: feature extraction/learning (e.g. DINO, autoencoders).
  - Supervised: fine-tuning, linear probing, or k-NN classification.

Taxonomy

- Unsupervised pre-processing, e.g.:
  - pre-training, clustering, etc.;
- **Wrapper methods, e.g.:**
  - self-training;
- Intrinsically semi-supervised, e.g.:
  - entropy minimisation;
  - virtual adversarial networks.

- Learning from synthetic data:
  - domain alignment;
  - consistency regularisation.

Inductive: Given a labelled and an unlabelled dataset, produce a classifier (operating on any input point).

Transductive: Given a labelled and an unlabelled dataset, produce labels for the unlabelled dataset.

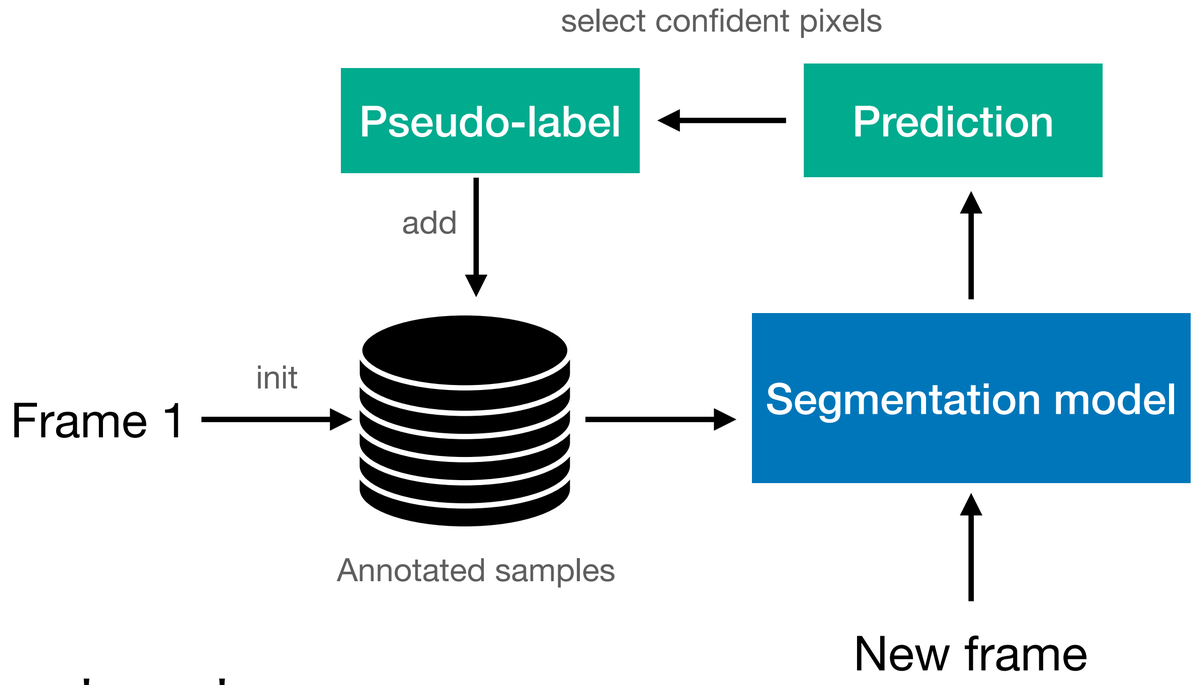
van Engelen and Hoos, “A survey on semi-supervised learning” (2020)

Self-training

- A single classifier trained jointly on labelled and self-labelled data from the unlabelled set
- We have already seen this (QUIZ).

OnAVOS: Online Adaptation

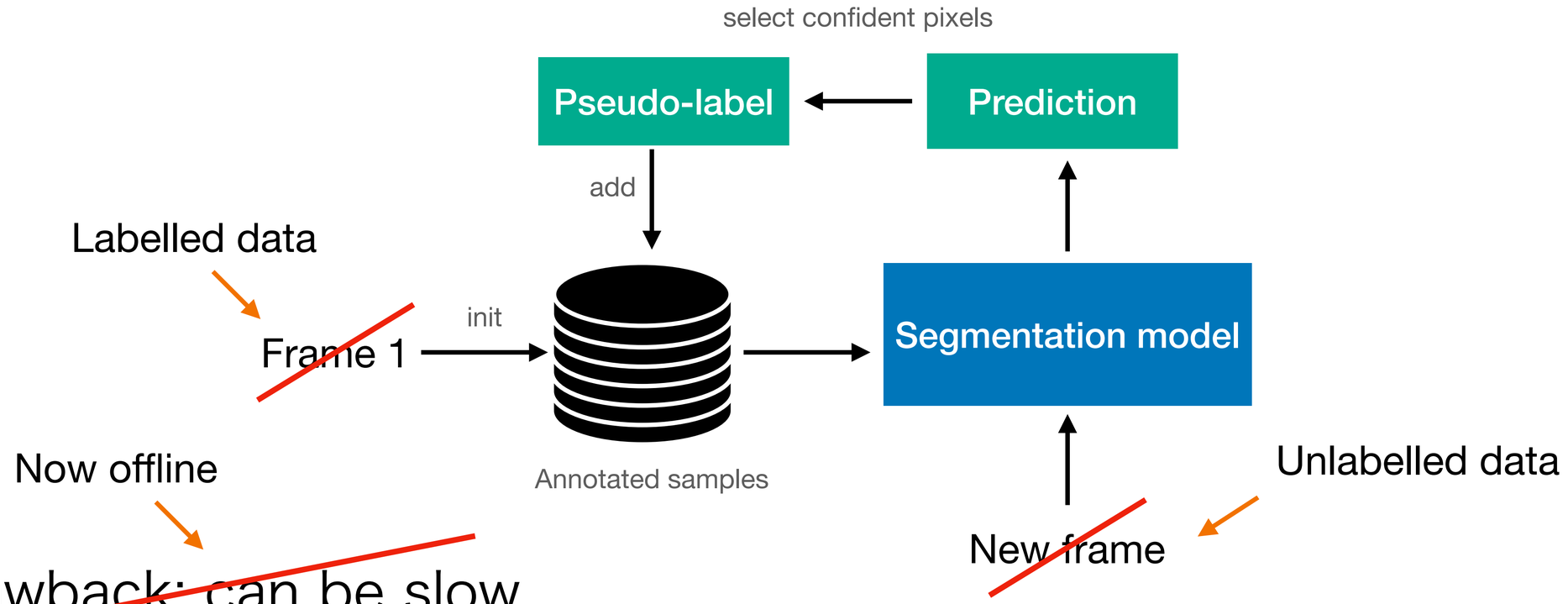
- Online adaptation: adapt model to appearance changes in every frame, not just the first frame.



- Drawback: can be slow.

OnAVOS: Online Adaptation

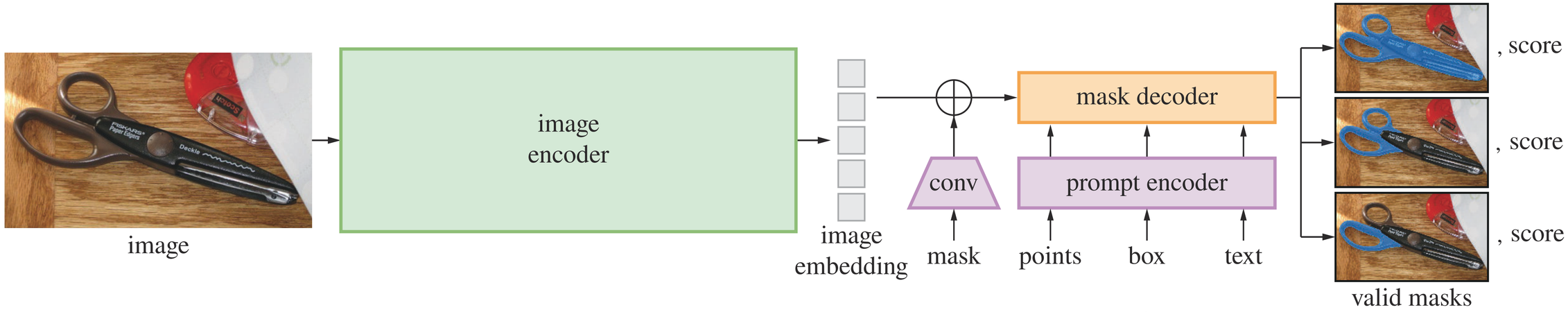
- Online adaptation: adapt model to appearance changes in every frame, not just the first frame.



- Drawback: can be slow.

Segment Anything

State-of-the-art segmentation:



Training data:

- 10.2M manually annotated masks
- 1.1B self-labelled masks

Kirillov et al., “Segment Anything” (2023)

Segment Anything

Point prompts

Text prompts (with bounding boxes)

Kirillov et al., “Segment Anything” (2023)

Self-training with pseudo labels

Open questions:

- How to select the labels (the confidence threshold)?
  - high vs low threshold trade-off (QUIZ)

Self-training with pseudo labels

Open questions:

- How to select the labels (the confidence threshold)?
  - high vs low threshold trade-off (QUIZ)
  - high: no learning signal (the gradient will be close to zero);
  - low: noisy labels → low accuracy.

Self-training with pseudo labels

Open questions:

- How to select the labels (the confidence threshold)?
  - high vs low threshold trade-off (QUIZ)
  - high: no learning signal (the gradient will be close to zero);
  - low: noisy labels → low accuracy.
- Tedious to train (multiple training rounds).

Self-training with pseudo labels

Open questions:

- How to select the pseudo-labels (the confidence threshold)?
  - high vs low threshold trade-off (QUIZ)
  - high: no learning signal (the gradient will be close to zero);
  - low: noisy labels → low accuracy.
- Tedious to train (multiple training rounds).
- Sensitive to the initial model:
  - fails if the initial predictions are largely inaccurate.

Self-training with pseudo labels

A general outline:

1. Train a strong baseline on the labelled set:
   - e.g. with heavy data augmentation (crops, photometric noise).
2. Predict pseudo-labels for the unlabelled set.
3. Continue training the network on both labelled and pseudo-labelled samples.
4. Repeat steps 2-3.

Taxonomy

- Unsupervised pre-processing, e.g.:
  - pre-training, clustering, etc.;
- Wrapper methods, e.g.:
  - self-training;
- **Intrinsically semi-supervised, e.g.**:
  - entropy minimisation;
  - virtual adversarial networks.

- Learning from synthetic data:
  - domain alignment;
  - consistency regularisation.

Inductive: Given a labelled and an unlabelled dataset, produce a classifier (operating on any input point).

Transductive: Given a labelled and an unlabelled dataset, produce labels for the unlabelled dataset.

van Engelen and Hoos, “A survey on semi-supervised learning” (2020)

Entropy minimisation

- Example:
  - Entropy minimisation for semantic segmentation (“self-training”).

$$\mathcal{L}(\{(x_i, y_i)\}_i, \{\hat{x}_i\}_i) = \sum_i \mathcal{L}_{\text{supervised}}(x_i, y_i) + \lambda \sum_i \mathcal{L}_{\text{unsupervised}}(\hat{x}_i)$$

- Objective: minimise the entropy of class distribution for each pixel:

$$\mathcal{L}_{\text{unsupervised}}(\hat{x}_i) := \mathbb{E}[-\log p(x_i)] \approx -\sum_i p(x_i) \log p(x_i)$$

Grandvalet and Bengio, “Semi-supervised Learning by Entropy Minimization” (2004).

Entropy minimisation

- Example:
  - Entropy minimisation for semantic segmentation (“self-training”).

$$\mathcal{L}(\{(x_i, y_i)\}_i, \{\hat{x}_i\}_i) = \sum_i \mathcal{L}_{\text{supervised}}(x_i, y_i) + \lambda \sum_i \mathcal{L}_{\text{unsupervised}}(\hat{x}_i)$$

- Objective: minimise the entropy of class distribution for each pixel:

$$\mathcal{L}_{\text{unsupervised}}(\hat{x}_i) := \mathbb{E}[-\log p(x_i)] \approx -\sum_i p(x_i) \log p(x_i)$$

QUIZ: Which of the 3 assumptions do we leverage here?

Grandvalet and Bengio, “Semi-supervised Learning by Entropy Minimization” (2004).

Virtual adversarial networks

Idea: Perturbations of the input should not change the label.

Quiz: Which assumption do we make here?

Let's consider a supervised case. We want to minimise:

$$D[q(y \mid x_*), p(y \mid x_* + r_{\text{adv}}, \theta)] \quad \text{where} \quad r_{\text{adv}} := \arg\max_{r:\|r\| \leq \epsilon} D[q(y \mid x_*), p(y \mid x_* + r, \theta)]$$

- $q(y \mid x_*)$: true posterior (ground-truth label)
- r_{adv}: perturbation
- θ: parameters we optimise for
- $\hat{\theta}$: parameters we fix in the current update

What if we do not have labels?

Replace $q(y \mid x_*)$ above with our current estimate, $p(y \mid x_*, \hat{\theta})$.

Taxonomy

- Unsupervised pre-processing, e.g.:
  - pre-training, clustering, etc.;
- Wrapper methods, e.g.:
  - self-training;
- Intrinsically semi-supervised, e.g.:
  - entropy minimisation;
  - virtual adversarial networks.

- **Learning from synthetic data:**
  - domain alignment;
  - consistency regularisation.

Inductive: Given a labelled and an unlabelled dataset, produce a classifier (operating on any input point).

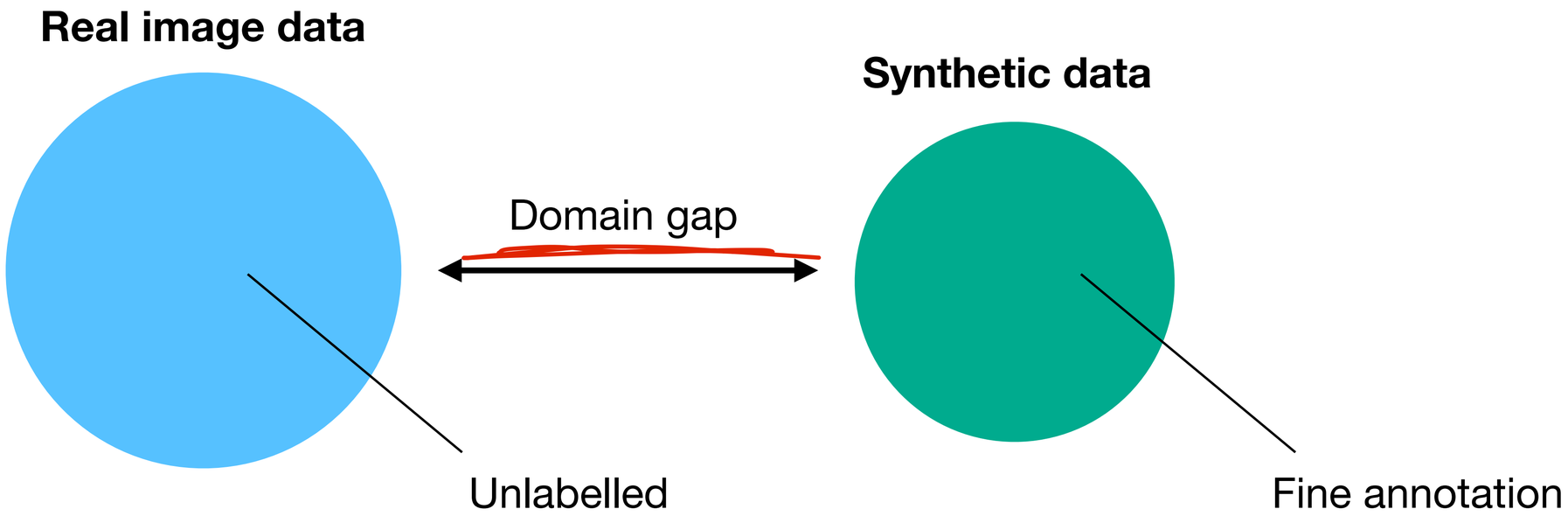
Transductive: Given a labelled and an unlabelled dataset, produce labels for the unlabelled dataset.

van Engelen and Hoos, “A survey on semi-supervised learning” (2020)

Learning from synthetic data

Labelled and unlabelled data may come from different distribution

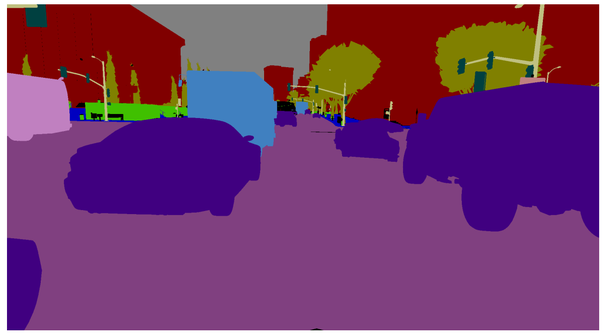
- e.g. due to differences in the synthetic and real appearance.

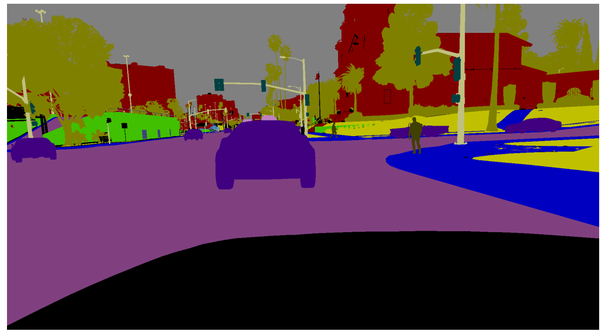



CRUCIAL FIX

Learning from synthetic data

- Labels are easier (hence cheaper) to obtain at a large scale.
- Consider it a special case of semi-supervised learning problem





Richter et al., “Playing for Data: Ground Truth from Computer Games” (2016).

Unsupervised Domain Adaptation

Model

train-time only

Unlabelled real-world data

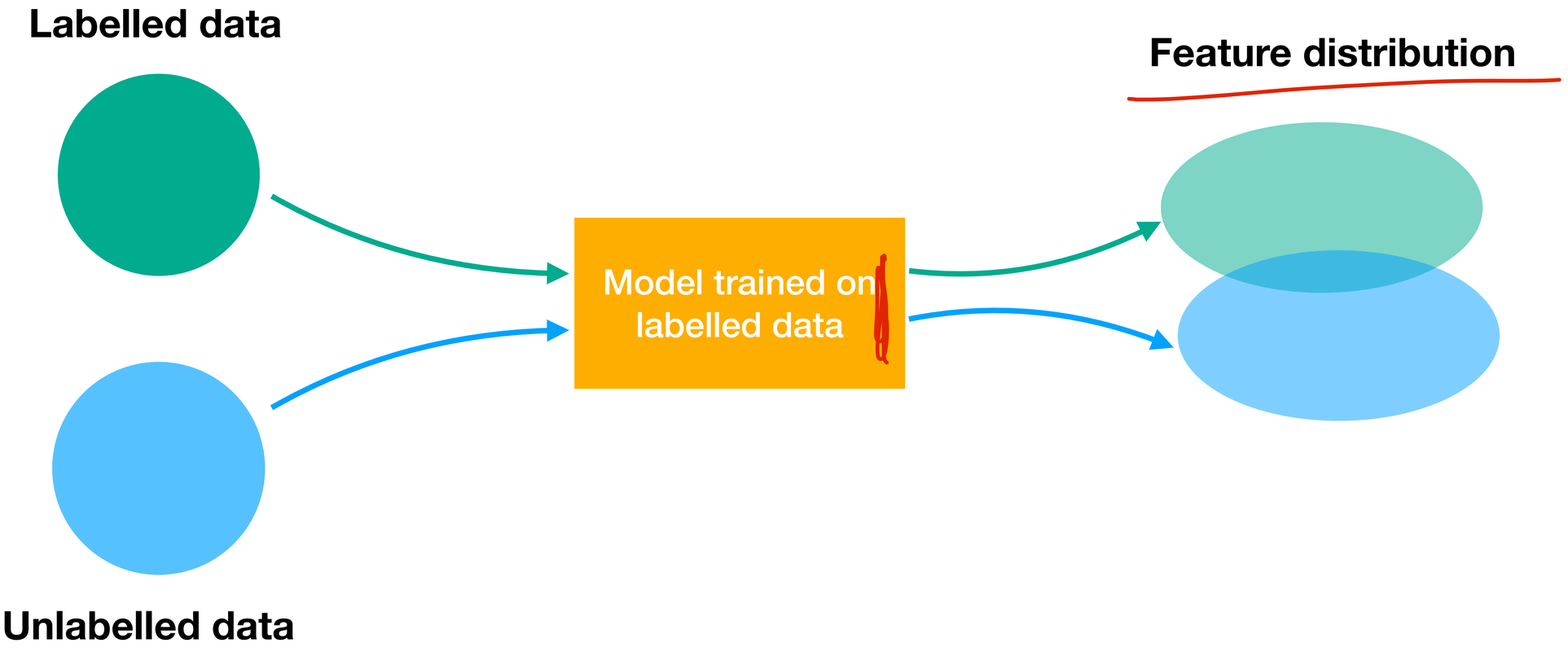
Semantic maps

Labelled synthetic data

Araslanov & Roth, 2021

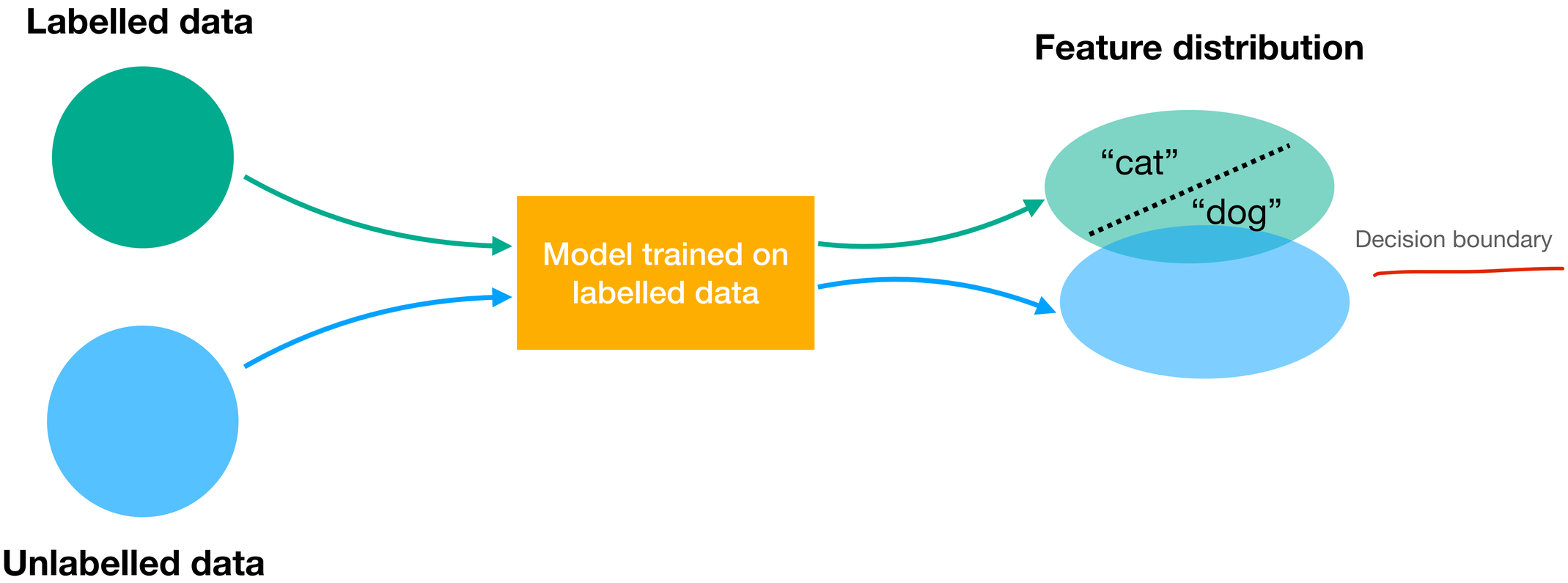
Domain alignment

This translates into disjoint feature distribution of a model trained only on the labelled data:



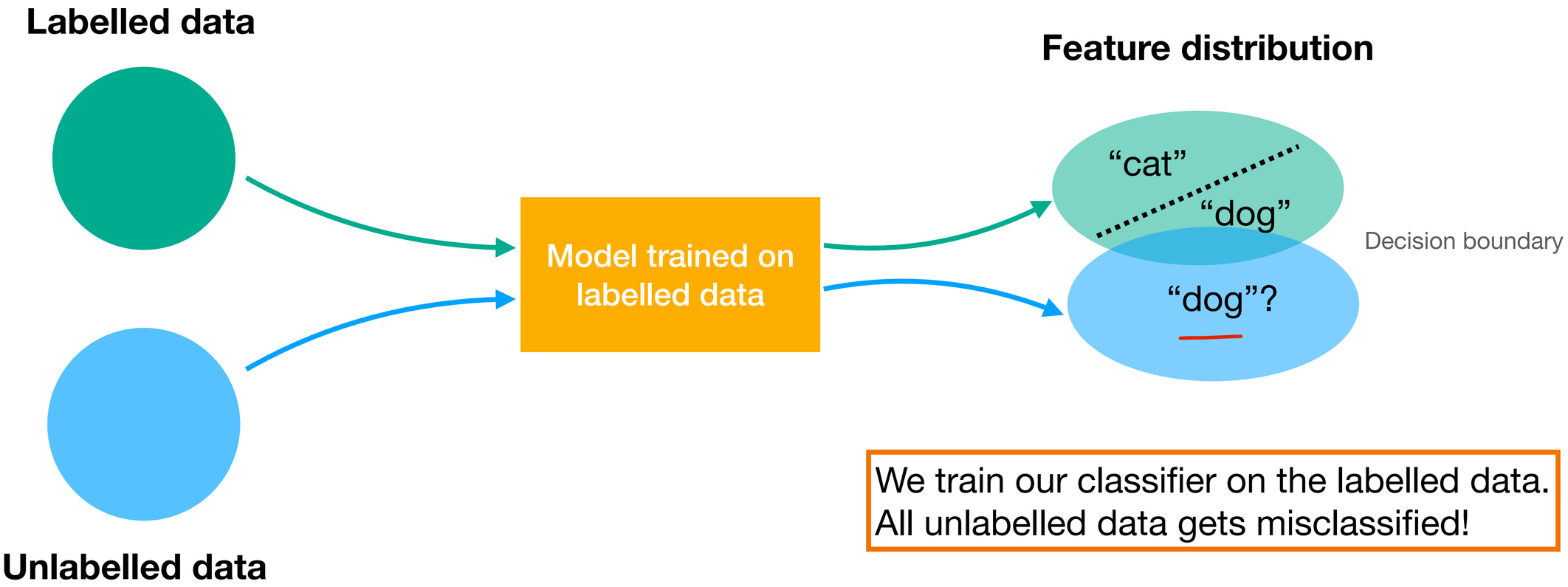
Domain alignment

This translates into disjoint feature distribution of a model trained only on the labelled data:



Domain alignment

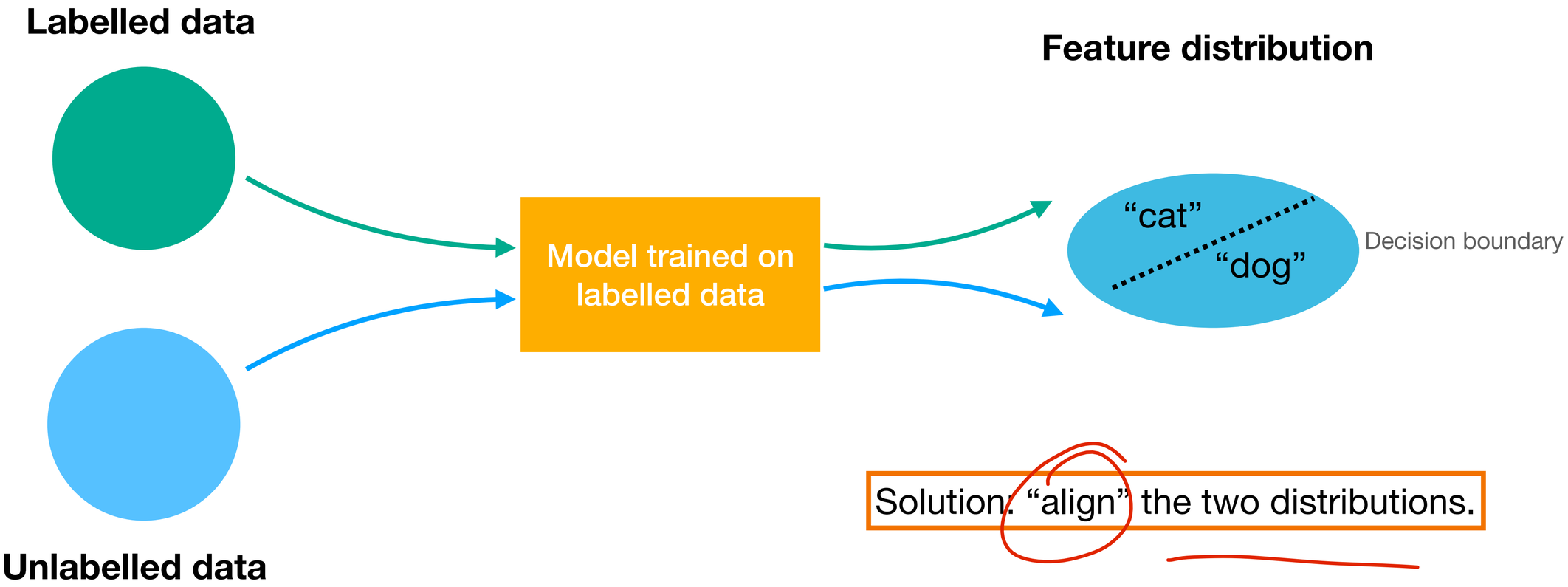
This translates into disjoint feature distribution of a model trained only on the labelled data:



We train our classifier on the labelled data.
All unlabelled data gets misclassified!

Domain alignment

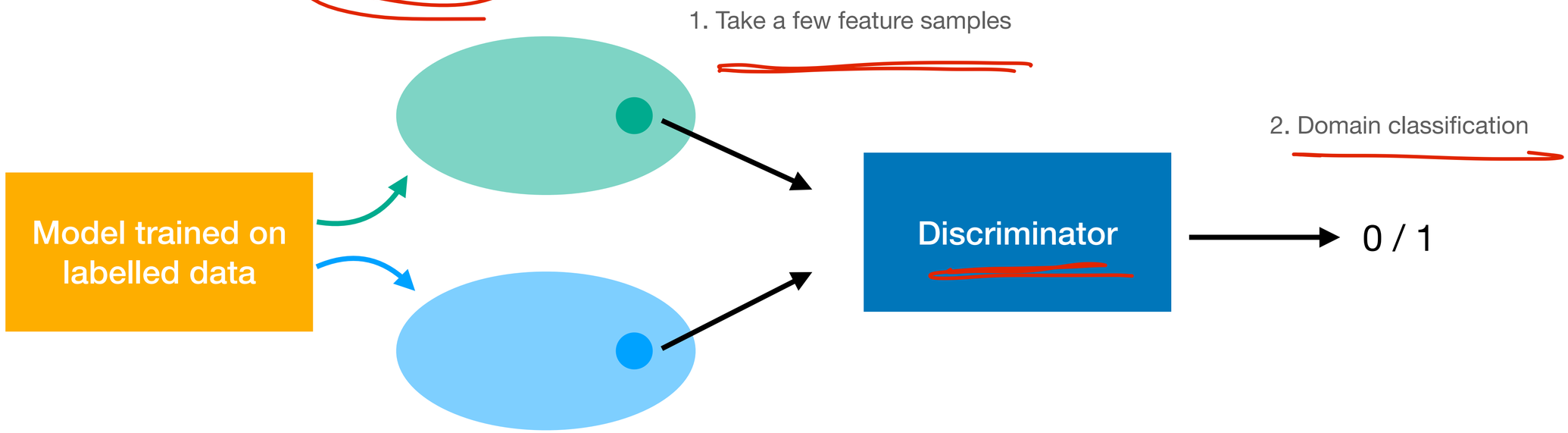
This translates into disjoint feature distribution of a model trained only on the labelled data:



Domain alignment

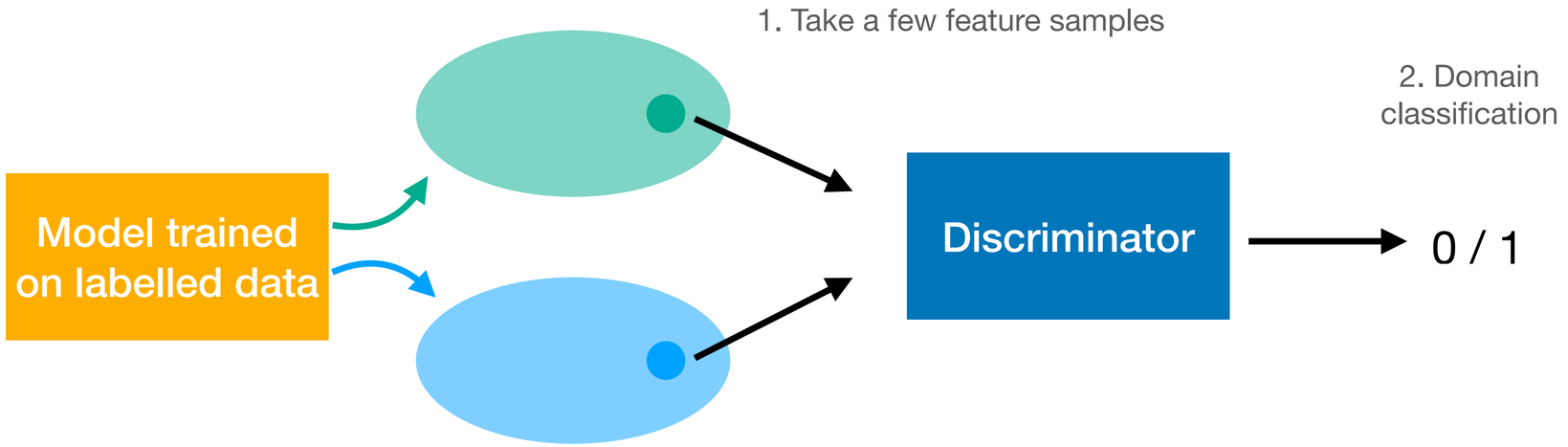
Domain alignment: make two feature distributions indistinguishable

- We could use a GAN:



Domain alignment

- Discriminator learns to classify the origin of the provided feature.
- The model learns
  - to classify the labelled images;
  - a feature representation that reduced discriminator accuracy.



Taxonomy

- Unsupervised pre-processing, e.g.:
  - pre-training, clustering, etc.;
- Wrapper methods, e.g.:
  - self-training;
- Intrinsically semi-supervised, e.g.:
  - entropy minimisation;
  - virtual adversarial networks.

- Learning from synthetic data:
  - domain alignment;
  - **consistency regularisation.**

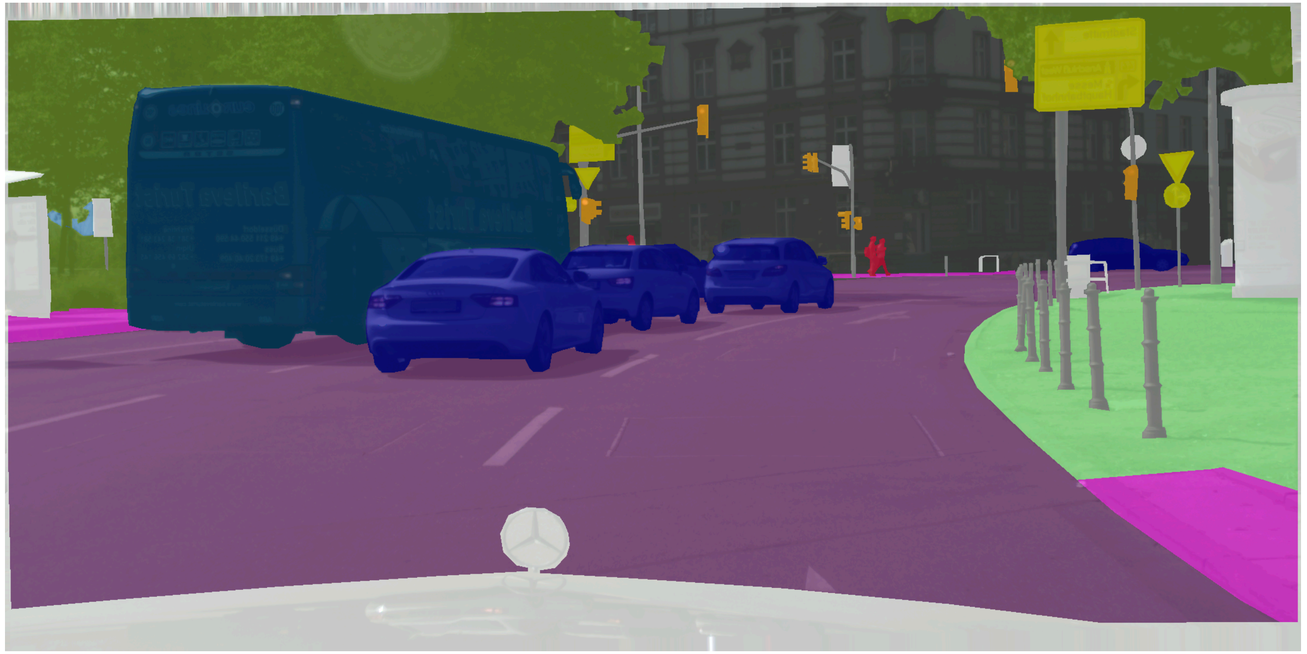
Inductive: Given a labelled and an unlabelled dataset, produce a classifier (operating on any input point).

Transductive: Given a labelled and an unlabelled dataset, produce labels for the unlabelled dataset.

van Engelen and Hoos, “A survey on semi-supervised learning” (2020)

Consistency regularisation

Consistent prediction across image transformations:



Araslanov & Roth, 2021

Consistency regularisation

Consistent prediction across image transformations:

Flipping

Araslanov & Roth, 2021

Consistency regularisation

Consistent prediction across image transformations:

Scaling

Araslanov & Roth, 2021

Consistency regularisation

Consistent prediction across image transformations:

Semantic meaning does not change, though not guaranteed in a deep net.

Araslanov & Roth, 2021

Consistency regularisation

Consistent prediction across image transformations:

permutation

QUIZ: Any similarity to what we have already seen in this lecture?

Semantic meaning does not change → use it as a consistency loss.

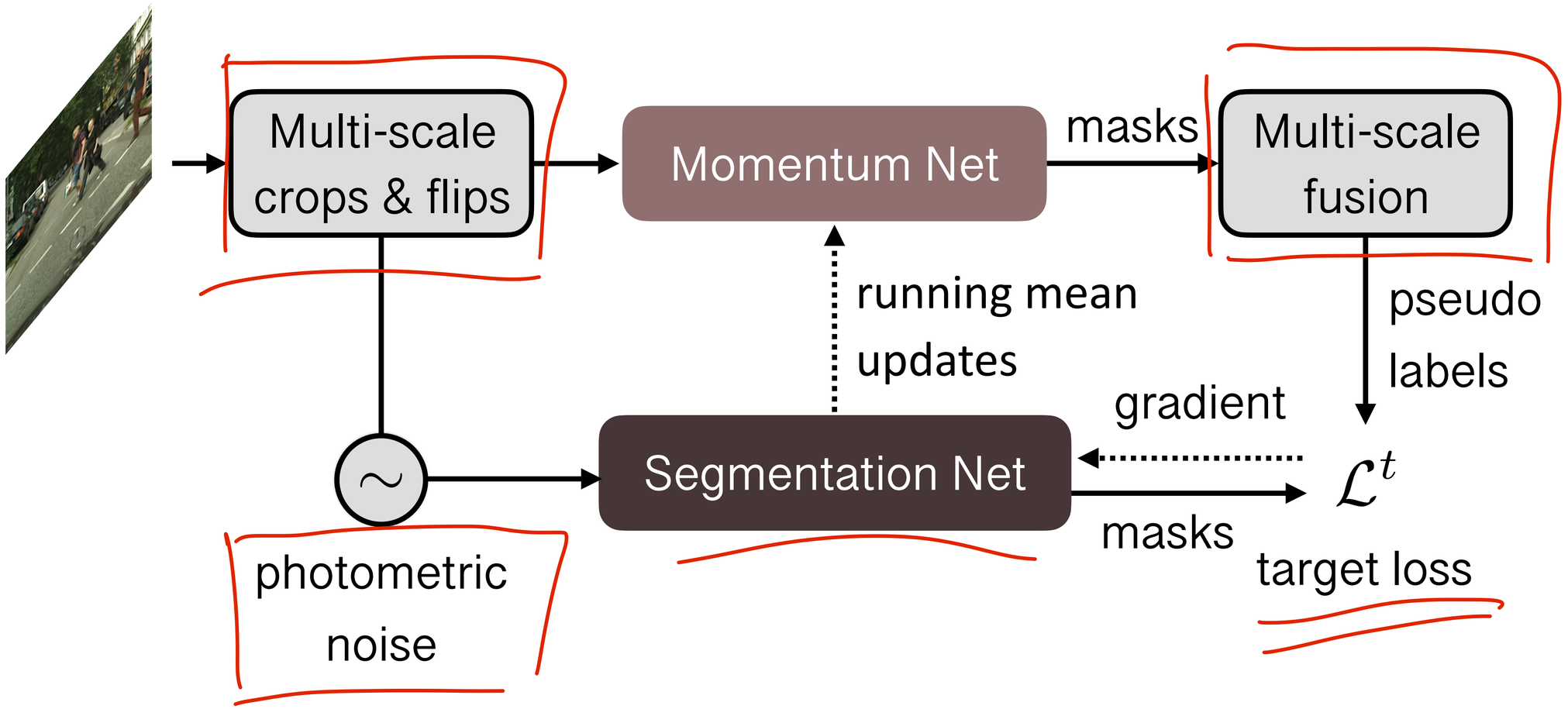
Araslanov & Roth, 2021

Some remarks

We have discussed a few techniques so far.

- These techniques are often complementary!
- State-of-the-art frameworks are typically a combination of multiple techniques.

Framework

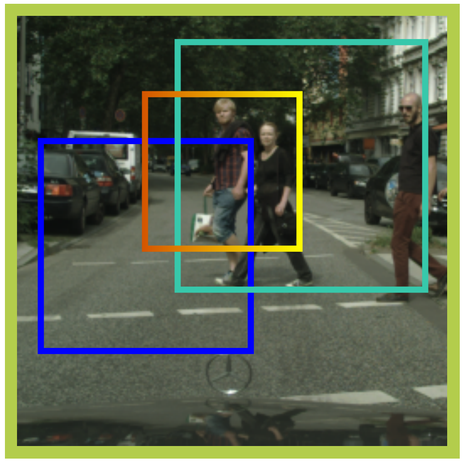


Araslanov & Roth, 2021

Momentum net

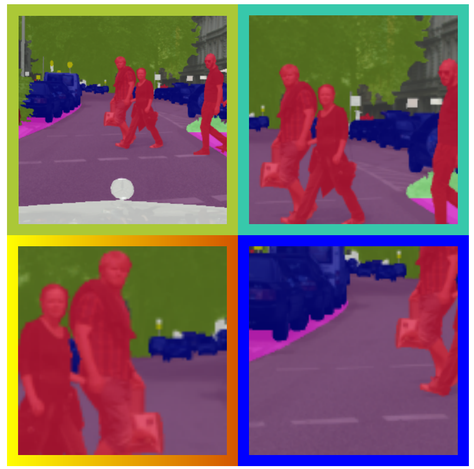
Test-time augmentation is applied online at training time:

Original

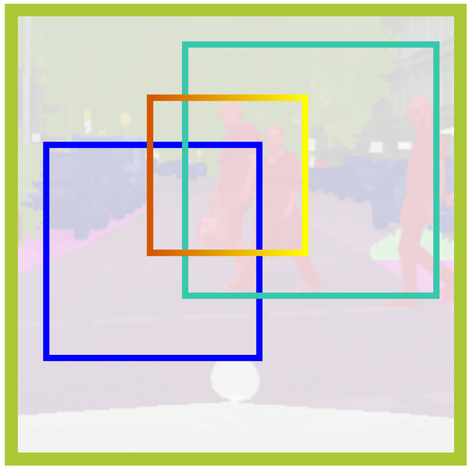


Input Batch

Predictions



Averaging



1. Random crops & horizontal flipping

2. Re-project and average predictions

3. Apply adaptive threshold

Araslanov & Roth, 2021


Before adaptation (Baseline)
After adaptation (Ours)
Araslanov & Roth, 2021

Additional reading

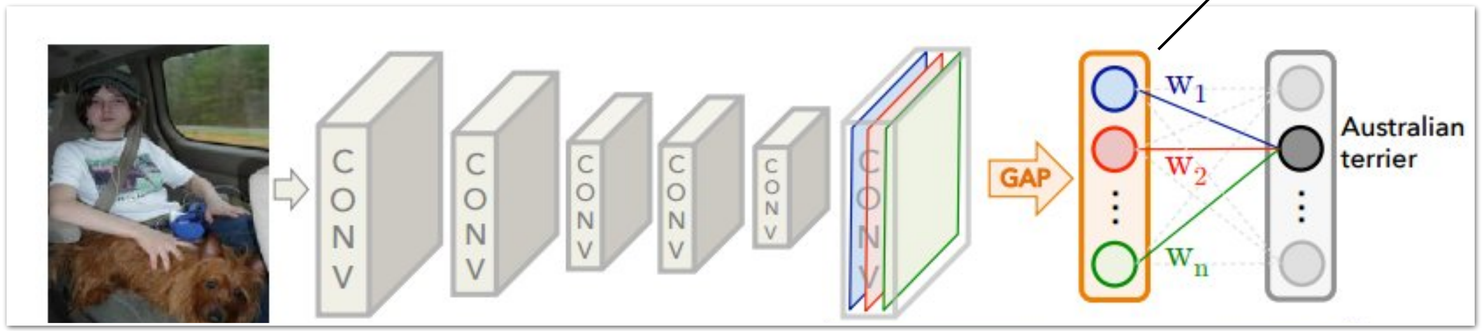
- Domain adaptation:
  - Ganin et al., “Domain-adversarial training of neural networks”. In JMLR, 2016.
  - Hoffman et al., “CyCADA: Cycle-consistent adversarial domain adaptation”. In ICML, 2018.
  - Richter et al., “Enhancing photorealism enhancement”. In TPAMI 2022.

- Learning from weak supervision:
  - Ahn et al., “Weakly supervised learning of instance segmentation with inter-pixel relations”. In CVPR 2019.
  - Araslanov and Roth. “Single-stage semantic segmentation”. In CVPR 2020.

Segmentation with image-level labels

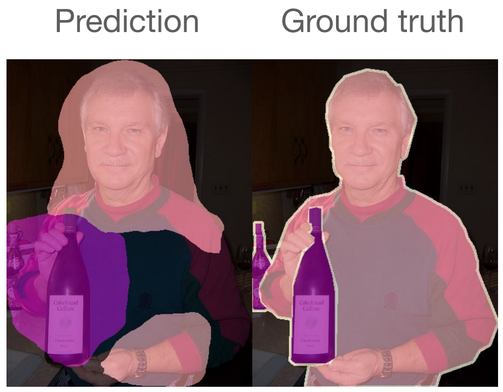
Idea: we can reuse the classifier weights to classify every pixel.

- Consider a classification network:



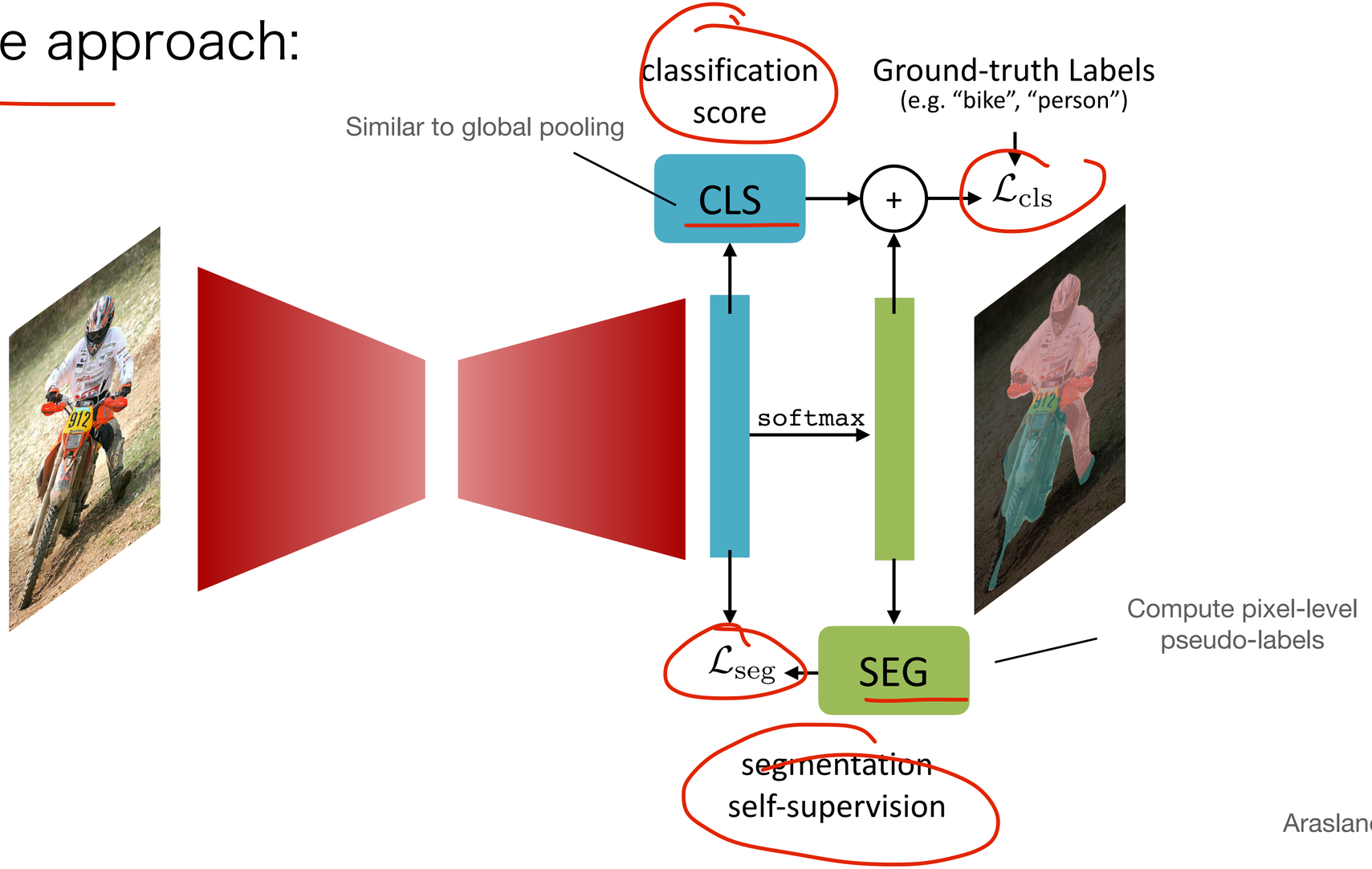
(Credit: Zhou et al., 2016)

- We can replace GAP and instead use 1x1 convolution.
- Insight: such classification turns out to be meaningful!



Segmentation with image-level labels

A single-stage approach:



Araslanov & Roth, 2020

Summary

- Entropy minimisation
  - improves accuracy, but leads to miscalibration.
- Virtual adversarial networks
  - generic treatment of supervised and unsupervised data.
- Consistency regularisation
  - can be effective, but is limited by available augmentation techniques.
- Self-training
  - simple and effective, but sensitive to initial pseudo-label quality.
- Unsupervised pre-training
  - simple and effective; this should be your first baseline.
- Domain alignment
  - typically less fine-tuning required, but can be still challenging to train (GAN).
- Coarse labels (weak supervision)
  - a cost-effective compromise between fully labelled and unlabelled data.

This is our last lecture

- **Next up**: Q&A on February 20 (details to be posted on Moodle).
  - Ask me anything about the lecture material and/or exam.
  - You can post questions in advance on Moodle.

Feeling challenged?

- We work on many exciting open research problems.
- Get in touch with me or TAs if you're interested!
  - Guided research, student research assistant (contract).
- Next semester (SoSe '24):
  - Practical course:

    Geometric Scene Understanding

  - Focus on a research-oriented project (closely guided by a mentor).

Computer Vision III:

Semi-supervised learning

Dr. Nikita Araslanov

16.01.2024