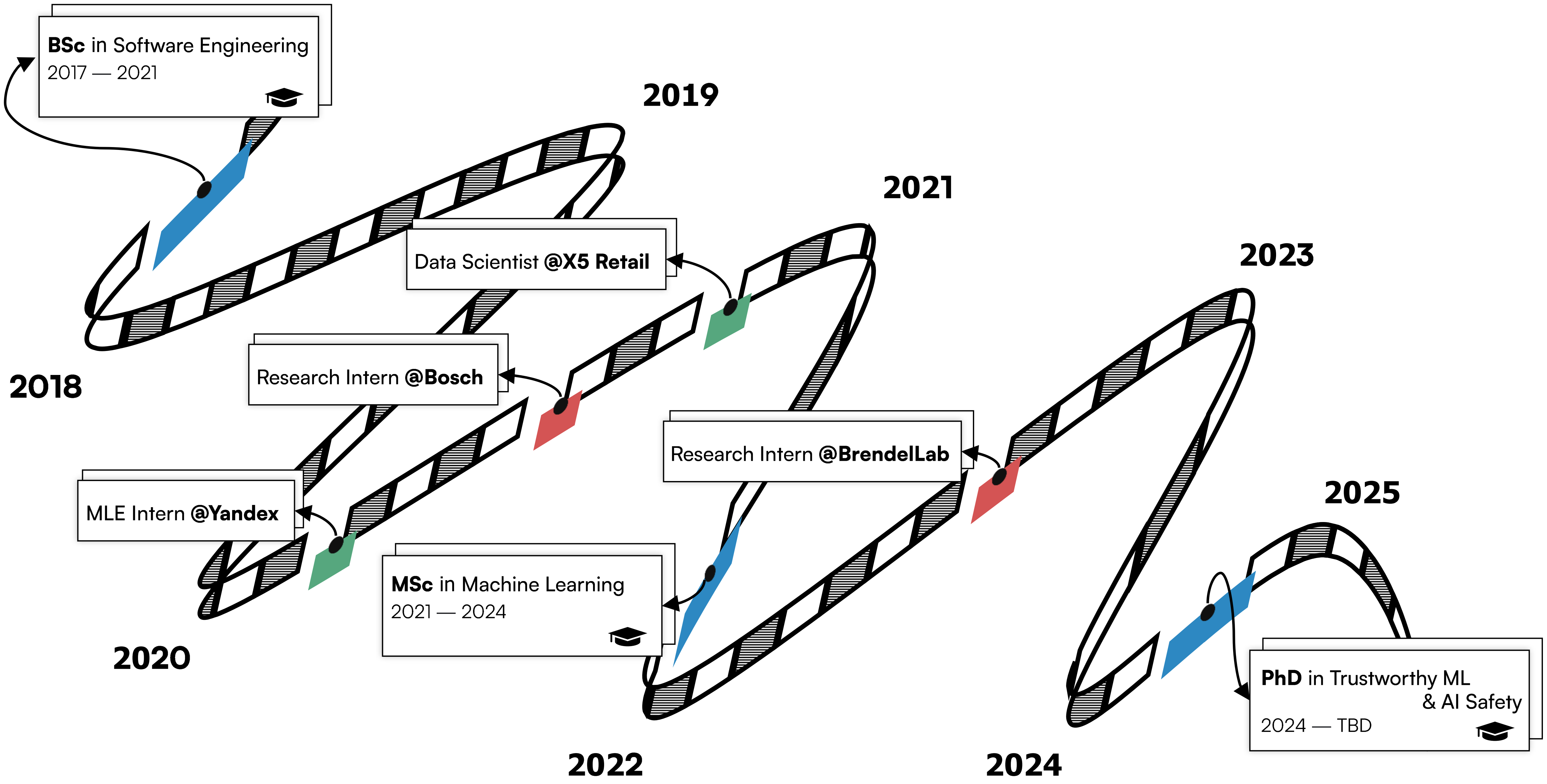
Out-of-the-(l_p)-Box:

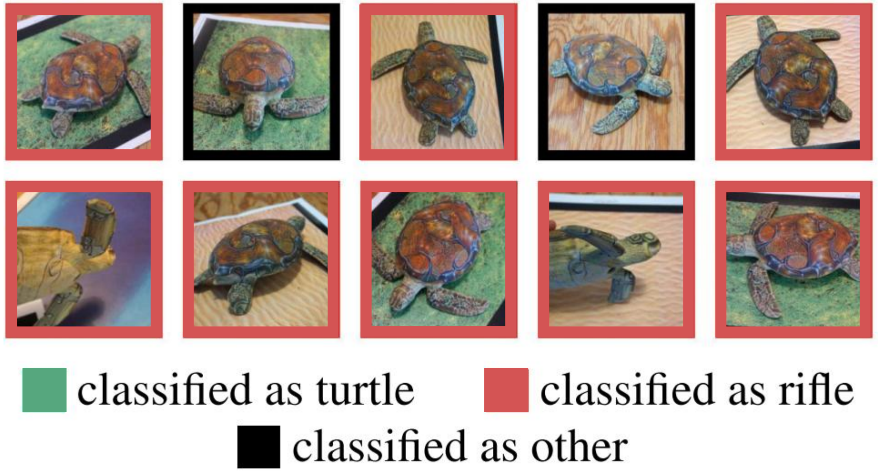
Exploiting Adversarials, Exploring Compositionality, and Exposing New AI Threats

IMPRS Scientific Talk by Alexander Panfilov
02/02/2024, Tübingen

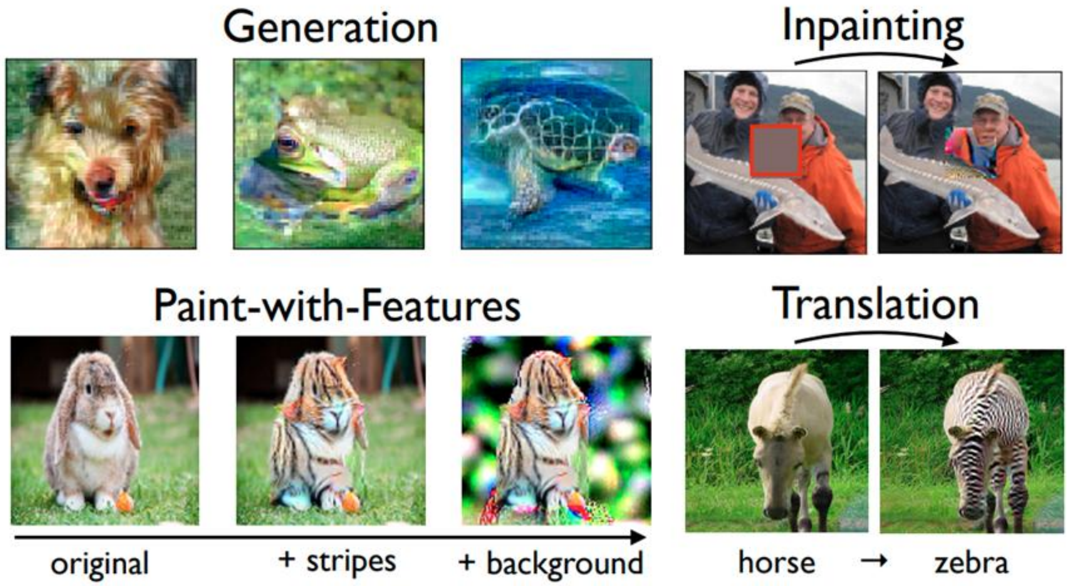
Background


BSc in Software Engineering
2017 — 2021
2019
2018
Data Scientist @X5 Retail
2021
Research Intern @Bosch
2023
Research Intern @BrendelLab
MLE Intern @Yandex
2025
MSc in Machine Learning
2021 — 2024
2020
PhD in Trustworthy ML
& AI Safety
2024 — TBD
2022
2024

Adversarial Threats and Robustness



Safety Problems[1]



Intriguing Properties[2]

[1] Athalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2018).
Synthesizing robust adversarial examples. In International Conference on Machine Learning (pp. 284-293). PMLR.

[2] Santurkar, S., Ilyas, A., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019).
Image synthesis with a single (robust) classifier. Advances in Neural Information Processing Systems, 32.

“Easy Features” for Domain Adaptation with OT

MNIST

$f(\mathcal{D}_{\text{source}})$

high accuracy

find OT map T

$f(T(\mathcal{D}_{\text{target}}))$

fair accuracy

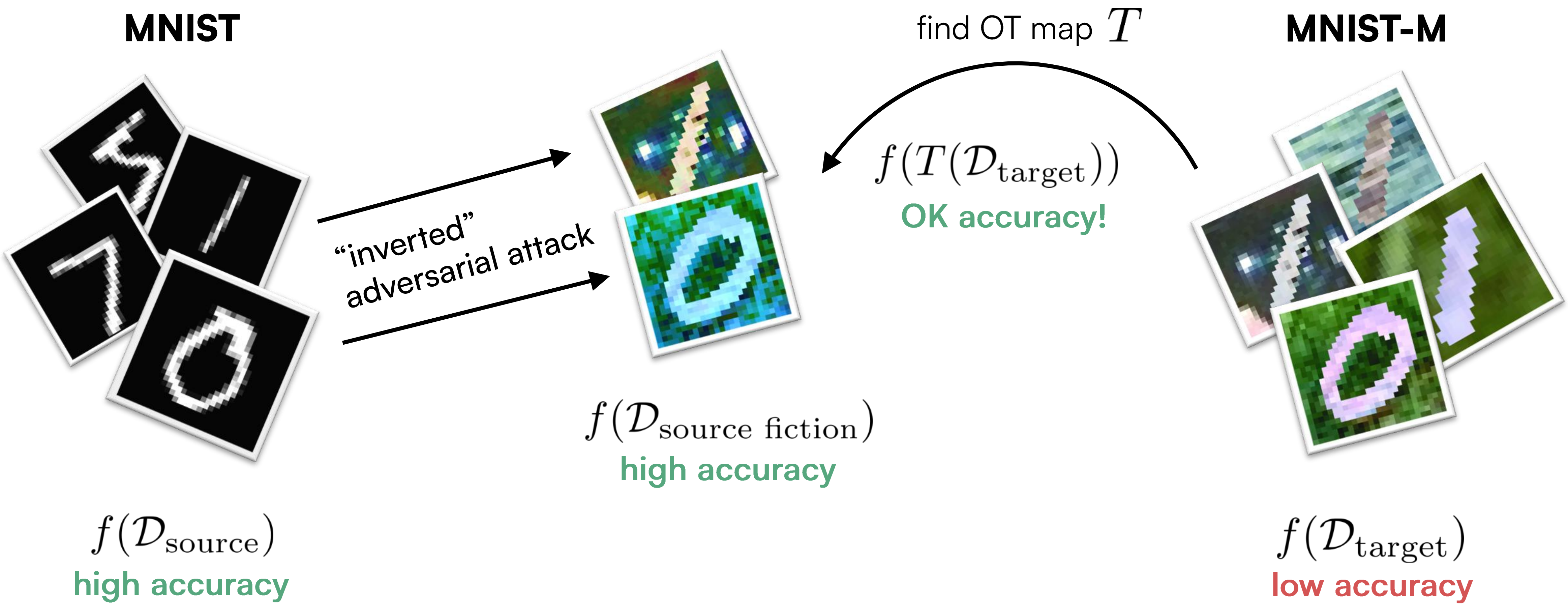
MNIST-M

$f(\mathcal{D}_{\text{target}})$

low accuracy

Asadulaev, A., Shutov, V., Korotin, A., Panfilov, A., Kontsevaya, V., & Filchenkov, A. (2023).
A Minimalist Approach for Domain Adaptation with Optimal Transport. Proceedings of The 2nd Conference on Lifelong Learning Agents, (vol. 232, pp. 1009-1024), PMLR.

"Easy Features" for Domain Adaptation with OT



Asadulaev, A., Shutov, V., Korotin, A., Panfilov, A., Kontsevaya, V., & Filchenkov, A. (2023).
A Minimalist Approach for Domain Adaptation with Optimal Transport. Proceedings of The 2nd Conference on Lifelong Learning Agents, (vol. 232, pp. 1009-1024), PMLR.

“Easy Features” for Domain Adaptation with OT

Method	M / S	S / M	M / U	M / MM
EMD	21.2 ±3	68.7±3	79.2±2	56.1±3
EMD(*sf*)	**23.0±3**	**86.3±3**	**83.1±2**	**62.7±2**
OTLin	21.8±4	69.9±4	84.1±7	62.3±1
OTLin(*sf*)	**25.5±4**	**88.4±4**	**89.3±6**	**64.5±3**
Sinkh	21.8±4	68.8±2	82.1±7	55.7±12
Sinkh(*sf*)	**25.5±4**	**86.2±4**	**83.8±6**	**62.9±4**
SinkhLp	21.8±4	68.8±6	84.8±16	55.7±19
SinkhLp(*sf*)	**25.5±4**	**86.3±7**	**88.3±19**	**63.0±27**
SinkhL2	21.8±4	68.8±4	84.8±2	55.7±4
SinkhL2(*sf*)	**25.5±4**	**86.3±2**	**88.3±2**	**63.0±2**

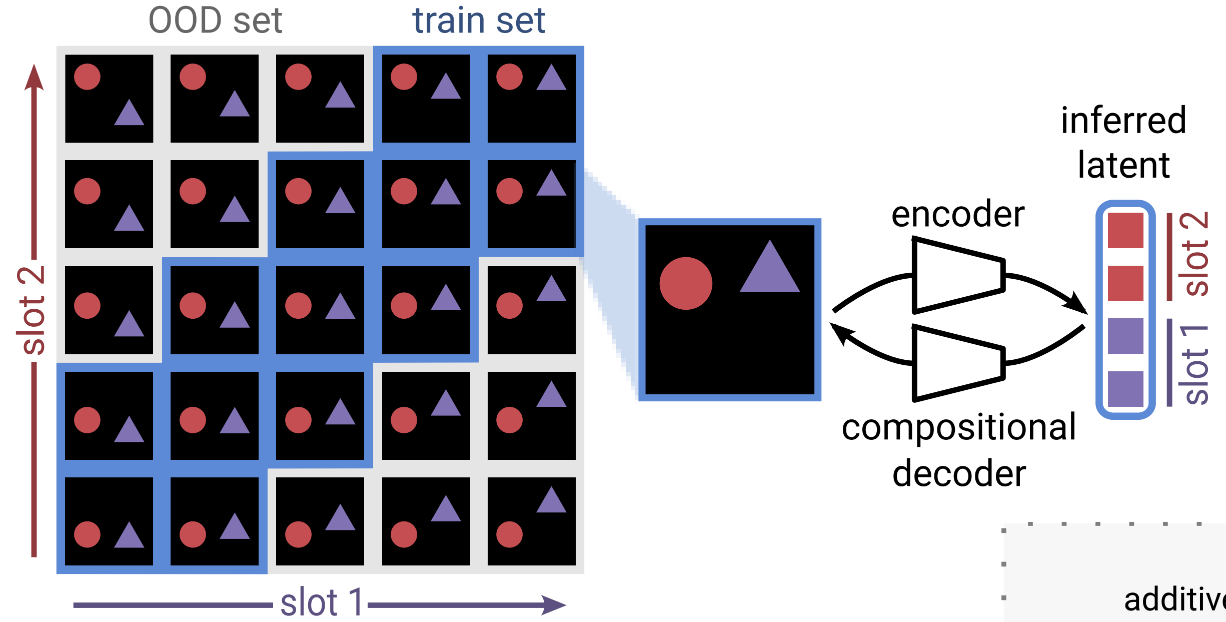
(a) Digits dataset domains.

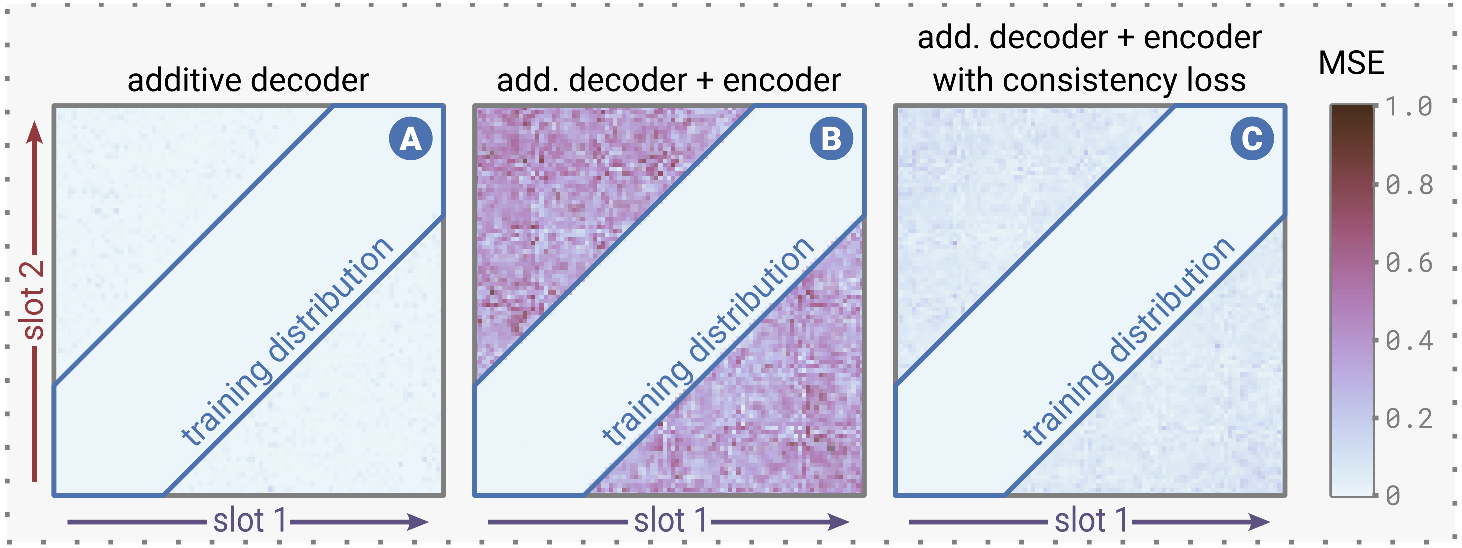
Method	A / S	S/ A	A / W	W / A	S / W	W/ S
EMD	38.4±3	9.3±5	45.2±3	45.6±5	13.6±3	36.7±3
EMD(*sf*)	**56.8±3**	**29.7±4**	**64.9±3**	**73.9±4**	**40.1±3**	**60.1±2**
OTLin	37.1±3	11.0±3	38.7±3	47.5±3	6.2±3	39.6±4
OTLin(*sf*)	**58.5±3**	**29.8±3**	**65.2±3**	**74.4±3**	**40.1±3**	**63.1±5**
Sinkh	38.0±3	10.1±4	44.7±6	45.5±3	13.1±7	37.2±3
Sinkh(*sf*)	**57.0±3**	**31.0±4**	**65.2±7**	**73.9±3**	**39.9±4**	**60.0±2**
SinkhLp	38.1±6	10.4±8	45.2±7	45.3±5	13.1±7	37.2±3
SinkhLp(*sf*)	**57.2±6**	**31.0±11**	**65.2±8**	**74.0±5**	**40.1±5**	**60.1±4**
SinkhL2	38.1±4	10.4±7	45.0±4	45.3±6	13.1±6	37.2±3
SinkhL2(*sf*)	**57.2±4**	**31.0±7**	**65.2±4**	**74.0±6**	**40.1±5**	**60.1±4**

(b) Modern Office-31 dataset domains.

Asadulaev, A., Shutov, V., Korotin, A., Panfilov, A., Kontsevaya, V., & Filchenkov, A. (2023).
A Minimalist Approach for Domain Adaptation with Optimal Transport. Proceedings of The 2nd Conference on Lifelong Learning Agents, (vol. 232, pp. 1009-1024), PMLR.

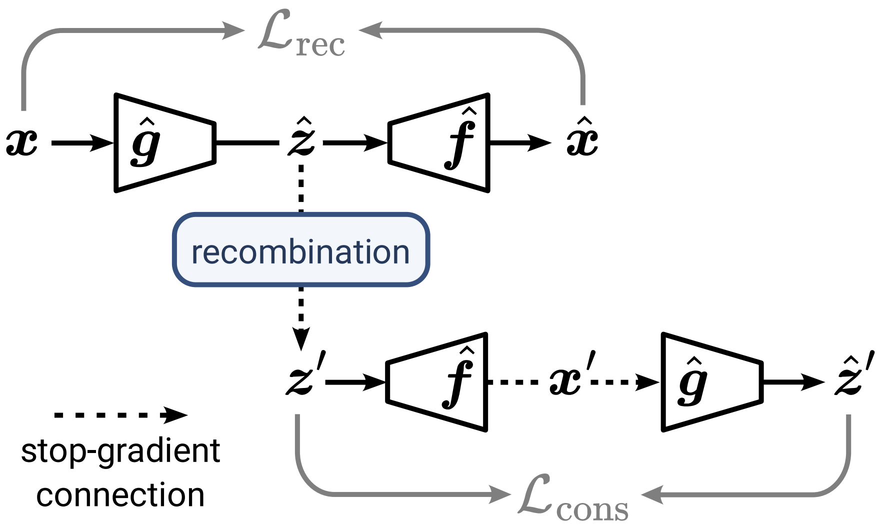
Provable Compositional Generalization for Object-Centric Learning

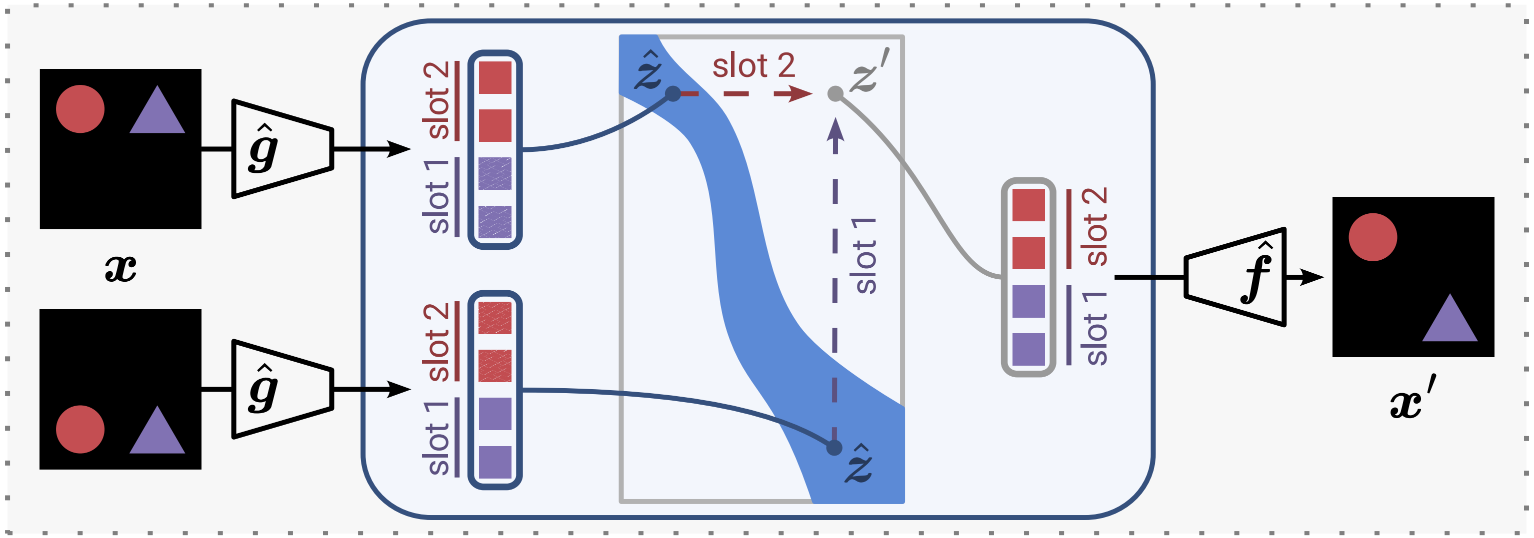




Wiedemer, T. *, Brady, J. *, Panfilov, A. *, Juhos, A. *, Bethge, M., & Brendel, W. (2023).
Provable Compositional Generalization for Object-Centric Learning. *arXiv preprint arXiv:2310.05327.*

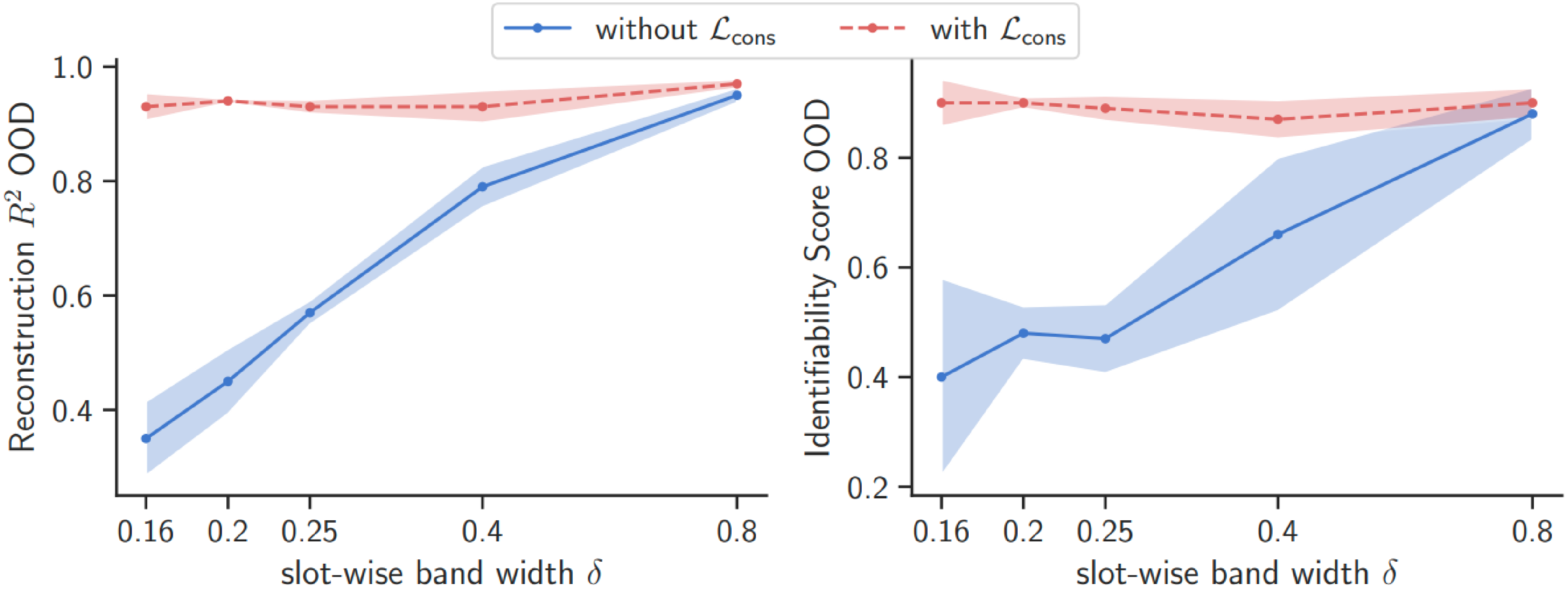
Provable Compositional Generalization for Object-Centric Learning





Wiedemer, T. *, Brady, J. *, Panfilov, A. *, Juhos, A. *, Bethge, M., & Brendel, W. (2023).
Provable Compositional Generalization for Object-Centric Learning. *arXiv preprint arXiv:2310.05327*.

Provable Compositional Generalization for Object-Centric Learning



Wiedemer, T. *, Brady, J. *, Panfilov, A. *, Juhos, A. *, Bethge, M., & Brendel, W. (2023).
Provable Compositional Generalization for Object-Centric Learning. *arXiv preprint arXiv:2310.05327.*

Topics of Interest

Topics of Interest

Different Aspects of Generalization

Safety & Security Problems

Topics of Interest

Different Aspects of Generalization

- Diffusion Models
  - Compositional (Concepts) Generalization
  - Shape Bias
  - Generative Classifiers

Safety & Security Problems

Topics of Interest

Different Aspects of Generalization

- Diffusion Models
  - Compositional (Concepts) Generalization
  - Shape Bias
  - Generative Classifiers

Safety & Security Problems

- Adversarial / Backdoor Attacks
- Data Memorization / Extraction
- Data Poisoning
- Watermarking

Topics of Interest

Different Aspects of Generalization

- Diffusion Models
  - Compositional (Concepts) Generalization
  - Shape Bias
  - Generative Classifiers
- Capabilities vs. Alignment

Safety & Security Problems

- Adversarial / Backdoor Attacks
- Data Memorization / Extraction
- Data Poisoning
- Watermarking
- New Threat Models
  - Author profiling
  - Deceptive Behavior
  - Malicious ChatBots

Topics of Interest

Different Aspects of Generalization

- Diffusion Models
  - Compositional (Concepts) Generalization
  - Shape Bias
  - Generative Classifiers
- Capabilities vs. Alignment

Safety & Security Problems

- Adversarial / Backdoor Attacks
- Data Memorization / Extraction
- Data Poisoning
- Watermarking
- New Threat Models
  - Author profiling
  - Deceptive Behavior
  - Malicious ChatBots

Topics of Interest

Some alignment researchers [..] believe that sufficiently advanced language models should be aligned to prevent an existential risk [...] to humanity: if this were true, **an attack** that causes such a model to become misaligned **would be devastating.** [1]

[1] Carlini, N., Nasr, M., Choquette-Choo, C. A., Jagielski, M., Gao, I., Awadalla, A., ... & Schmidt, L. (2023). Are aligned neural networks adversarially aligned?. arXiv preprint arXiv:2306.15447.

Thank you!