Paper 1

RILS LAB

Robot Intelligence Learning & System

최성용

2026. 07. 23

01

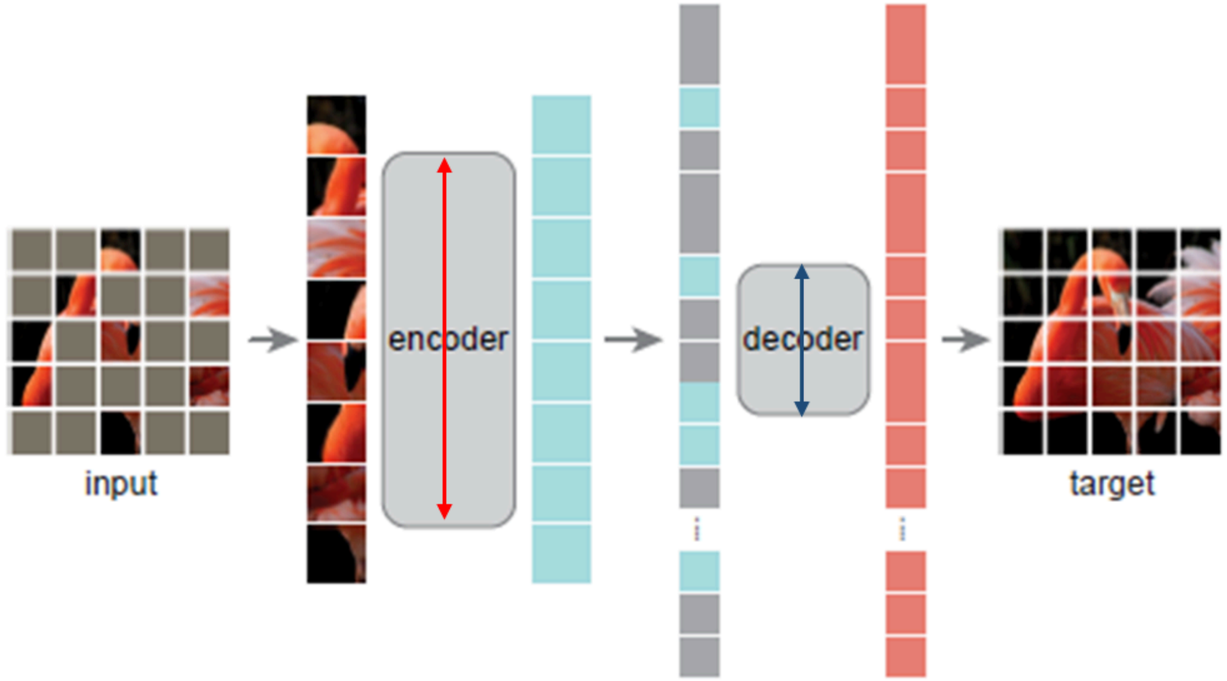
In Pursuit of Pixel Supervision for Visual Pre-training

Lihe Yang, Shang-Wen Li, Yang Li, Xinjie Lei,
Dong Wang, Abdelrahman Mohamed, Hengshuang Zhao, Hu Xu

CVPR / 2026

MAE

이미지의 **일부를 가리고**, 보이는 부분만 이용해 가려진 **픽셀을 복원**하도록 학습하는 **자기지도학습** 방법



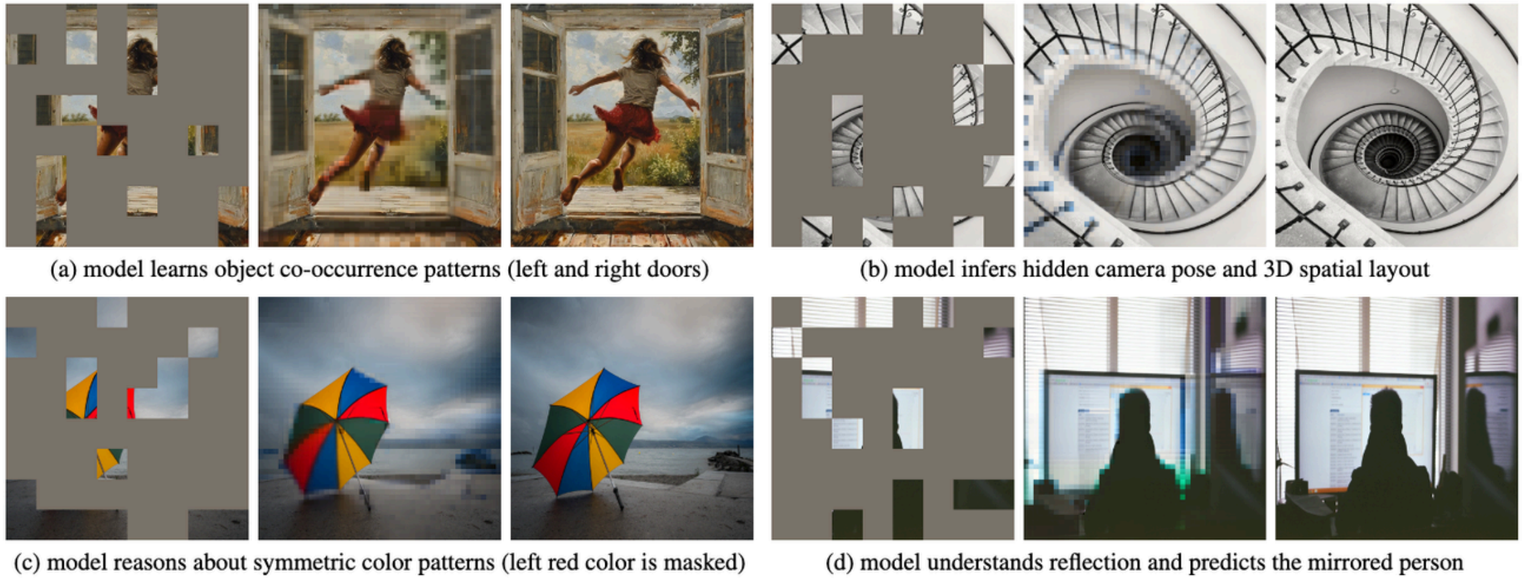
- Encoder: 이미지에서 중요한 시각 정보를 추출해 feature 벡터로 변환(보이는 패치만 처리)
- Decoder: 가려진 픽셀 복원

- Patch tokens: 이미지의 지역 정보(각 패치를 벡터로 변환)
- Class tokens: 이미지 전체의 전역 정보

1. Problem Setting

Vision Representation Learning 흐름:

Raw Images
↓
Self-Supervised Learning
↓
Visual Representation
↓
Downstream Tasks:
Depth · 3D · Segmentation · Robot Learning



(a) model learns object co-occurrence patterns (left and right doors)

(b) model infers hidden camera pose and 3D spatial layout

(c) model reasons about symmetric color patterns (left red color is masked)

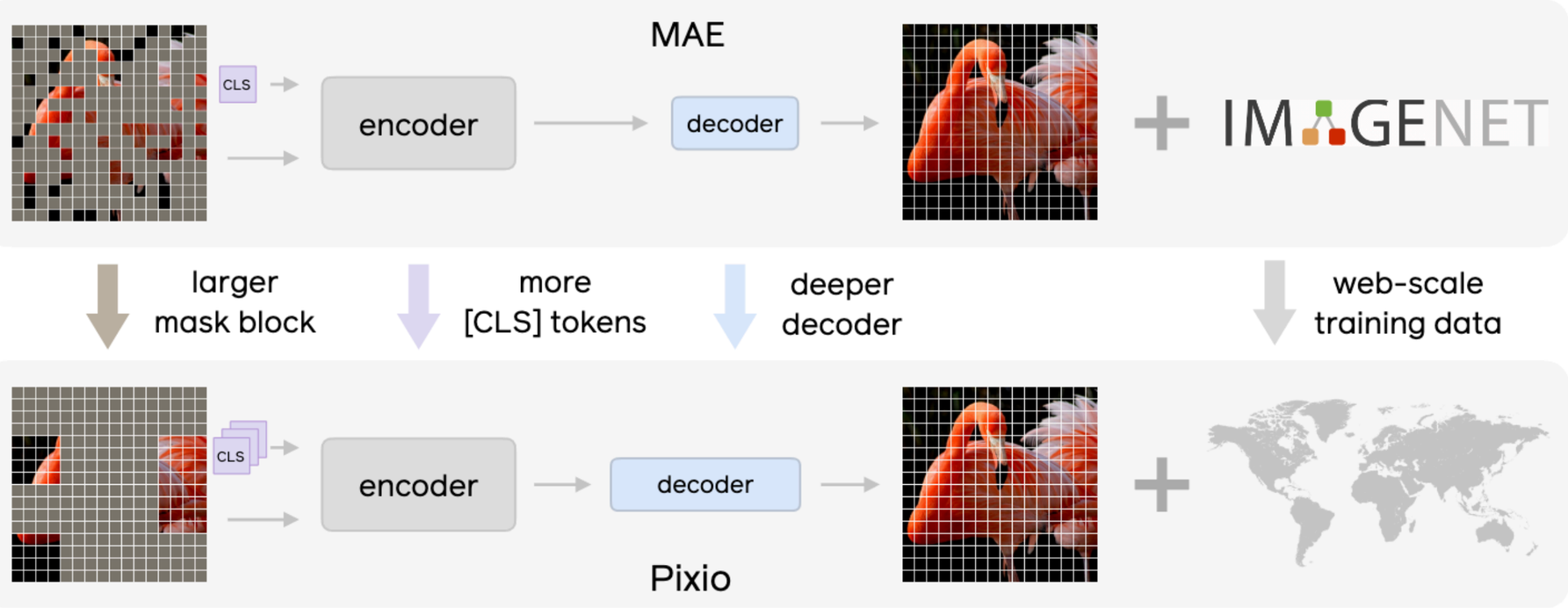
(d) model understands reflection and predicts the mirrored person

Problem

- 최근에는 latent-space learning이 주류(DINO, JEPA)
- Pixel-space learning은 상대적으로 덜 주목받음(MAE)
- 기존 MAE는 좋은 범용 표현을 학습하기에 모델 설계가 충분히 강력하지 않을 수 있음

📌 **Q: 픽셀을 직접 복원하는 것(pixel-space learning)만으로도 범용적인 시각 표현을 학습할 수 있는가?**

2. Solution: Pixio



Model 개선

1. Larger mask blocks
- 주변 패치를 복사하는 shortcut을 줄이고 더 넓은 문맥을 학습

2. Deeper decoder
- 픽셀 복원은 decoder가 담당하도록 해, encoder는 표현학습에 집중

3. Multiple class tokens
- 서로 다른 전역 정보와 시각적 속성을 분산해 표현

Dataset 개선

Raw Web Data
↓
Self-Curation
↓
Pre-training

3. Contribution

다양한 downstream task에서 경쟁력 확인

- Monocular depth estimation
- Feed-forward 3D reconstruction
- Semantic segmentation
- Robot learning

논문의 세 가지 기여

1. Pixel supervision의 재평가
   - 픽셀 복원만으로도 범용 시각 표현 학습 가능성 제시
2. MAE의 간단한 개선
   - 더 어려운 masking, 더 강한 decoder, 여러 class token
3. 데이터와 표현학습의 새로운 방향
   - 대규모·다양한 데이터와 최소한의 self-curation 활용

핵심 메시지: Pixel-space learning은 latent-space learning의 대안이자 보완재가 될 수 있다

02

INSID3: Training-Free In-Context Segmentation with DINOv3

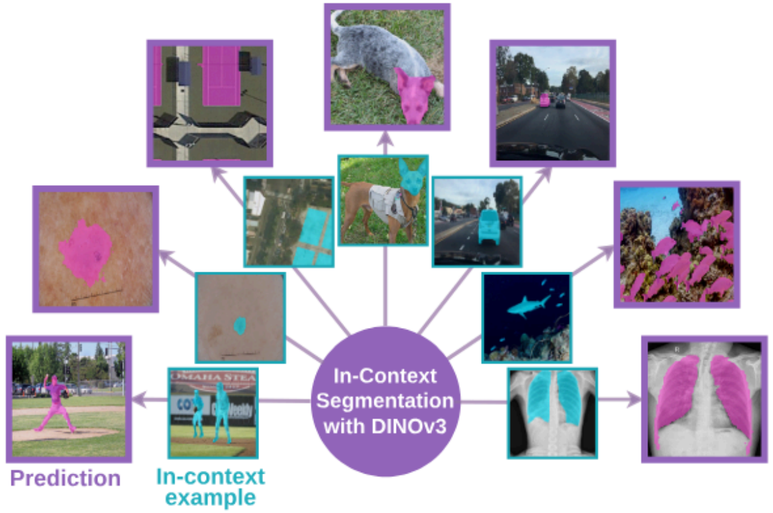
Claudia Cuttano, Gabriele Trivigno, Christoph Reich, Daniel Cremers,
Carlo Masone, Stefan Roth

CVPR / 2026

In-context Segmentation(ICS)

일반적인 class-based segmentation

- 학습: 고양이 이미지와 고양이 mask를 대량으로 사용
- 추론: 새로운 이미지에서 “고양이” class를 분할



- Input:
  - Reference image
  - Annotated mask
  - Target image
- Output:
  - Target image에서 동일한 개념의 segmentation mask

ICS

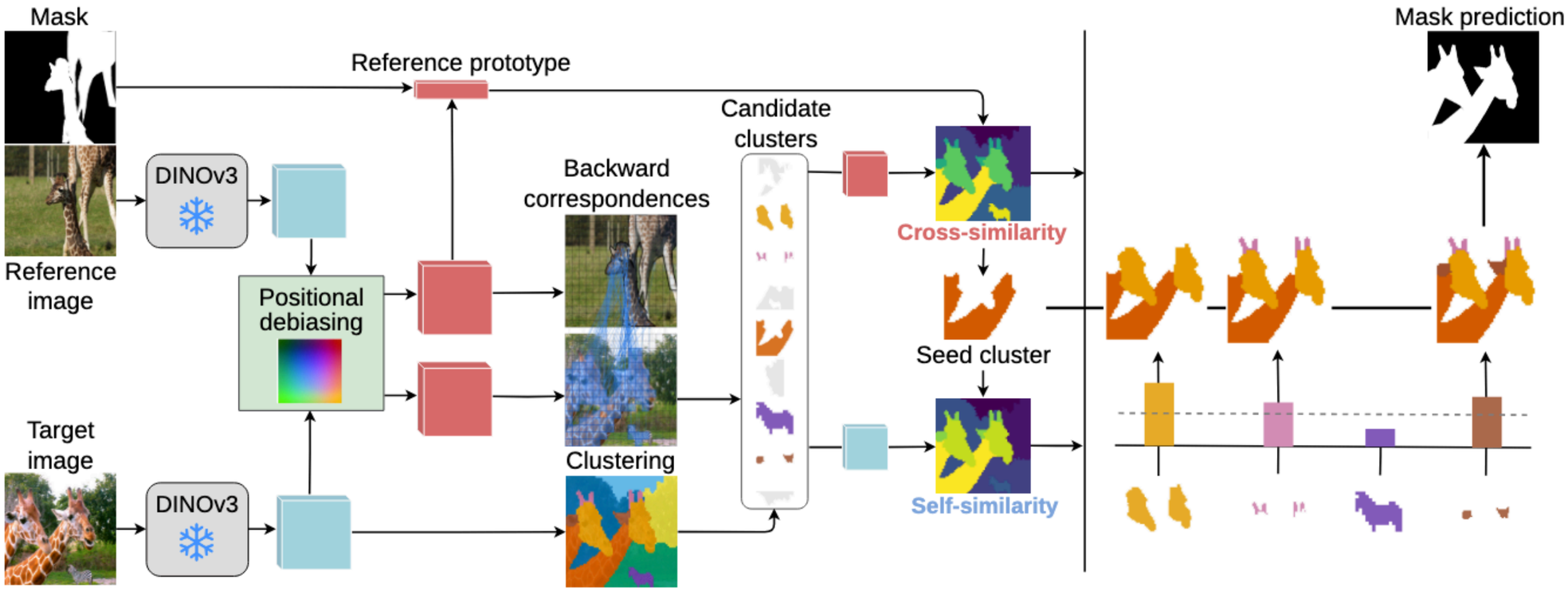
- 모델을 새로운 데이터로 추가 학습시키지 않고, inference 단계에서 mask가 포함된 예시 이미지 한 장을 제공하여 타겟 이미지에서 같은 개념을 찾아 분할하도록 함.
- 단, 특정 class에 대한 fine-tuning을 하지 않는 것이지, DINOv3와 같은 Backbone은 pre-training 되어 있음.

1. Problem Setting

Problem

- 기본적인 Object뿐 아니라 part / personalized instance까지 처리해야함.
- 새로운 도메인과 개념에 대한 일반화 필요
- 기존 방법의 한계
  - Fine-tuning: 높은 in-domain 성능, but 낮은 generalization
  - DINOv2 + SAM: 강한 generalization, but 복잡한 multi-model 구조

2. Solution(Methodology)



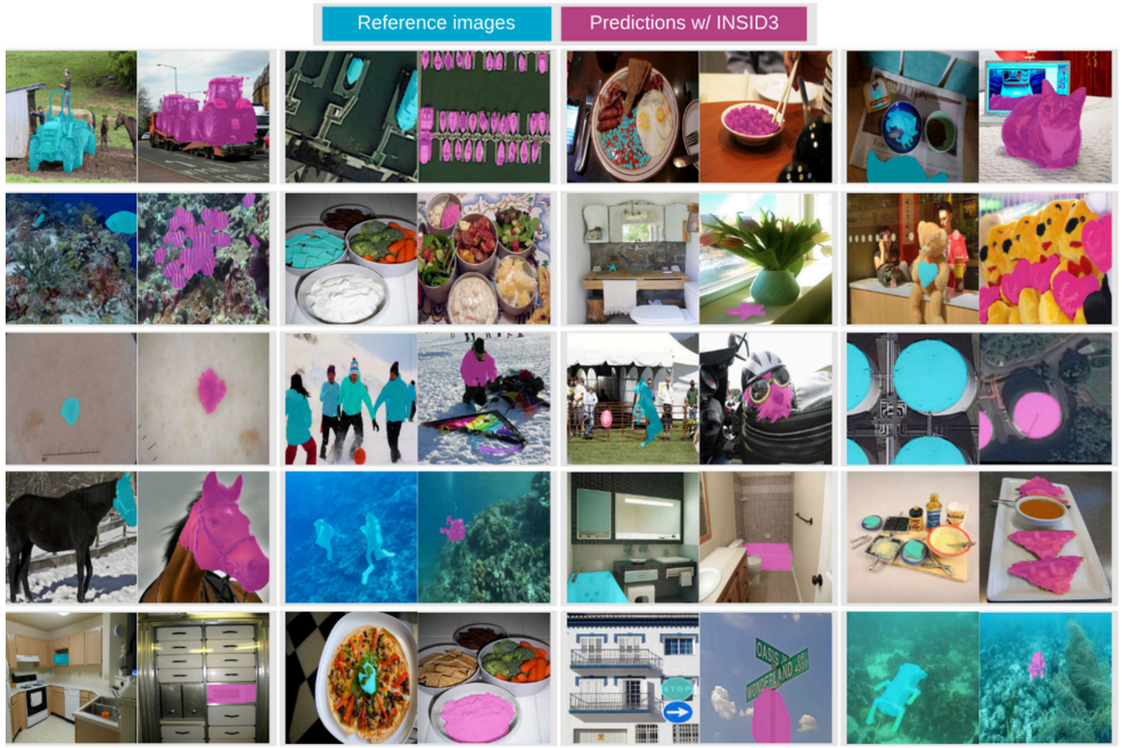
유사도	비교대상	결과
Cross-image similarity	Reference prototype ↔ Candidate cluter prototype	Seed Cluster(가장 유사도가 높은 cluster)
Self-similarity	Seeld Cluster ↔ Targert image Cluster	-

최종 cluster score

$$S_k = s_k^{\text{cross}} \cdot s_k^{\text{intra}}$$

→ threshold α 이상인 cluster를 seed와 합침

3. Contribution



1. Training-free ICS

- Frozen DINOv3 only
- No decoder
- No fine-tuning
- No mask-level supervision

2. Strong Generalization

- Semantic segmentation
- Part segmentation
- Personalized segmentation
- Medical / aerial / underwater domains

3. Efficient and Extensible

- 304M parameters
- 302 ms/image
- 5-shot으로 자연스럽게 확장
- Empty-mask prediction 가능

Thank you

Question & Disscusion

RILS LAB