# Test-Time Training with Self-Supervision for Generalization under Distribution Shifts

How to Ace the Test

**허세민**

# Contents

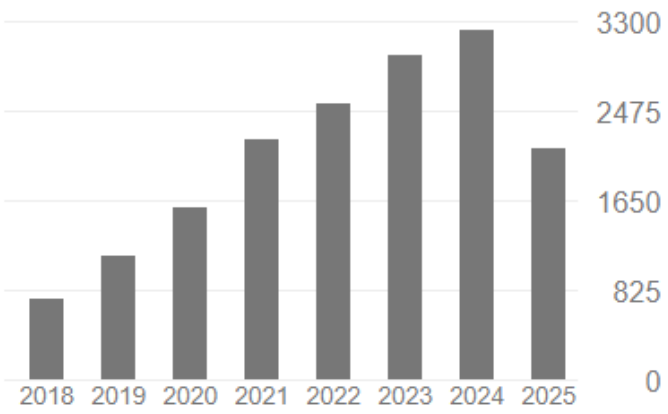
# Author

## Yu Sun

Postdoc, Stanford University
cs.stanford.edu의 이메일 확인됨 - 홈페이지

Artificial Intelligence  Machine Learning

| 인용 | | 모두 보기 |
|---|---|---|
| | 전체 | 2020년 이후 |
| 서지정보 | 17244 | 14723 |
| h-index | 16 | 16 |
| i10-index | 16 | 16 |




| | | |
|---|---|---|
| On calibration of modern neural networks<br>C Guo*, G Pleiss*, Y Sun*, KQ Weinberger<br>ICML 2017 | 7827 | 2017 |
| From word embeddings to document distances<br>M Kusner, Y Sun, N Kolkin, K Weinberger<br>International Conference on Machine Learning, 957-966 | 3105 | 2015 |
| Deep networks with stochastic depth<br>G Huang*, Y Sun*, Z Liu, D Sedra, KQ Weinberger<br>European Conference on Computer Vision, 646-661 | 3069 | 2016 |
| Test-time training with self-supervision for generalization under distribution shifts<br>Y Sun, X Wang, Z Liu, J Miller, AA Efros, M Hardt<br>International Conference on Machine Learning (ICML) | 1133 | 2020 |
| An empirical study on evaluation metrics of generative adversarial networks<br>Q Xu, G Huang, Y Yuan, C Guo, Y Sun, F Wu, K Weinberger<br>arXiv preprint arXiv:1806.07755 | 426 | 2018 |
| Adversarial deep averaging networks for cross-lingual sentiment classification<br>X Chen, Y Sun, B Athiwaratkun, C Cardie, K Weinberger<br>Transactions of the Association for Computational Linguistics 6, 557-570 | 375 | 2018 |
| Unsupervised Domain Adaptation through Self-Supervision<br>Y Sun, E Tzeng, T Darrell, AA Efros<br>arXiv preprint arXiv:1909.11825 | 290 | 2019 |
| Supervised Word Mover's Distance<br>G Huang*, C Guo*, MJ Kusner, Y Sun, F Sha, KQ Weinberger<br>Advances in Neural Information Processing Systems, 4862-4870 | 264 | 2016 |
| Test-Time Training with Masked Autoencoders<br>Y Gandelsman*, Y Sun*, X Chen, AA Efros<br>Advances in Neural Information Processing Systems | 218 | 2022 |
| Self-supervised policy adaptation during deployment<br>N Hansen, R Jangir, Y Sun, G Alenyà, P Abbeel, AA Efros, L Pinto, ...<br>ICLR 2021 | 209 | 2020 |
| Learning to (Learn at Test Time): RNNs with Expressive Hidden States<br>Y Sun*, X Li*, K Dalal*, J Xu, A Vikram, G Zhang, Y Dubois, X Chen, ...<br>ICML 2025 | 142 | 2024 |

# Background

# Distribution shift

✓ 학습 분포와 테스트 분포가 달라지는 현상
- **시간 경과** : 사용자 트렌드 변화, 계절적 요인
- **환경의 변화** : 병원의 새로운 의료 장비 도입, 다른 국가에 모델 배포
- **데이터 수집 오류 :** 센서 결함, 샘플링 방식의 변경

| 유형 | 무엇이 변하나 | 수학적 표현 |
|---|---|---|
| **Covariate shift** | 입력 분포 | $P(x)$변화, $P(y\|x)$동일 |
| **Label (prior) shift** | 레이블 비율 | $P(y)$변화, $P(y\|x)$동일 |
| **Concept/conditional shift** | 조건부 관계 | $P(y\|x)$ 변화 |

# Distribution shift

✓ 학습 분포와 테스트 분포가 달라지는 현상
- **시간 경과** : 사용자 트렌드 변화, 계절적 요인
- **환경의 변화** : 병원의 새로운 의료 장비 도입, 다른 국가에 모델 배포
- **데이터 수집 오류 :** 센서 결함, 샘플링 방식의 변경

| 유형 | 무엇이 변하나 | 수학적 표현 |
|---|---|---|
| **Covariate shift** | 입력 분포 | $P(x)$변화, $P(y\|x)$동일 |
| **Label (prior) shift** | 레이블 비율 | $P(y)$변화, $P(y\|x)$동일 |
| **Concept/conditional shift** | 조건부 관계 | $P(y\|x)$ 변화 |

# How Have We Been Solving It?

- ✓ Adversarial Robustness
  - 모델이 **의도적인 적대적 교란**에도 예측을 안정적으로 유지


- ✓ Domain Adaptation
  - 학습 데이터와 테스트 데이터의 **도메인 분포 차이**를 사전에 줄임

# Test-Time Training

# Main Idea

- ✓ 테스트 중에 모델을 학습 시킴
  - 라벨이 없는 테스트 입력에 대해 보조 과제를 수행하며 파라미터 업데이트
  - 사전 수집 불가능한 환경 변화(Distribution Shift)에 대해 즉각 적응
  - 별도의 추가 라벨링 없이도 성능 유지 또는 향상 가능
  - 분포 변화에 즉각 대응, 라벨 없이도 성능 보존

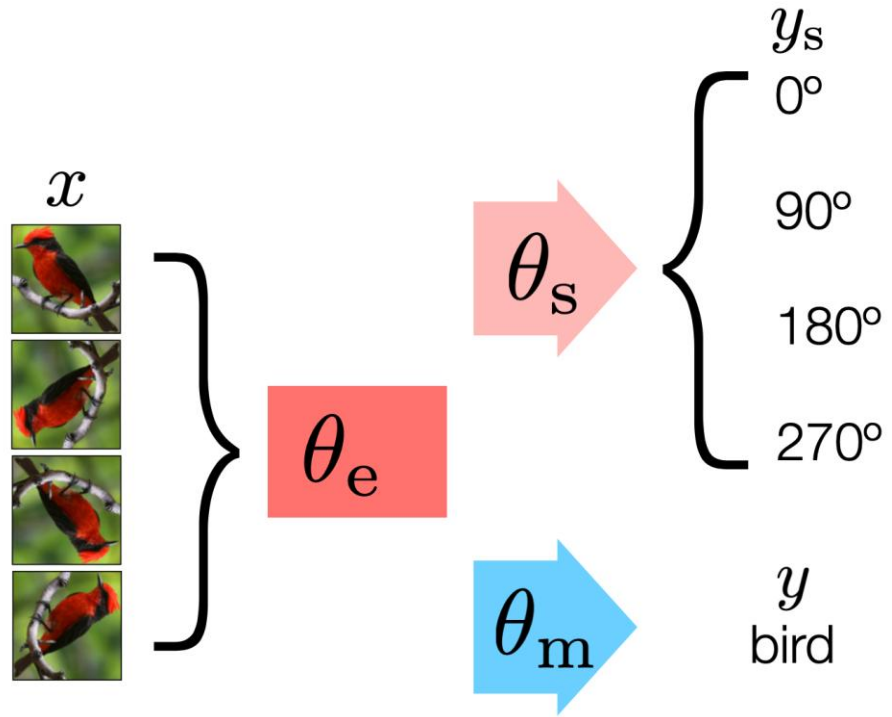
training

$$\min_{\theta_e,\theta_s,\theta_m} \mathbb{E}_P \left[ \begin{array}{l} \ell_m(x, y; \theta_e, \theta_m) \\ +\ell_s(x, y_s; \theta_e, \theta_s) \end{array} \right]$$

testing

$$\min_{\theta_e,\theta_s} \mathbb{E}_Q \left[ \ell_s(x, y_s; \theta_e, \theta_s) \right]$$

$\rightarrow \theta(x)$: make prediction on $x$



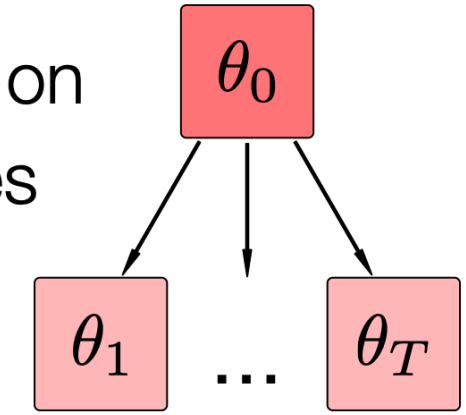
# Test-Time Training: Standard vs. Online

multiple test samples $x_1, \ldots, x_T$

$\theta_0$: parameters after joint training

**standard version**

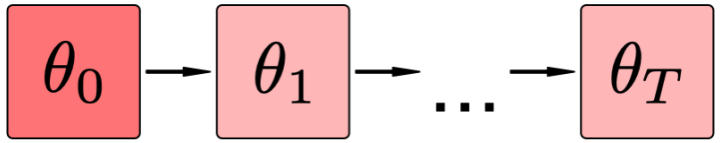
no assumption on the test samples




**online version**

$x_1, \ldots, x_T$ come from the same $Q$

or smoothly changing $Q_1, \ldots, Q_T$



# Batch Normalization

**Input:** Values of $x$ over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: $\gamma$, $\beta$

**Output:** $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

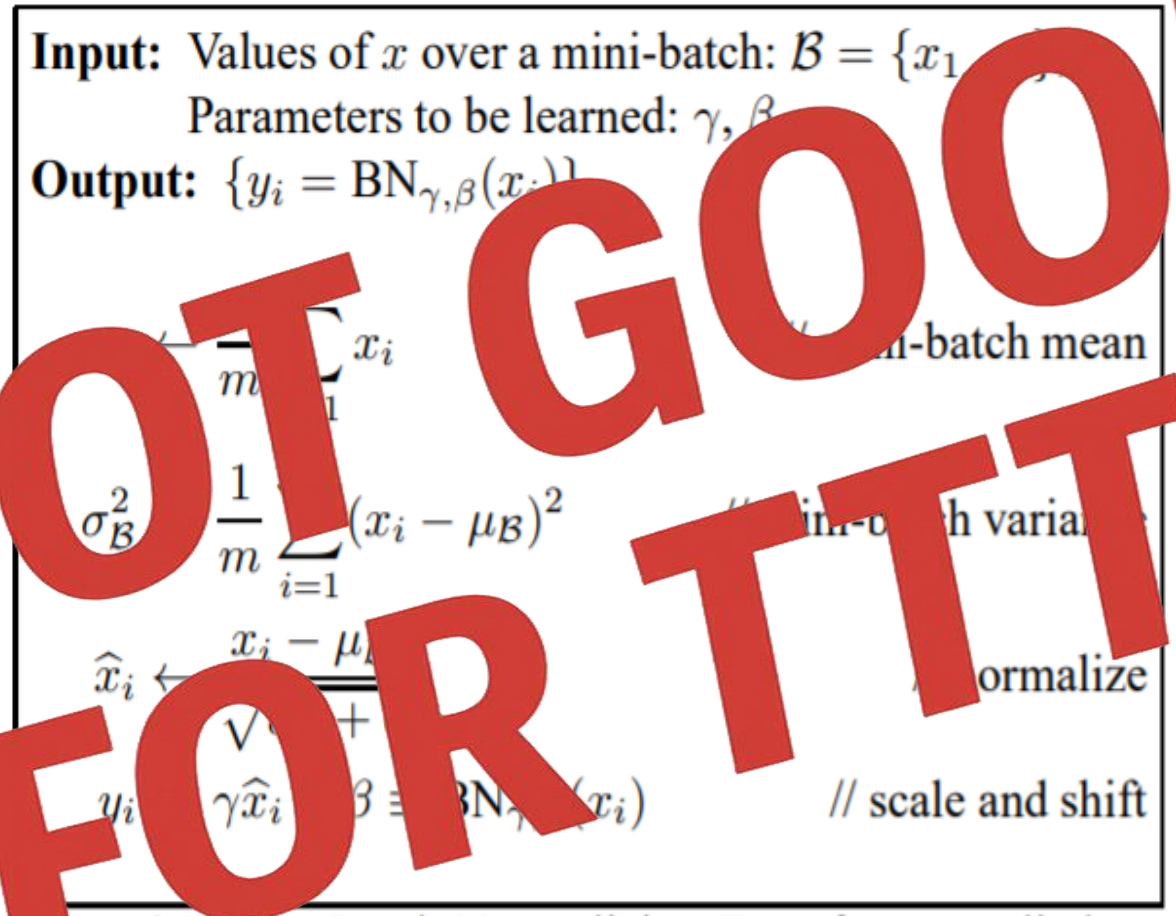
$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

**Algorithm 1:** Batch Normalizing Transform, applied to activation $x$ over a mini-batch.

# Batch Normalization



**Input:** Values of $x$ over a mini-batch: $\mathcal{B} = \{x_1$
Parameters to be learned: $\gamma$, $\beta$
**Output:** $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$ // mini-batch mean

$\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2$ // mini-batch variance

$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + \epsilon}}$ // normalize

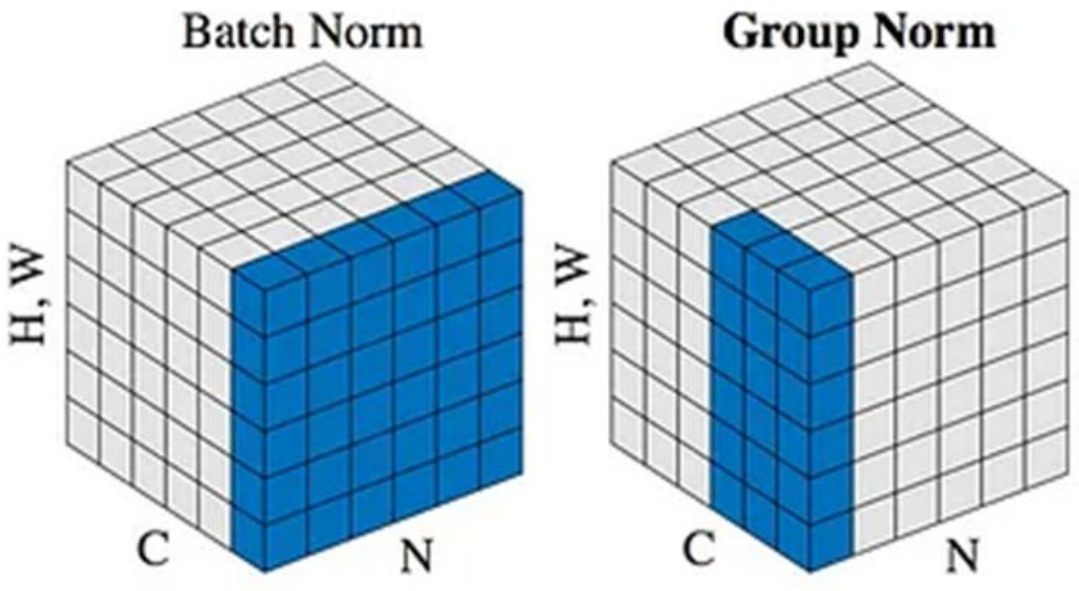
$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i)$ // scale and shift

**Algorithm 1:** Batch Normalizing Transform, applied to activation $x$ over a mini-batch.

NOT GOOD FOR TTT!

# Group Normalization

- Batch Size와 무관하게 동작
- 채널을 그룹 단위로 정규화 -> batch=1 에서도 안정적
- 학습/추론 모두 같은 방식으로 일관성 유지



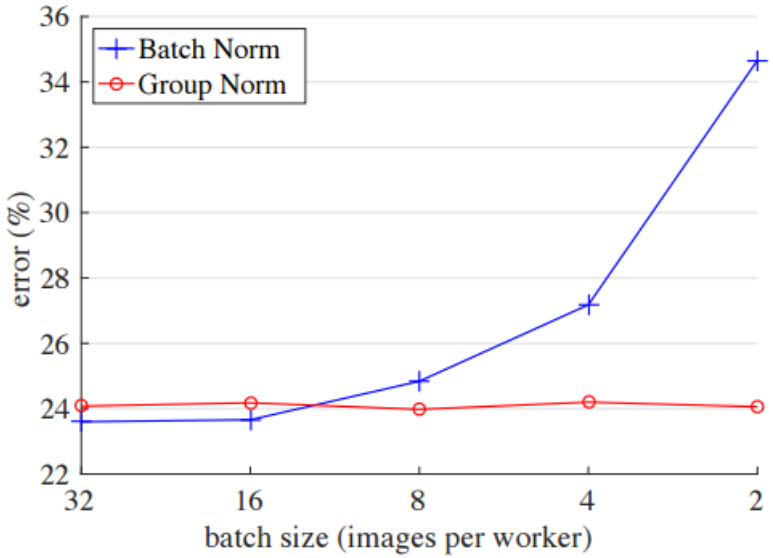
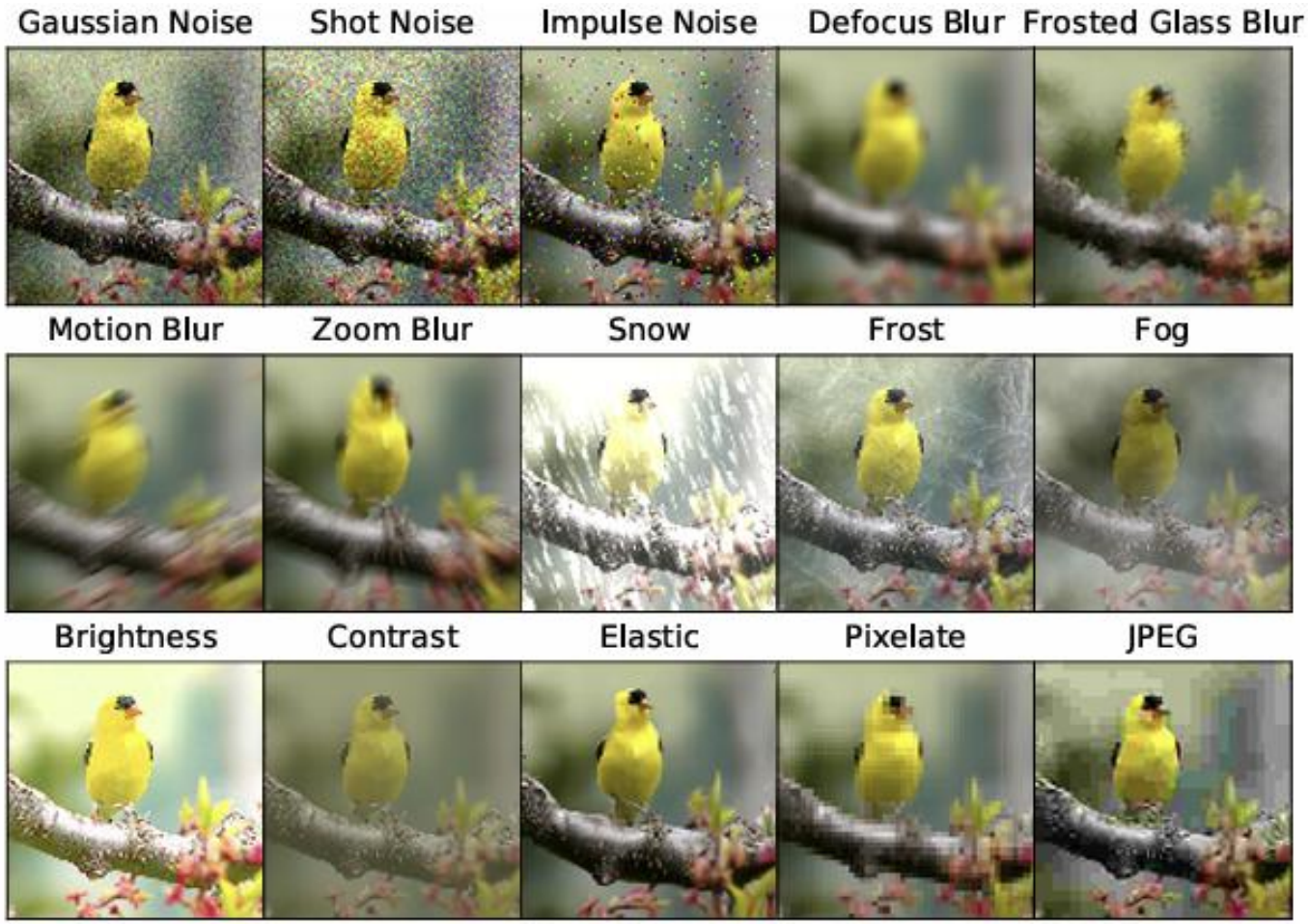




Figure 1. **ImageNet classification error *vs.* batch sizes**. This is a ResNet-50 model trained in the ImageNet training set using 8 workers (GPUs), evaluated in the validation set.

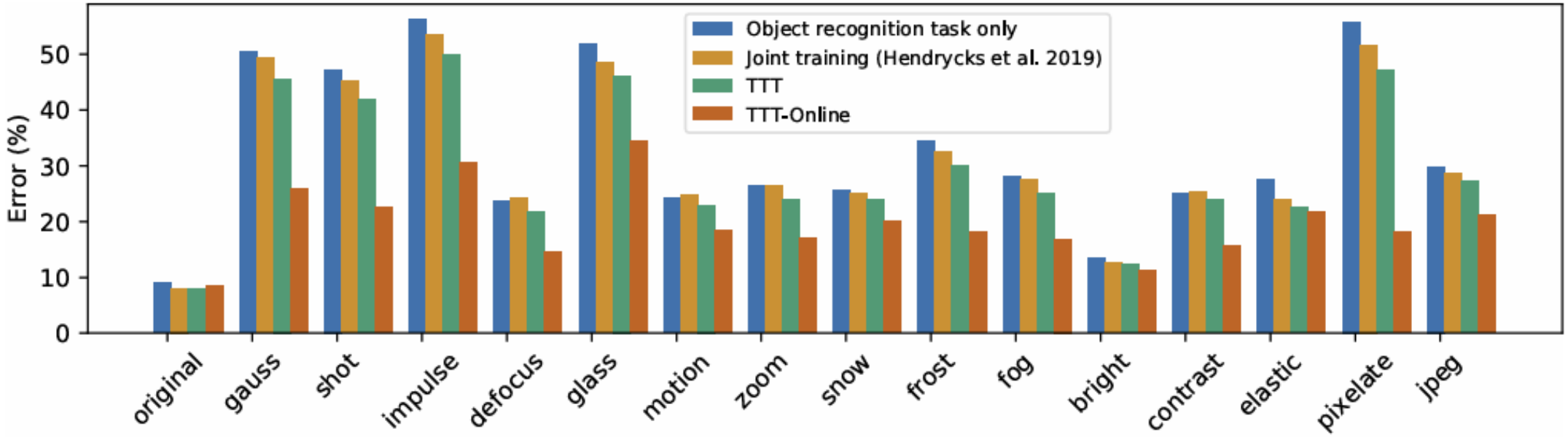
# Experiments

# Corrupted Image

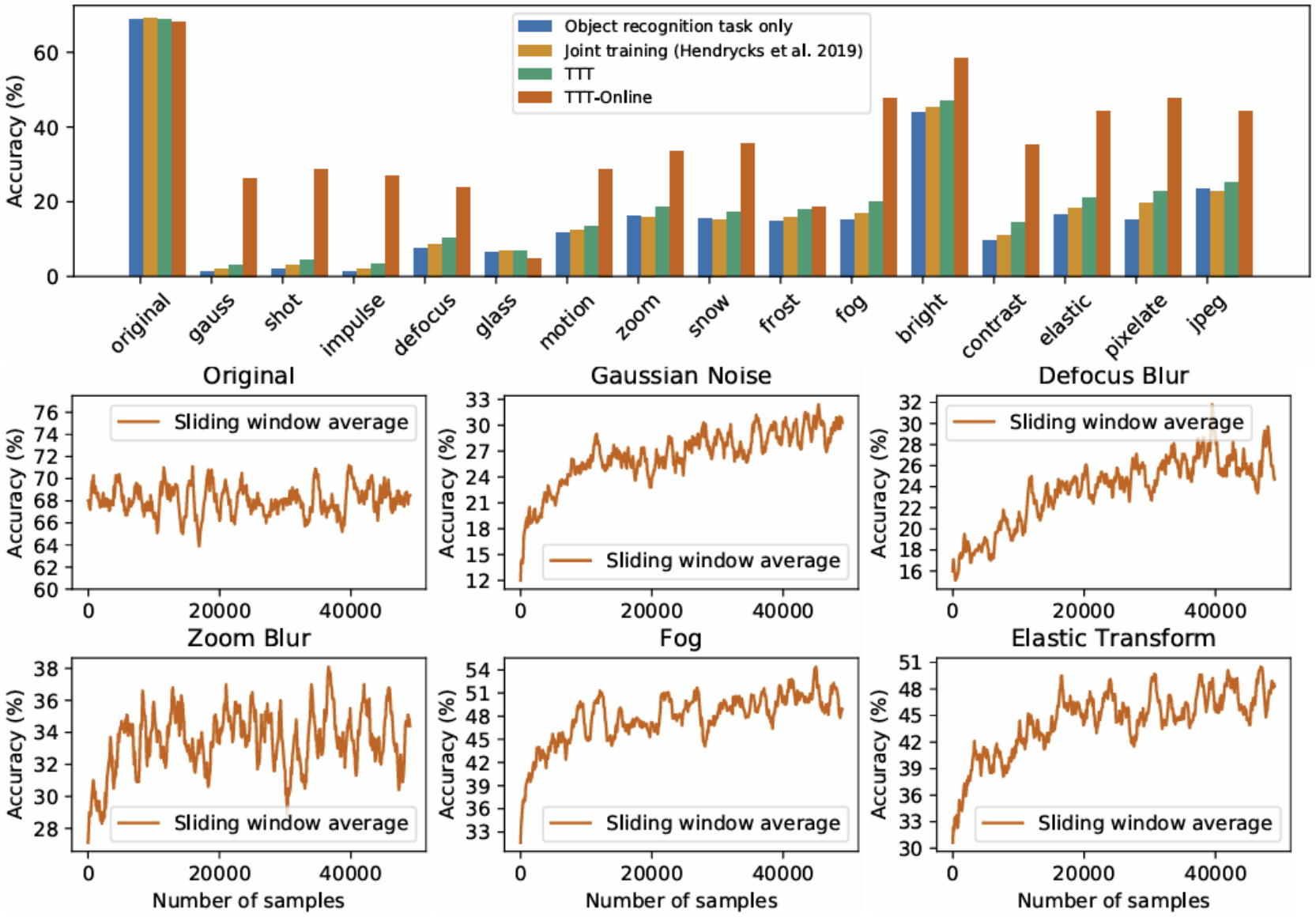
| Category | Corruption | 설명 |
|---|---|---|
| **Noise** | Gaussian Noise | 정규분포에서 뽑은 픽셀 노이즈 |
| | Shot Noise | 빛의 입자 수 변동(포아송 분포) |
| | Impulse Noise | Salt-and-Pepper, 픽셀 무작위 흑/백 변환 |
| **Blur** | Defocus Blur | 초점이 맞지 않아 전체적으로 흐림 |
| | Frosted Glass Blur | 유리창 너머처럼 랜덤한 흐림 |
| | Motion Blur | 카메라/피사체 움직임에 의한 직선 흐림 |
| | Zoom Blur | 줌 인/아웃 중 발생하는 방사형 흐림 |
| **Weather** | Snow | 눈 입자가 시야를 가리는 효과 |
| | Frost | 성에 낀 듯 일부 영역이 가려짐 |
| | Fog | 대비 저하 + 흐릿해짐 |
| **Digital** | Brightness | 전체 밝기 변화 |
| | Contrast | 색 대비 강도 변화 |
| | Elastic | 픽셀 변형, 물체가 휘어짐 |
| | Pixelate | 해상도 저하, 픽셀 크게 보임 |
| | JPEG Compression | 압축 아티팩트, 블록 노이즈 |



# CIFAR-10-C (ResNet-26)



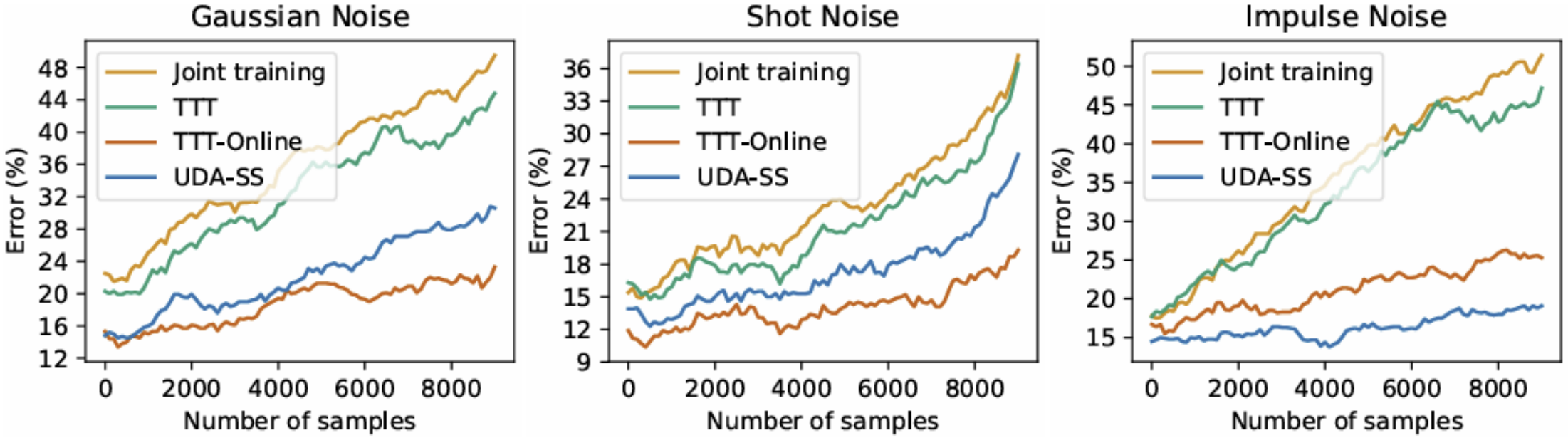
# ImageNet-C (ResNet18)



# vs. Unsupervised domain adaption

| | orig | gauss | shot | impul | defoc | glass | motn | zoom | snow | frost | fog | brit | contr | elast | pixel | jpeg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TTT-Online | **8.2** | **25.8** | **22.6** | 30.6 | **14.6** | **34.4** | **18.3** | **17.1** | **20.0** | **18.0** | **16.9** | **11.2** | 15.6 | **21.6** | **18.1** | **21.2** |
| UDA-SS | 9.0 | 28.2 | 26.5 | **20.8** | 15.6 | 43.7 | 24.5 | 23.8 | 25.0 | 24.9 | 17.2 | 12.7 | **11.6** | 22.1 | 20.3 | 22.6 |

# Gradually Changing Distribution Shifts



# Applications & Perspecitves

# Speech Enhancement

✓ 문제 배경
- 학습 시 보지 못한 잡음 환경 → 성능 급격히 저하
- 기존 도메인 적응법은 **타깃 잡음 데이터 필요**

✓ 제안 방법
- NSP (Masked Spectrogram Prediction)
- NyTT (Noisy-target Training)

✓ TTT 전략
- Standalone / Online / Online-batch / Bias-Only update

✓ 결과
- 합성 및 실제 잡음 환경 모두에서 성능 향상
- NyTT-real: 음질 보존 ↑
- NyTT-gaussian: 잡음 억제 ↑
- Bias-only 업데이트 → **엣지 디바이스 적용 가능성**

Behera, Avishkar, et al. "Test-Time Training for Speech Enhancement." *arXiv preprint arXiv:2508.01847* (2025).

# Depression Detection

✓ 문제 배경
- 우울증 탐지 모델은 **잡음, 성별, 데이터셋 차이**로 인한 분포 불일치에 취약
- 실제 환경 적용 시 성능 급락

✓ 제안 방법
- Audio-MAE 기반 TTT
- 테스트 시 입력 음성을 자기지도 학습으로 인코더 미세조정

✓ 실험 환경
- 분포 불일치 유형: 배경 잡음 / 성별 / 데이터셋

✓ 결과
- 모든 불일치 상황에서 성능 저하 최소화
- Macro F-score 지속적으로 개선
- Fine-tuning보다 안정적, **실제 환경 적용 가능성 입증**

Dumpala, Sri Harsha, et al. "Test-Time Training for Depression Detection." *arXiv preprint arXiv:2404.05071* (2024).

# Connections to Other Paradigms

| 학습 패러다임 | 핵심 개념 | TTT와의 연결점 |
|---|---|---|
| **Self-Supervised** | **라벨 없이**보조 과제로 표현 학습 | 회전 예측 등 **자기지도 과제**로 테스트 시 적응 |
| **Non-Adversarial Robustness** | 날씨·잡음 등 **자연스러운 분포 변화**에 견고 | **테스트 환경 변화**에 대응 |
| **Domain Adaptation (UDA)** | Source 라벨 + Target 무라벨로 **분포 맞추기** | **사전 Target 없이** 샘플 단위 적응 |
| **Domain Generalization** | 여러 도메인 학습 → **새로운 도메인 일반화** | **도메인 가정 없이** 테스트 시 적응 |
| **Few-shot** | 1~몇 개 샘플로 새로운 태스크 학습 | **단일 샘플(one-shot)**로 자기지도 적용 |
| **Continual Learning** | 순차적 학습, **기존 지식 유지** | 과거 샘플 기억 불필요 → **즉시 재적응** |

# conclusion

# Conclusion

✓ 핵심 아이디어
- 라벨 없는 **테스트 샘플을 자기지도 과제로 활용**
- 분포 변화 상황에서도 **즉시 적응** 가능

✓ 실험적 성과
- 다양한 벤치마크에서 **일관된 성능 향상**
- Joint Training, UDA보다 **더 안정적이고 일반화 우수**

✓ 의의와 확장성
- **고정된 모델 경계**를 넘어 테스트 시점 학습을 실현
- 다른 분야(Task)에도 적용될 수 있는 **확장 잠재력**

# 감사합니다