뇌 자기공명영상을 통한 치매 진단의 시각적 설명 기술

윤지석

wltjr1007@korea.ac.kr

고려대학교

기계지능연구실 (지도교수: 석흥일 교수)

설명 가능한 인공지능

- 인공지능의 **해석성 부여**
- 블랙박스 **인공지능 분해**
- **인간 수준의 사고** 모델링
  - 사후 가정 사고*를 하는 인공지능

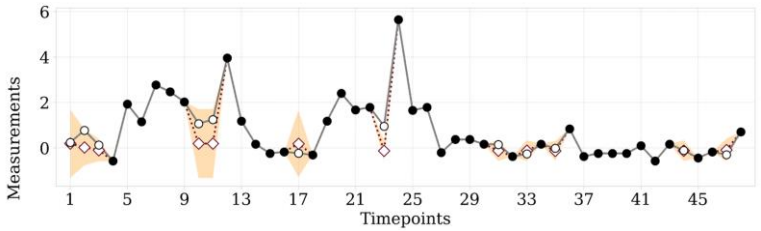
Interpretability

X

A

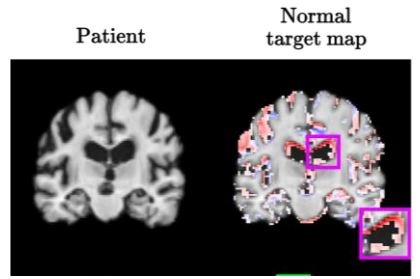
I

Explainability



전자의료기록

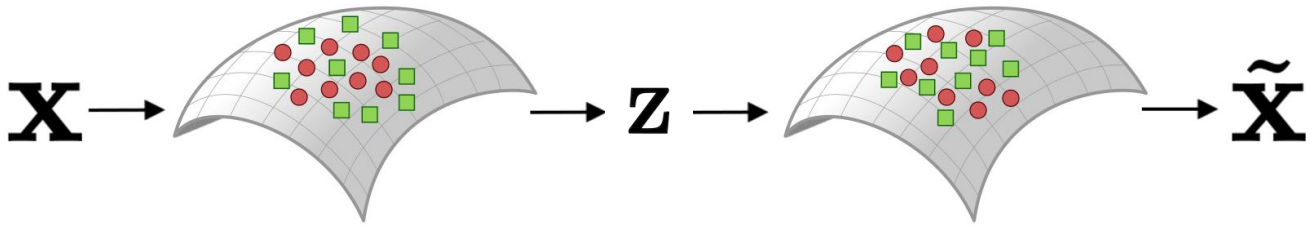
자연 영상



자기공명영상 (MRI)

*Counterfactual Thinking, 현실에서 일어난 일과 다른 상상을 통한 인간의 사고 방법

XAI Building Block: 인공지능의 해석성 부여



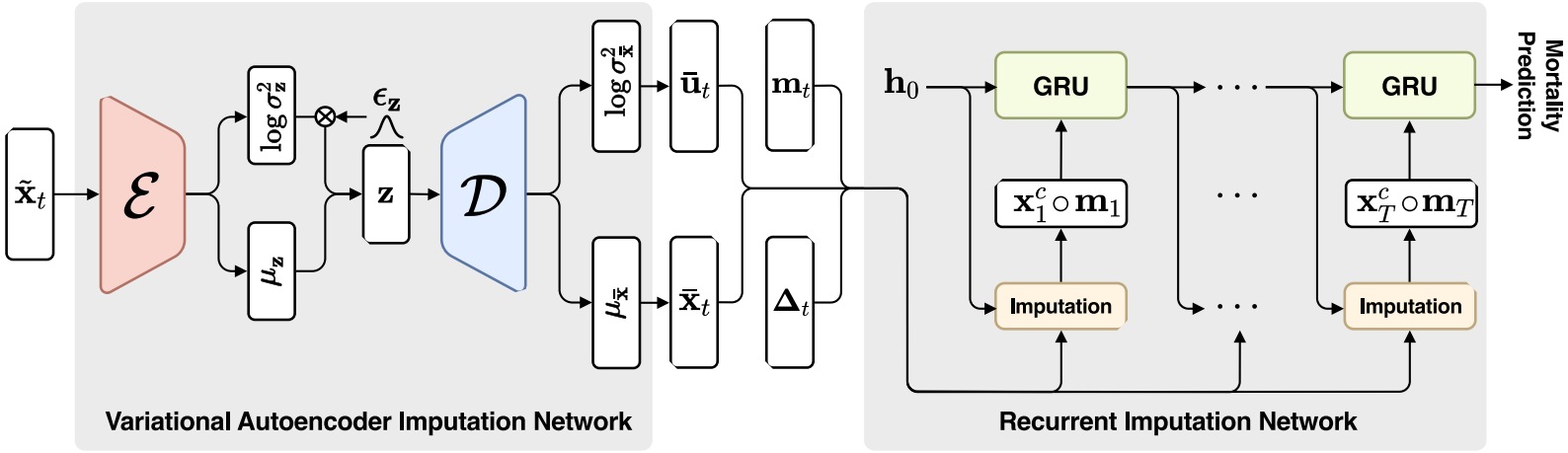
- Goal: 인공지능의 추론 과정 해석
- 즉, 추론 중 생성되는 feature embedding **z** 에 대한 해석

인공지능의 해석성 부여

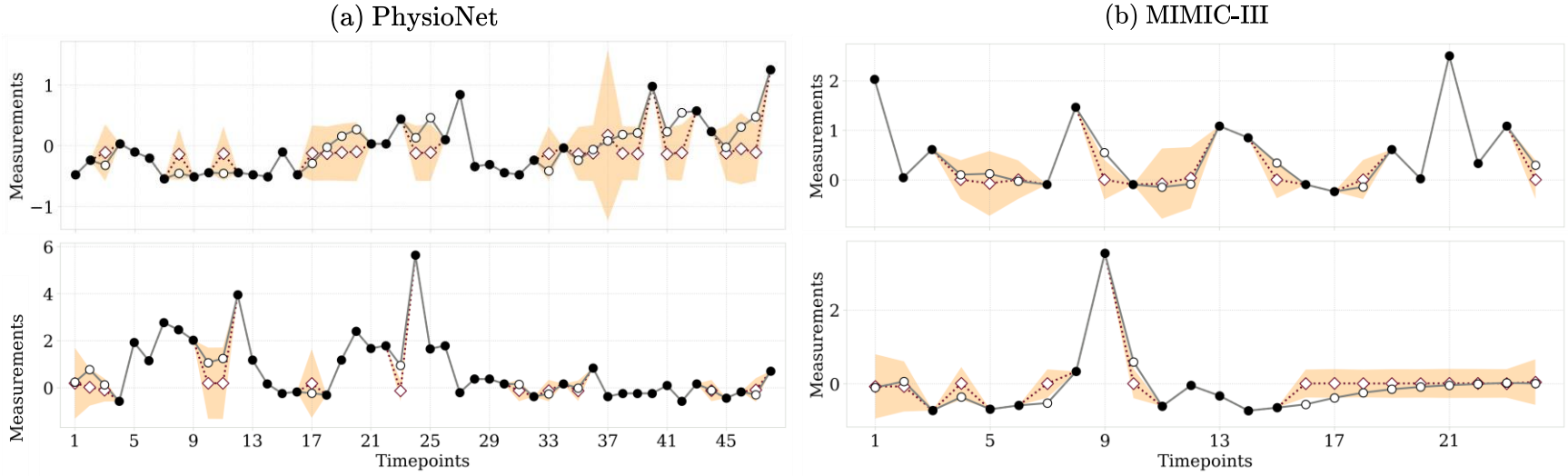
불확실성을 활용한 결측 데이터 대체 모델 개발

$$\mu_{\mathbf{z}} = \mathcal{E}_{\mu}(\tilde{\mathbf{x}}_t; \phi), \quad \log \sigma_{\mathbf{z}}^2 = \mathcal{E}_{\sigma}(\tilde{\mathbf{x}}_t; \phi)$$

$$\mu_{\bar{\mathbf{x}},t} = \mathcal{D}_{\mu}(\mathbf{z}; \theta), \quad \log \sigma_{\bar{\mathbf{x}},t}^2 = \mathcal{D}_{\sigma}(\mathbf{z}; \theta)$$

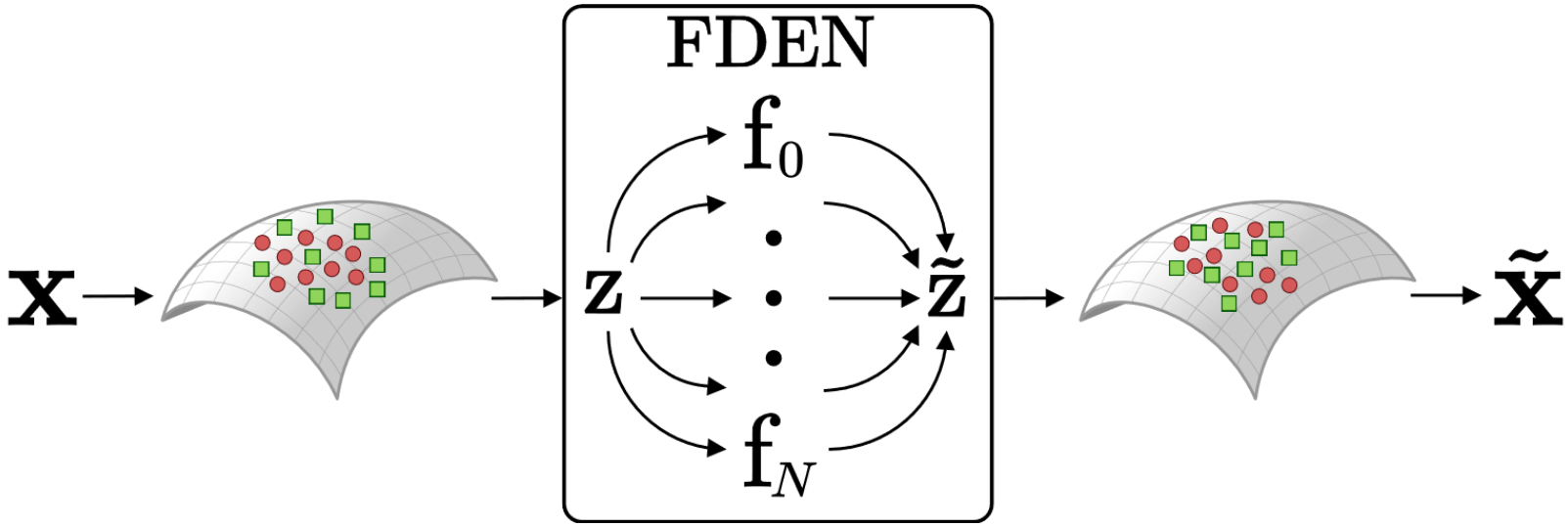


제안하는 모델



Mulyadi, Jun, and Suk, "Uncertainty-Aware Variational-Recurrent Imputation Network for Clinical Time Series," Submitted to IEEE Trans. On Cybernetics (IF 11.079).

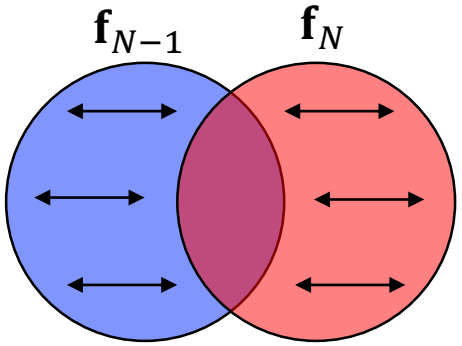
XAI Building Block: 인공지능의 분해



- Goal: 인공지능의 중요한 인자 분해 및 합성

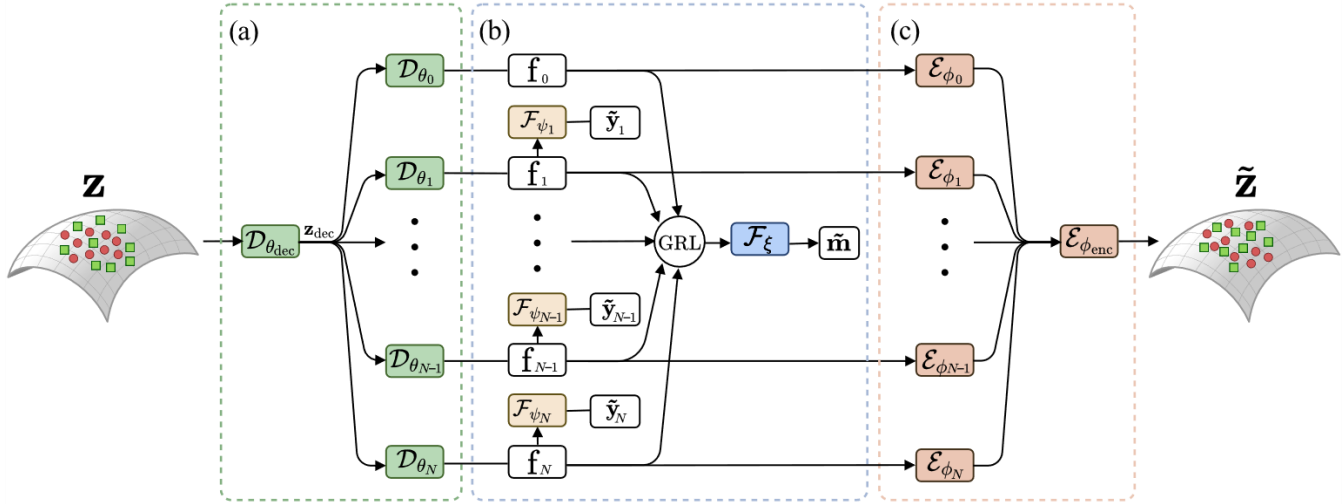
- 인공지능의 추론 과정 분해
  - 즉, feature embedding $\mathbf{z}$ → 인자 $\mathbf{f}_0, \ldots, \mathbf{f}_N$로 분해
  - 예: $\mathbf{f}_0$ - 눈, $\mathbf{f}_1$ - 코, …, $\mathbf{f}_N$ - 머리 색깔

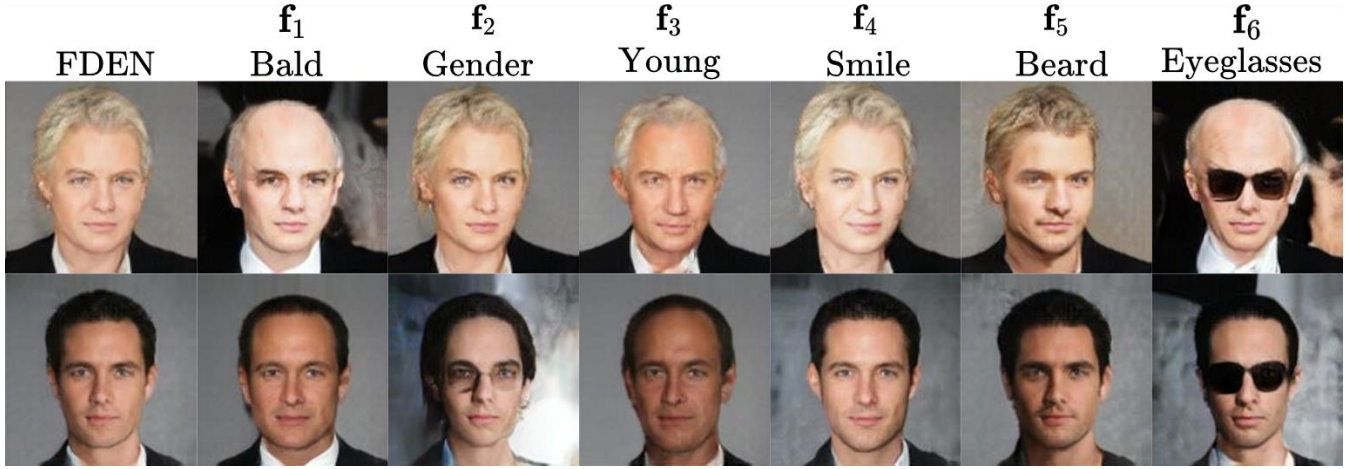
인공지능의 분해



Mutual Information 통한 인자들의 교집합 최소화

$$\mathcal{L}_M = \sup_{\xi} \mathbb{E}_{\mathbb{P}_{0,\ldots,N}} [\mathcal{F}_\xi] - \log \left(\mathbb{E}_{\mathbb{P}_0 \otimes \ldots \otimes \mathbb{P}_N} [\exp (\mathcal{F}_\xi)] \right)$$





인공지능의 추론 과정 인자 단위로 분해

인자를 통해 인공지능의 설명 가능성 부여

Machine Intelligence Lab

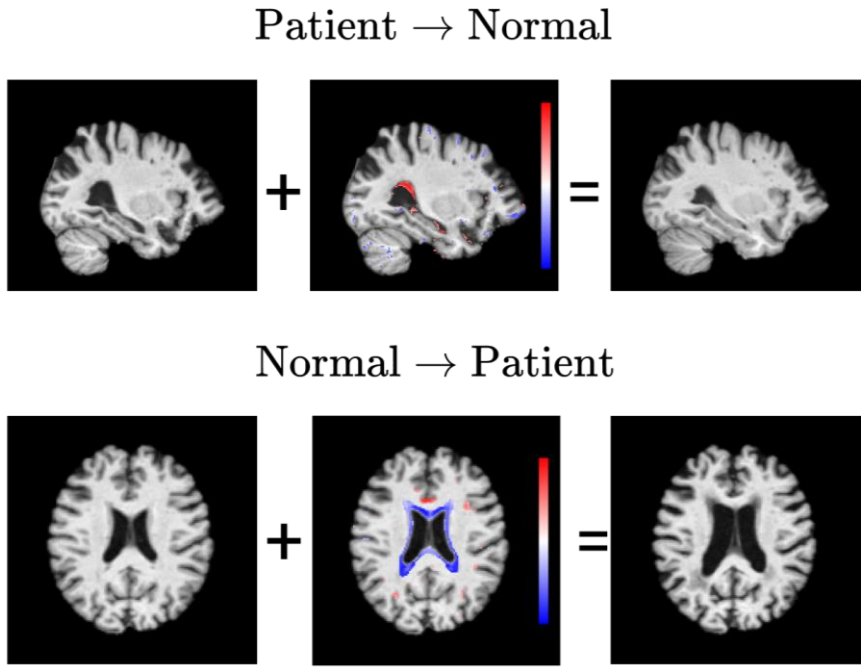
xaic eXplainable Artificial Intelligence Center

Yoon, Roh, and Suk, "A Plug-in Method for Representation Factorization in Connectionist Models," Submitted to IEEE Trans. Neural Networks and Learning Systems (IF 8.793).

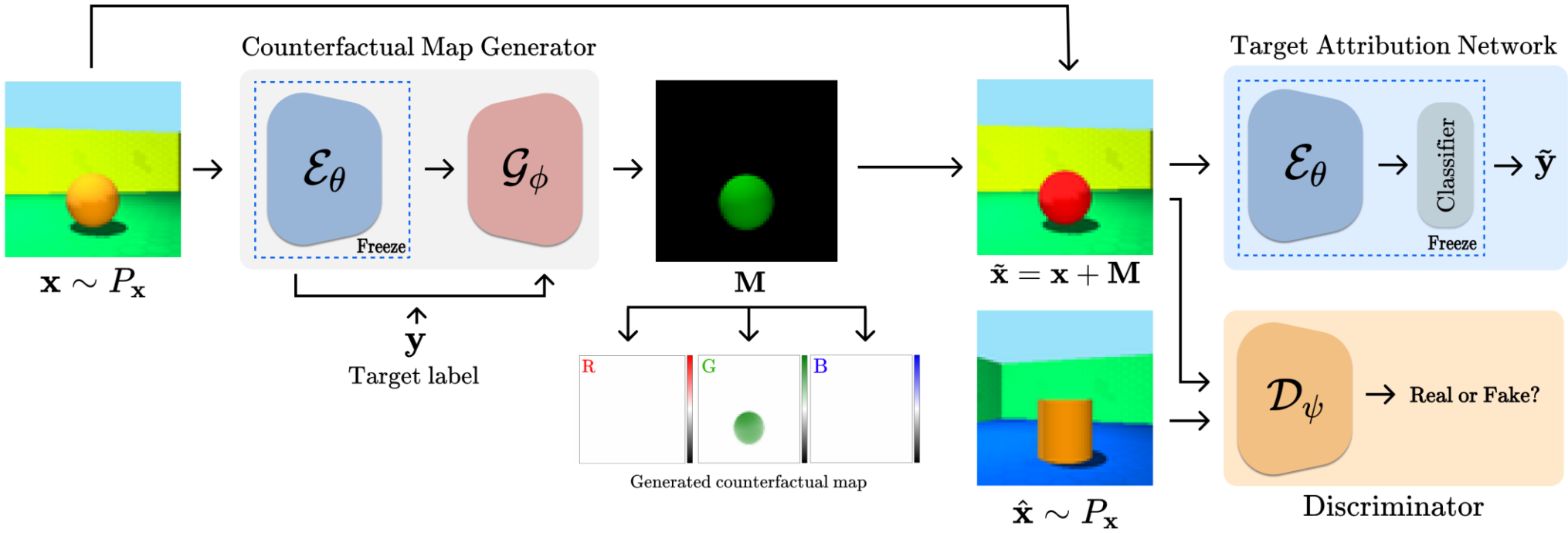
인간 수준의 사고 모델링

- 사후 가정 사고 (Counterfactual Reasoning)
  - 현실에서 일어난 일과 다른 상상을 통한 인간의 사고 방법

- 예: 정상인의 뇌영상에서 어떤 이상 증상이 나타나야 치매로 진단 될까...?

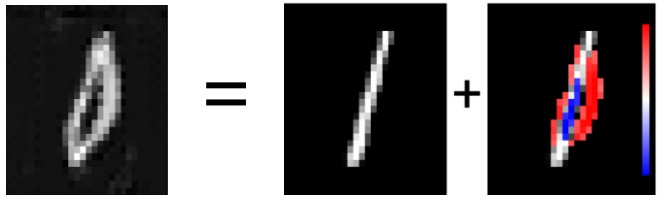


인간 수준의 사고 모델링



$$\mathbf{M}_{\mathbf{x},\mathbf{y}} = \mathcal{G}_\phi\left(\mathcal{E}_\theta(\mathbf{x}), \mathbf{y}\right)$$

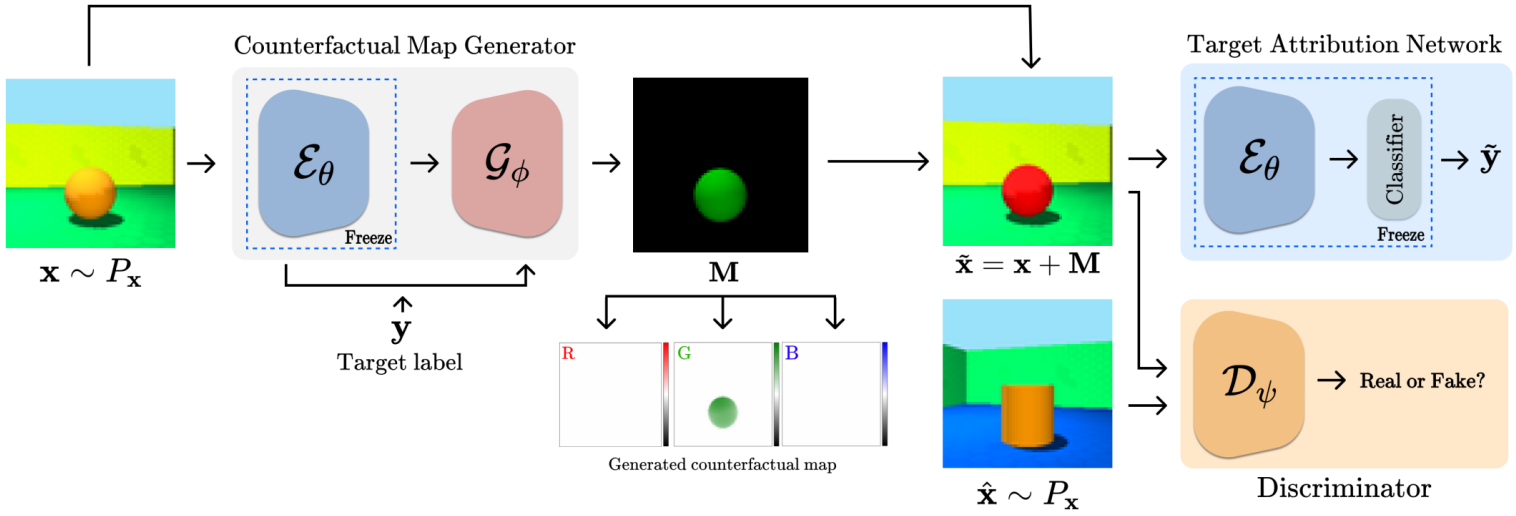
$$\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{M}_{\mathbf{x},\mathbf{y}}$$



Machine Intelligence Lab

xaic eXplainable Artificial Intelligence Center

인간 수준의 사고 모델링



Target Attribution Loss

$$\mathcal{L}_{cls} = \mathbb{E}_{\mathbf{x}\sim P_\mathbf{x}, \mathbf{y}\sim P_\mathbf{y}}[CE(\mathbf{y}, \tilde{\mathbf{y}})]$$

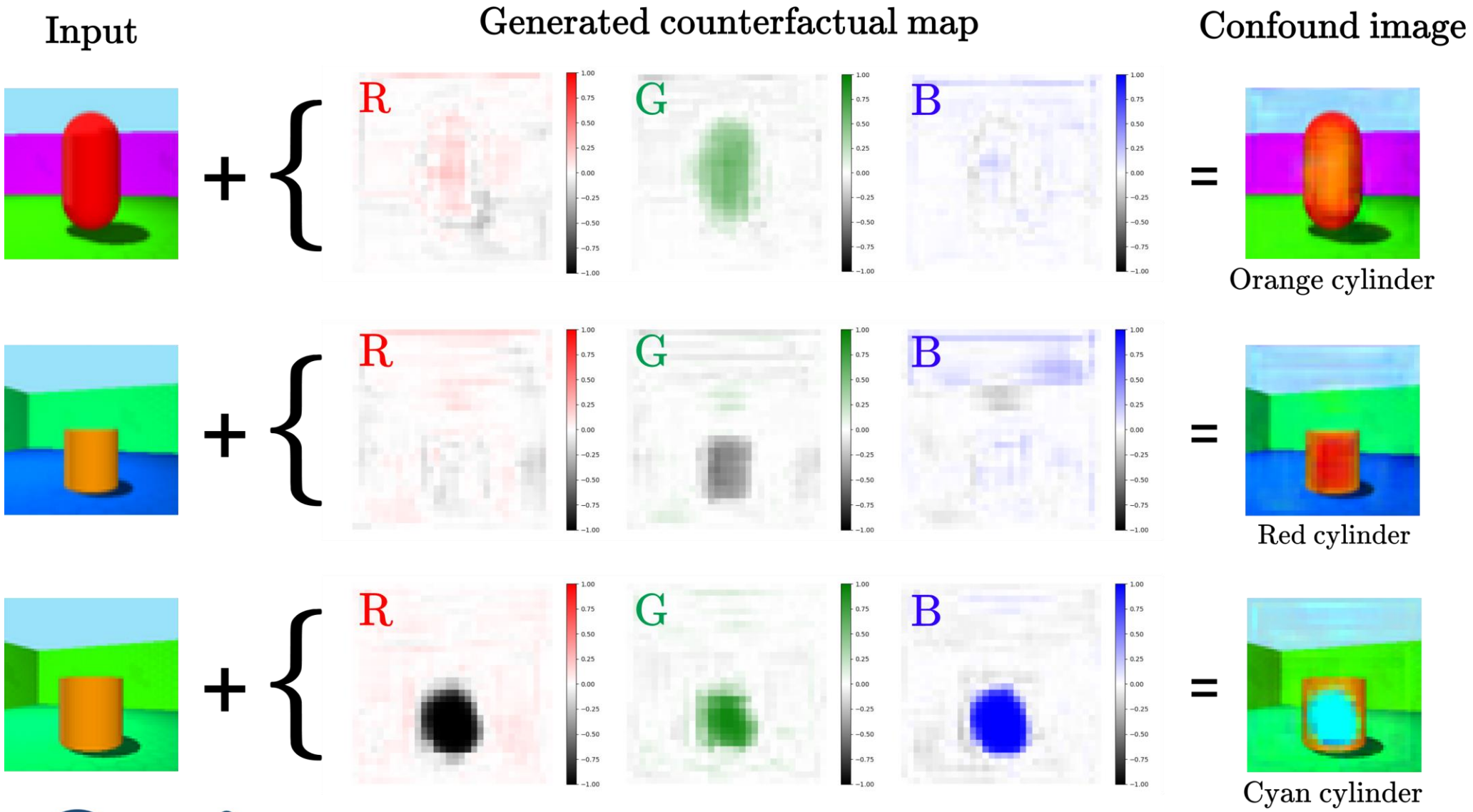
Counterfactual Map Loss

$$\mathcal{L}_{map} = \mathbb{E}_{\mathbf{x}\sim P_\mathbf{x}, \mathbf{y}\sim P_\mathbf{y}}[\lambda_1 \|\mathbf{M}_{\mathbf{x},\mathbf{y}}\|_1 + \lambda_2 \|\mathbf{M}_{\mathbf{x},\mathbf{y}}\|_2]$$

Machine Intelligence Lab

xaic eXplainable Artificial Intelligence Center

인간 수준의 사고 모델링

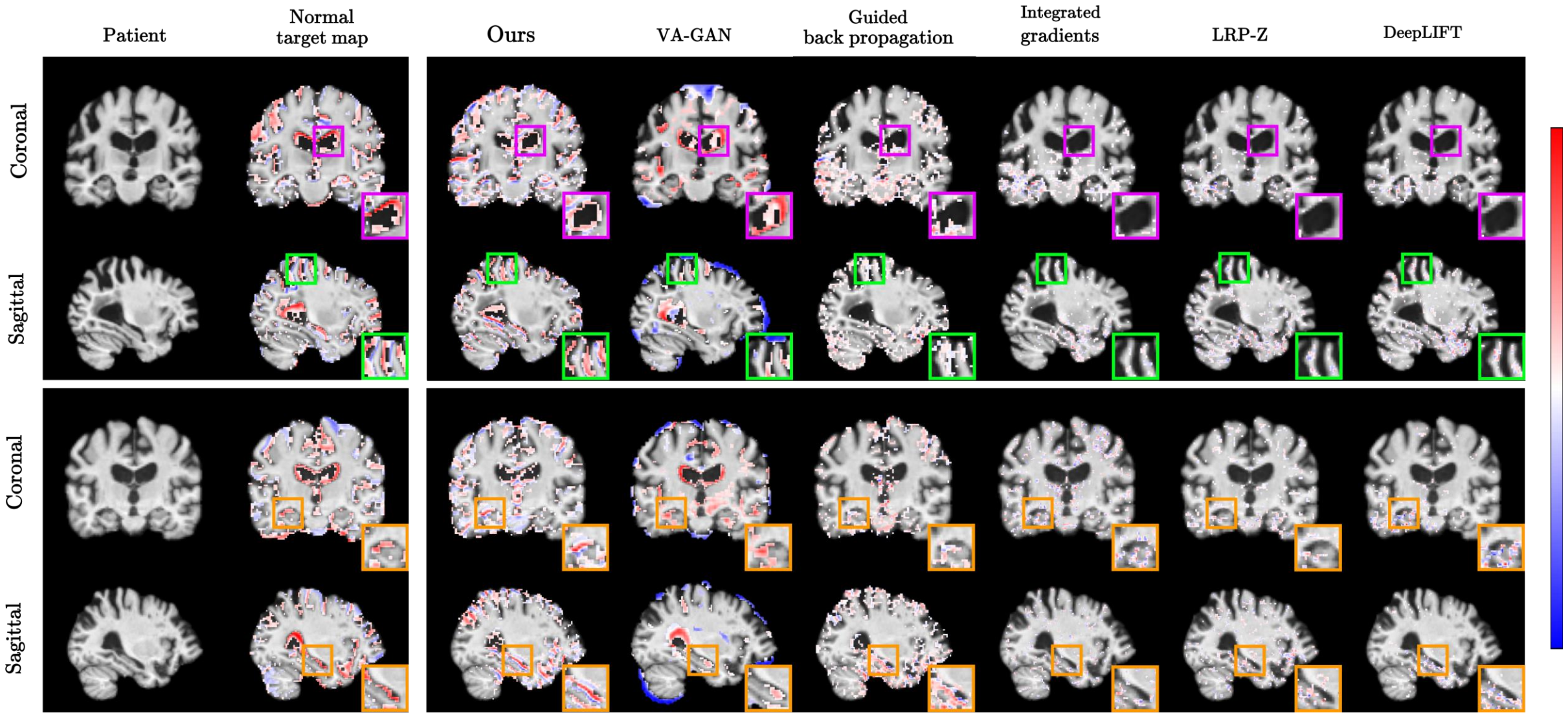


Machine Intelligence Lab

xaic
eXplainable Artificial Intelligence Center

Machine Intelligence Lab

xaic eXplainable Artificial Intelligence Center



Machine Intelligence Lab

xaic
eXplainable Artificial Intelligence Center

감사합니다

https://milab.korea.ac.kr

https://www.jsyoon.kr