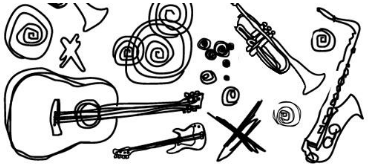
Tampere University
Tampere University of Applied Sciences

Unsupervised Adversarial Domain Adaptation Based On The Wasserstein Distance For Acoustic Scene Classification

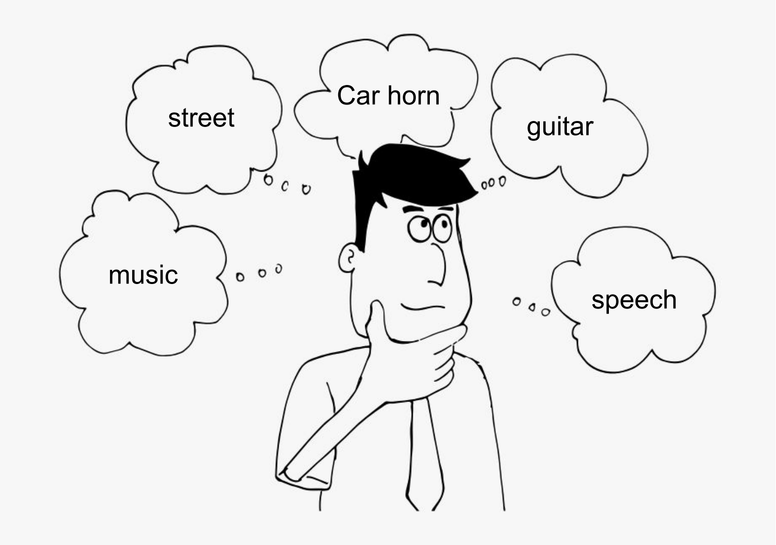
K. Drossos, P. Magron, T. Virtanen
Audio Research Group, Tampere University

Audio classification - Data collection



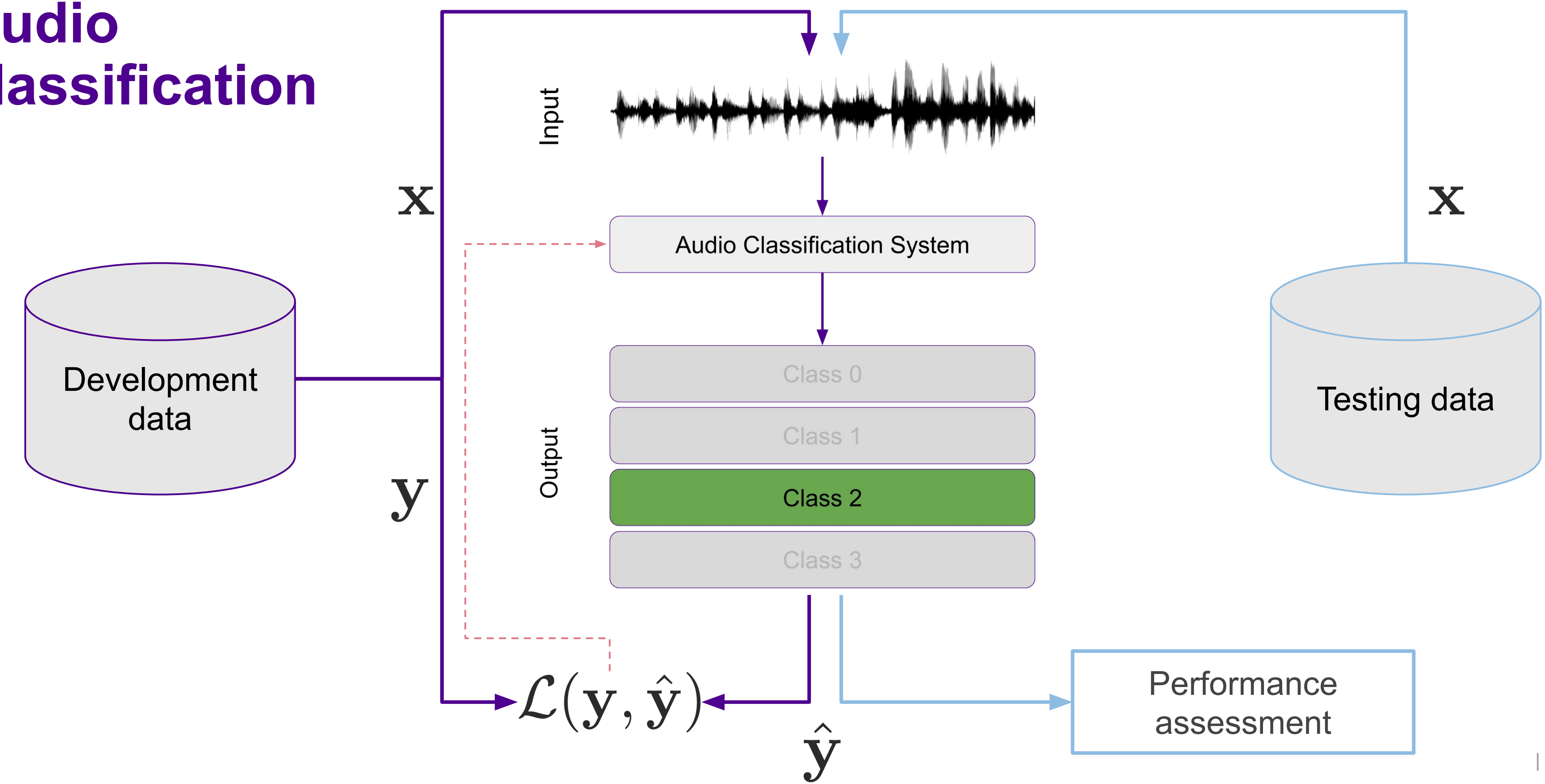
Audio signals

Recording and recording devices



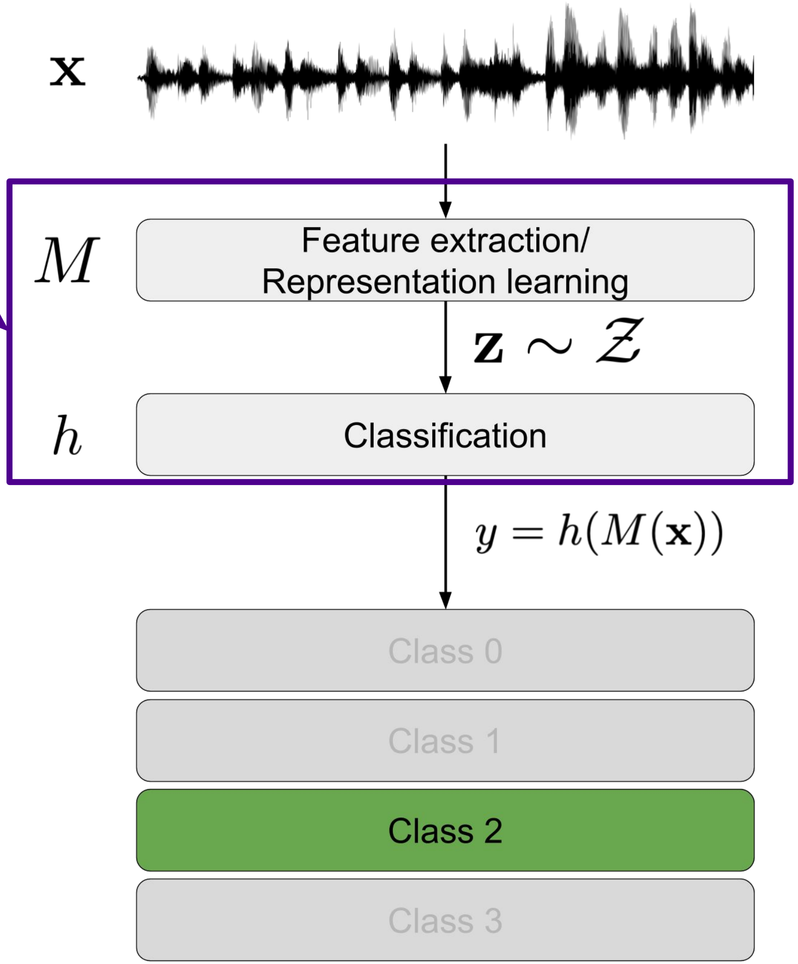
Human annotation

Audio classification



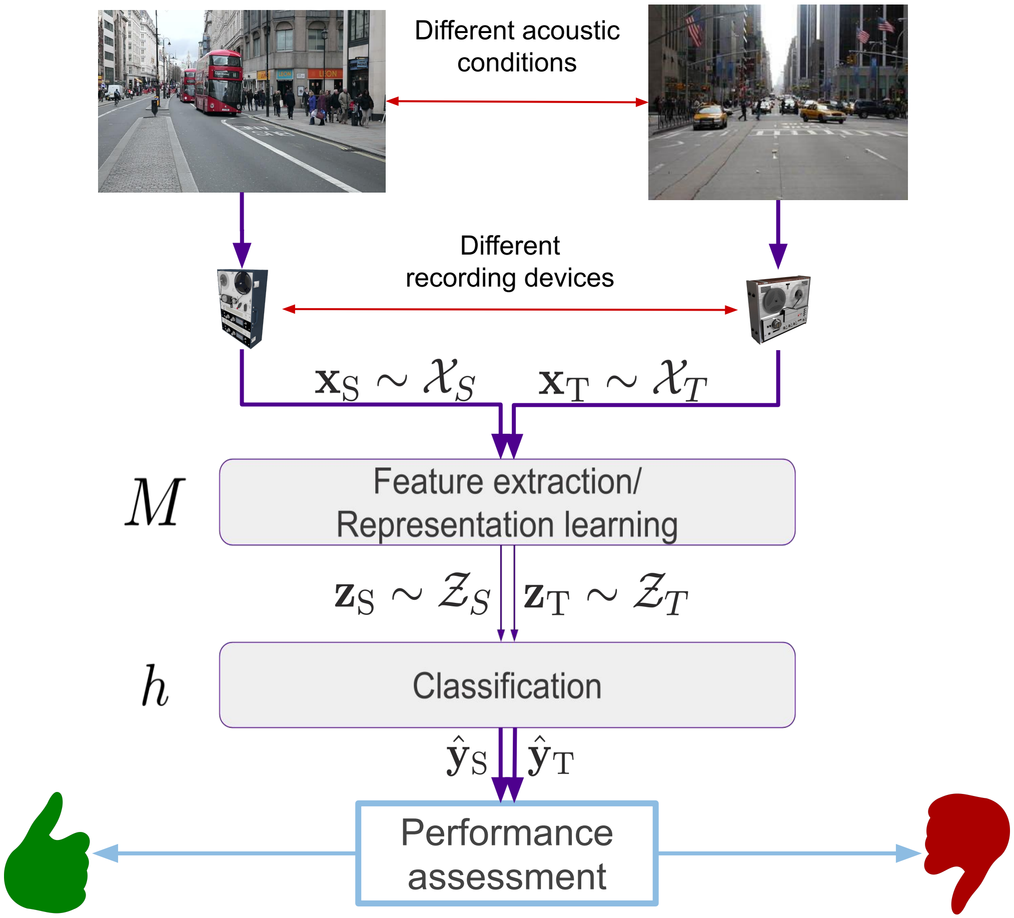
Audio Classification System

Audio classification system



- Input audio segment $\mathbf{x}$
- Feature extractor M
    - Learned representation $\mathbf{z}$
- Classifier h
- Optimized on training data
    - $\mathbf{x}_S \sim \mathcal{X}_S$
- Assess performance on testing data
    - $\mathbf{x}_T \sim \mathcal{X}_T$

Mismatching conditions



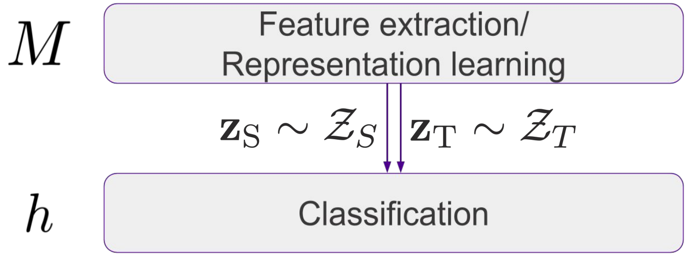
- Mismatch between $\mathcal{X}_S$ and $\mathcal{X}_T \rightarrow$ **poor performance** of h
- Promising tackling way $\rightarrow$ **domain adaptation**

Some definitions

- A **domain** $\mathcal{D} = \langle \mathcal{Z}, f \rangle$
  - A distribution $\mathcal{Z} \rightarrow \mathbf{z} \sim \mathcal{Z}$
  - A labeling process $f : \mathbf{z} \mapsto y,\ \ y \rightarrow$ ground truth label(s)
- Two domains
  - **Source** → ***optimization*** on
  - **Target** → ***adaptation*** on
- Average disagreement of h and f is $\epsilon(h, f) = \mathbb{E}_{\mathbf{z}}[\mathcal{L}(h(\mathbf{z}), f(\mathbf{z}))]$
  - Source domain → $\mathcal{D}_S = \langle \mathcal{Z}_S, f_S \rangle,\ \ \epsilon_S(h, f_S)$
  - Target domain → $\mathcal{D}_T = \langle \mathcal{Z}_T, f_T \rangle,\ \ \epsilon_T(h, f_T)$

S. Ben-David et al., "A theory of learning from different domains," Machine Learning, vol. 79, no. 1, pp. 151–175, May 2010.

Domain adaptation (DA)

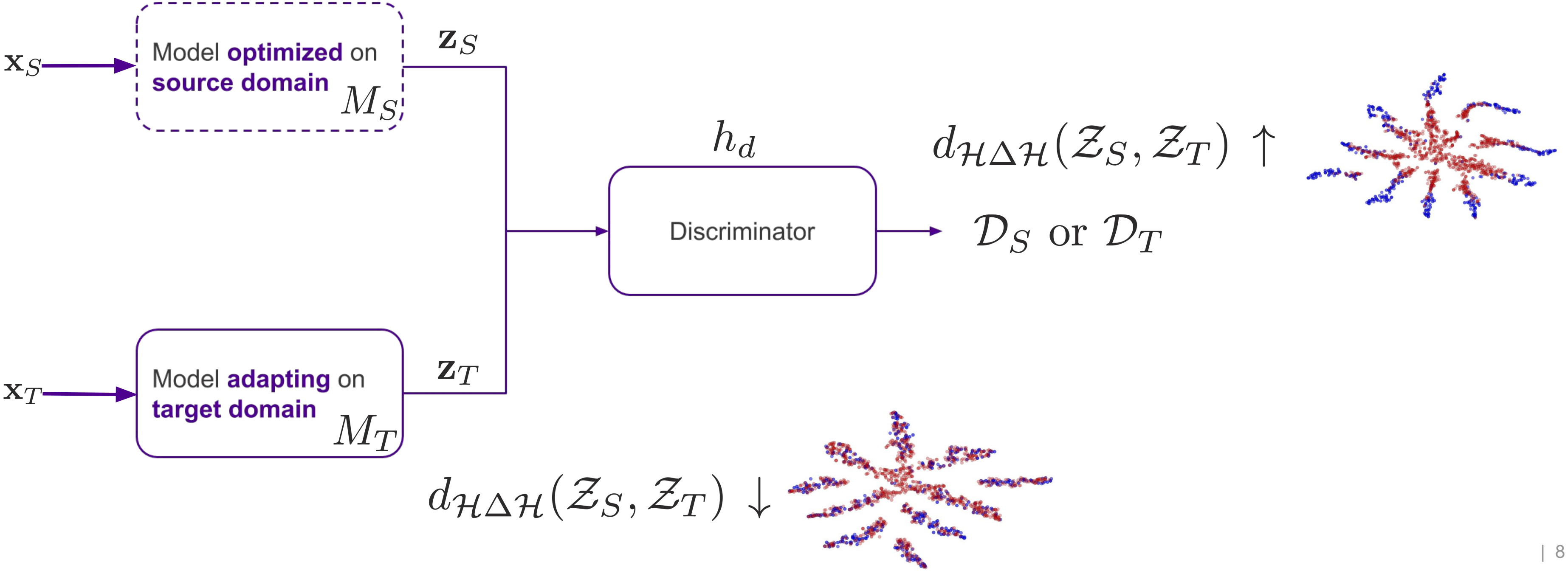


- If $\mathcal{Z}_S$ and $\mathcal{Z}_T$ are close $\rightarrow$ ϵ_S and ϵ_T will be close
  - $\mathcal{H}\Delta\mathcal{H}$ distance $\rightarrow$ discrepancy between $\mathcal{Z}_S$ and $\mathcal{Z}_T$
    - $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{Z}_S, \mathcal{Z}_T) = 2 \sup_{h,h'} \left|\mathrm{P}_{\mathbf{z}\sim\mathcal{Z}_S}[h(\mathbf{z}) \neq h'(\mathbf{z})] - \mathrm{P}_{\mathbf{z}\sim\mathcal{Z}_T}[h(\mathbf{z}) \neq h'(\mathbf{z})]\right|$
  - Use $\mathcal{H}\Delta\mathcal{H}$ to upper bound the error on $\mathcal{D}_T$
    - $\epsilon_T(h, f_T) \leq \epsilon_S(h, f_S) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{Z}_S, \mathcal{Z}_T) + \lambda$
- **Goal of DA** $\rightarrow$ Having a model that yields low ϵ_S, adapt and yield low ϵ_T
  - No labels from $\mathcal{D}_T$, unsupervised
  - Some labels from $\mathcal{D}_T$, semi-supervised

$$\lambda = \epsilon_S(h^\star, f_S) + \epsilon_T(h^\star, f_T)$$
$$h^\star = \operatorname*{argmin}_h \left(\epsilon_S(h, f_S) + \epsilon_T(h, f_T)\right)$$

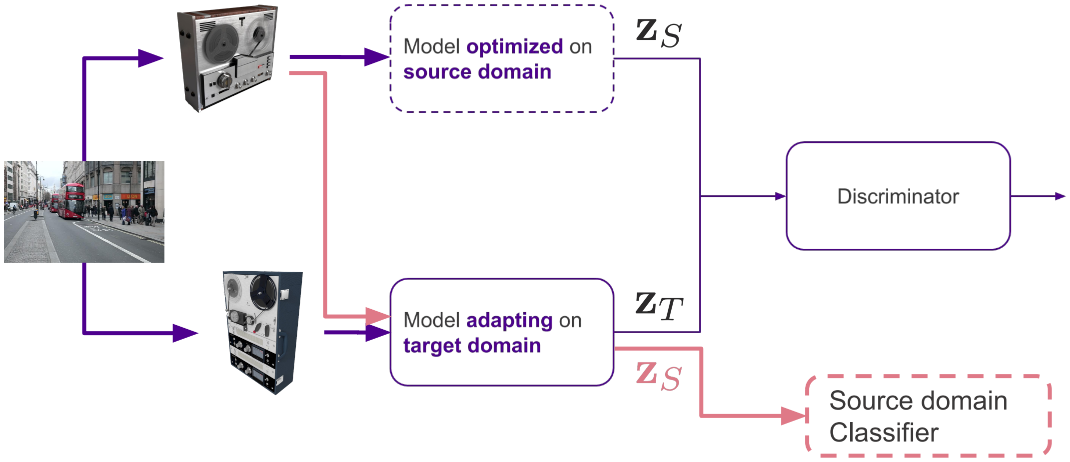
J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman, “Learning bounds for domain adaptation,” NeurIPS 2008
Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” ICML 2015

Adversarial DA



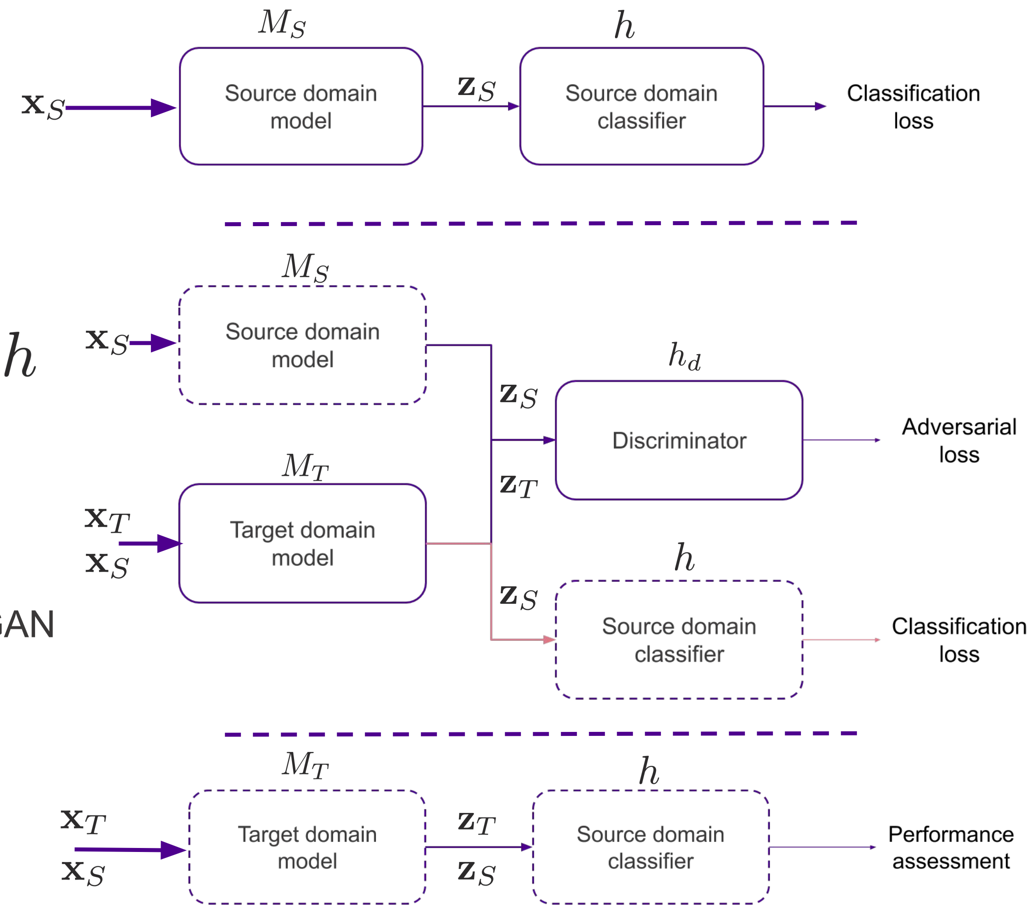
Proposed method overview

- Target problem
  - Acoustic scene classification
  - Mismatched recording devices
- Proposed solution
  - Adversarial unsupervised DA
- Key points
  - Wasserstein distance
  - Extra learning signal

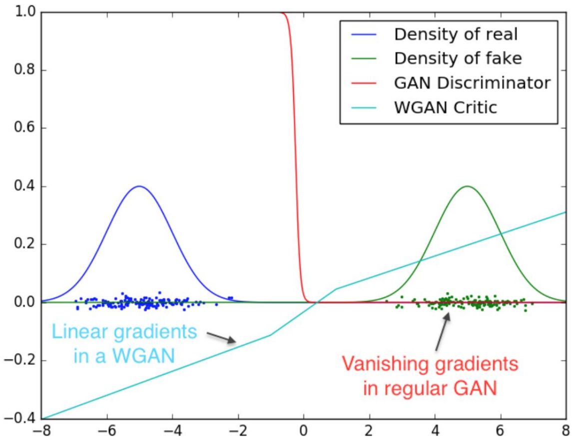


Method details

- Step 1 → **Optimize on $\mathcal{D}_S$**
  - Yielding model M_S and classifier h
  - Typical classification task
- Step 2 → **Adapt to $\mathcal{D}_T$**
  - Initialize M_T to M_S
  - Adapt M_T
  - Adversarial learning based on WGAN
- Step 3 → **Test on $\mathcal{D}_T$**
  - Use M_T and h



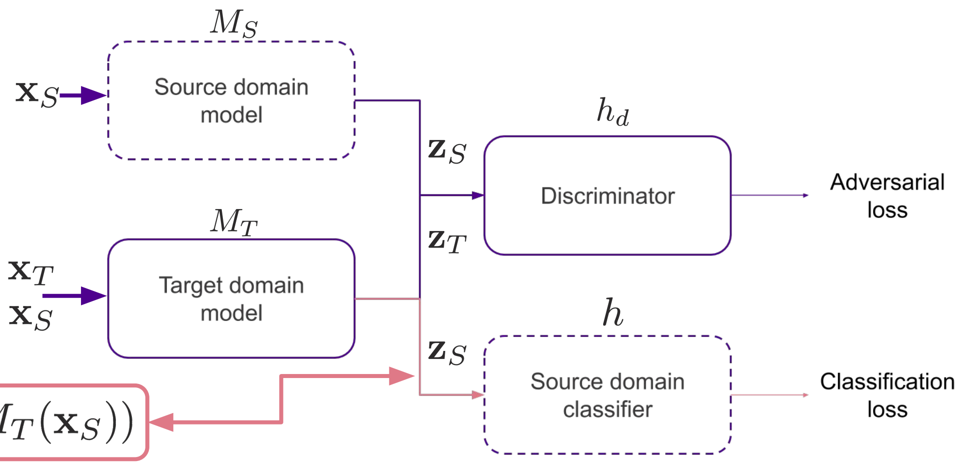
Adaptation details - Target domain



- Previous approach → typical GAN
  - Jensen–Shannon divergence minimization
- Training/Optimization problems
  - When discriminator is not fooled
  - When densities are away
- Wasserstein distance
  - Iterative minimization of
    - Discriminator → $\sum_{\mathbf{x}\sim\mathcal{X}_S} h_d(M_S(\mathbf{x})) - \sum_{\mathbf{x}\sim\mathcal{X}_T} h_d(M_T(\mathbf{x}))$
    - Target model → $\sum_{(\mathbf{x}_S,\mathbf{x}_T)\sim(\mathcal{X}_S,\mathcal{X}_T)} h_d(M_T(\mathbf{x}_T)) + \mathcal{L}_{\text{class}}(h, M_T(\mathbf{x}_S))$

Adaptation details - Source domain

- Retain performance on source domain
- Extra learning signal → M_T
- $$\sum_{(\mathbf{x}_S,\mathbf{x}_T)\sim(\mathcal{X}_S,\mathcal{X}_T)} h_d(M_T(\mathbf{x}_T)) + \boxed{\mathcal{L}_{\text{class}}(h, M_T(\mathbf{x}_S))}$$



Evaluation details - process

- **Source** domain → **professional** recording device
- **Target** domain → **consumer** recording devices
  - Synchronized recordings
  - Freely available dataset → DCASE 2018
- No pre-training → M_S from previous SOTA method
  - M_S is **available online** (**link in the paper**)
  - Previous SOTA → adversarial DA with typical GAN setting
- **Input** log mel-band energies → **output** acoustic scene label

Evaluation details - model

- Model → CNN based → **available online** (**link in the paper**)
  - Two CNN blocks (CNN, batch norm, max pooling, ReLU)
  - Two CNNs + ReLU
  - One CNN block
- Acoustic scene classifier
  - Two Linear layers + ReLU
- Domain classifier
  - Three CNN blocks

Results - source

Proposed

True label \ Predicted label

	airport	bus	metro	metro_station	park	public_square	shopping_mall	street_pedestrian	street_traffic	tram
airport	0.72	0.00	0.00	0.10	0.00	0.02	0.08	0.04	0.03	0.01
bus	0.00	0.60	0.20	0.01	0.00	0.00	0.00	0.00	0.00	0.19
metro	0.02	0.03	0.55	0.20	0.00	0.00	0.00	0.00	0.02	0.19
metro_station	0.09	0.02	0.04	0.73	0.00	0.02	0.02	0.03	0.01	0.05
park	0.02	0.00	0.00	0.02	0.84	0.06	0.00	0.01	0.01	0.04
public_square	0.01	0.01	0.03	0.07	0.09	0.50	0.00	0.05	0.19	0.06
shopping_mall	0.38	0.00	0.00	0.04	0.00	0.04	0.45	0.08	0.00	0.00
street_pedestrian	0.09	0.00	0.00	0.06	0.00	0.25	0.02	0.50	0.04	0.04
street_traffic	0.00	0.01	0.01	0.04	0.01	0.04	0.00	0.06	0.83	0.00
tram	0.01	0.09	0.15	0.06	0.01	0.00	0.01	0.02	0.00	0.66

Non-adapted model

True label \ Predicted label

	airport	bus	metro	metro_station	park	public_square	shopping_mall	street_pedestrian	street_traffic	tram
airport	0.69	0.0	0.03	0.08	0.0	0.03	0.12	0.02	0.02	0.01
bus	0.0	0.63	0.17	0.02	0.01	0.0	0.0	0.0	0.0	0.17
metro	0.0	0.03	0.63	0.18	0.0	0.01	0.0	0.0	0.02	0.12
metro_station	0.07	0.02	0.08	0.69	0.0	0.01	0.03	0.05	0.01	0.03
park	0.03	0.0	0.0	0.01	0.84	0.05	0.0	0.03	0.01	0.02
public_square	0.0	0.01	0.01	0.01	0.12	0.5	0.0	0.11	0.2	0.04
shopping_mall	0.37	0.0	0.01	0.03	0.0	0.03	0.52	0.04	0.0	0.0
street_pedestrian	0.05	0.0	0.0	0.06	0.01	0.23	0.05	0.56	0.02	0.01
street_traffic	0.0	0.0	0.01	0.04	0.01	0.04	0.0	0.05	0.85	0.0
tram	0.0	0.17	0.14	0.04	0.0	0.0	0.01	0.01	0.0	0.62

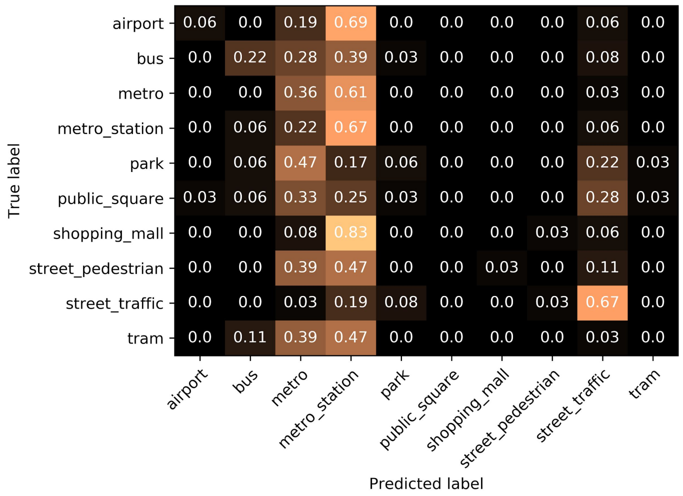
Prv SOTA

True label \ Predicted label

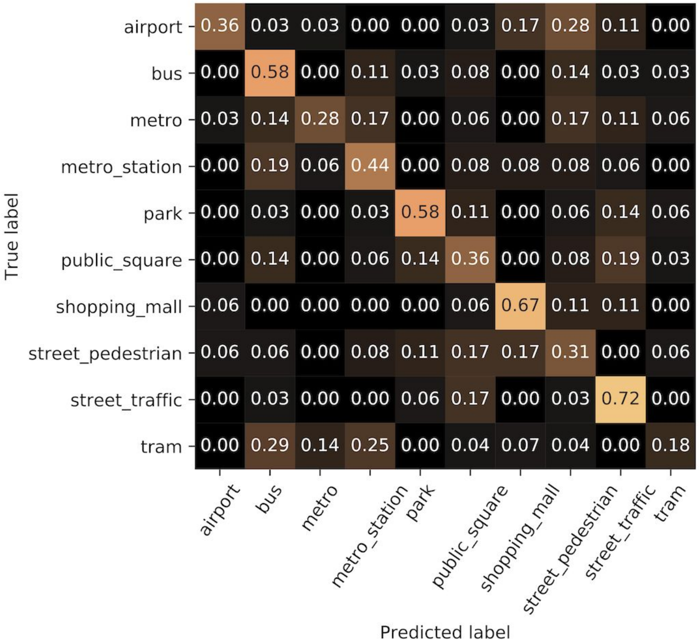
	airport	bus	metro	metro_station	park	public_square	shopping_mall	street_pedestrian	street_traffic	tram
airport	0.39	0.0	0.02	0.07	0.0	0.02	0.42	0.03	0.05	0.01
bus	0.0	0.62	0.24	0.03	0.02	0.0	0.0	0.0	0.0	0.09
metro	0.0	0.01	0.75	0.15	0.0	0.01	0.0	0.0	0.03	0.05
metro_station	0.07	0.02	0.07	0.67	0.0	0.01	0.06	0.08	0.02	0.01
park	0.02	0.01	0.0	0.0	0.82	0.07	0.0	0.03	0.01	0.02
public_square	0.0	0.0	0.04	0.02	0.24	0.36	0.0	0.12	0.19	0.02
shopping_mall	0.12	0.0	0.0	0.0	0.0	0.06	0.8	0.01	0.0	0.0
street_pedestrian	0.03	0.01	0.01	0.03	0.01	0.17	0.06	0.64	0.04	0.0
street_traffic	0.0	0.0	0.01	0.02	0.02	0.03	0.0	0.04	0.88	0.0
tram	0.0	0.09	0.22	0.07	0.0	0.0	0.02	0.02	0.0	0.57

Results - target

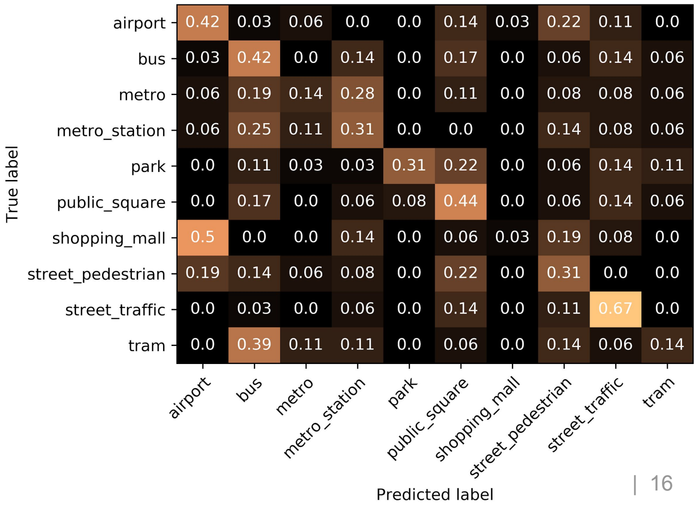
Non-adapted model



Proposed



Prv SOTA



Results - mean accuracy

	Non-adapted	Previous SOTA	Proposed
Source domain	0.65	0.65	0.64
Target domain	0.20	0.32	**0.45**

Conclusions and future research

- First approach for **unsupervised** DA for general audio **with WGAN**
  - Superpassed previous SOTA **by 13%** (mean accuracy)
- Expand unsupervised DA method for machine listening
- Future research:
  - Usage of bigger datasets
  - Multi-label problems
    - Sound event detection → releasing novel dataset: VOICe

Reproducibility

zenodo

- Binary files
  - Models
  - Dataset

- Code

PyTorch

- FOSS DL Framework

Paper on arXiv: https://arxiv.org/abs/1904.10678

Thank you!

Questions?