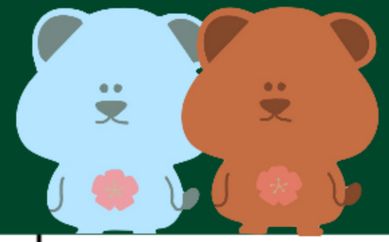


Masked Autoencoders Are Scalable Vision Learners

Research 정은채

EWHA EURON

목차

#01 Introduction

#02 Related Work

#03 제안 방법론

#04 실험 및 결과

#05 결론 & Discussion

EWHA EURON

Introduction

EWHA EURON

#1-1. 컴퓨터 비전에서의 표현 학습 한계

“컴퓨터 비전은 왜 항상 라벨이 있어야만 학습이 되는가?”

“자연어 처리처럼, 라벨 없이도 표현을 학습할 수는 없을까?”

🔷 기존 컴퓨터 비전의 흐름

- 대부분 감독 학습(supervised learning)에 의존
- 성공 사례: CNN 기반의 ImageNet 사전학습 (ResNet, EfficientNet 등)
- 모델이 좋아질수록 **더 많은 라벨링 데이터**가 필요

🔷 현실적인 문제

- **수백만~수억 장의 이미지 라벨링**은 고비용 & 시간 소모
- 프라이버시, 저작권, 도메인 한계 등으로 **데이터 확장성 제한**
- 예: ViT 같은 대형 모델은 **1M 이미지로 과적합**,

→ JFT-300M 같이 외부 대규모 데이터 필요 (대부분 비공개)

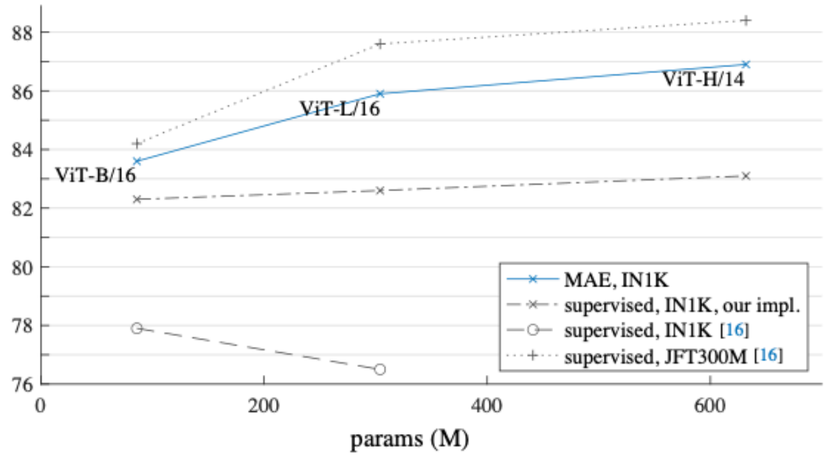


Figure 8. **MAE pre-training *vs.* supervised pre-training**, evaluated by fine-tuning in ImageNet-1K (224 size). We compare with the original ViT results [16] trained in IN1K or JFT300M.

#1-2. NLP의 성공적인 Self-Supervised Learning 사례

🔷 NLP의 대전환: Self-supervised Learning

- 전통적으로는 번역, 개체명 인식 등에 **라벨링 데이터** 필요
- 하지만 지금은 대부분의 NLP 모델이 **자가 지도 학습(pretraining) → 파인튜닝** 구조

🔷 대표 사례 1: BERT (2018, Google)

- 입력 문장에서 일부 단어를 [MASK]로 가림
- 문맥 기반 복원 (Masked Language Modeling)을 통해 의미 파악
- 이후 다양한 언어 과제에 파인튜닝 → **폭발적인 성능 향상**

🔷 대표 사례 2: GPT 시리즈 (2018~2023)

- 문장의 앞 단어들을 보고 **다음 단어 예측** (Autoregressive 방식)
- 수천억 토큰 학습 → **인간 수준의 언어 생성 가능**
- 현재 대부분의 생성형 AI는 이 구조 사용

"단어를 복원하는 간단한 과제 하나만으로,
복잡한 의미 표현과 세계 지식을 학습할 수 있다"

Vision 분야로 확장될 수 있을까?

"NLP처럼, 이미지도 일부를 가리고 복원하면
강력한 시각 표현을 학습할 수 있을까?"

✅ **GPT / BERT 학습 방식 비교**

- **BERT**: "I want to [MASK] pizza" → "[eat]"
- **GPT**: "I want to eat" → "pizza"

#1-3. Vision 분야에서 MAE가 필요한 이유

🔷 **비전 분야: Self-supervised 학습은 여전히 복잡함**

- contrastive 방식: positive/negative 쌍 필요 (ex. SimCLR, MoCo)
- clustering 방식: 임의 증강을 여러 개 만들어 군집화 (ex. SwAV, DINO)
- 대부분 **고복잡도 파이프라인**, **많은 데이터 증강**, **많은 계산량** 필요

🔷 **BEiT와 같은 token prediction 방식도 등장했지만:**

- discrete visual token 사전(dVAE) 필요
- 복잡한 사전 학습 단계 & 품질 제약

🔷 **Vision에도 BERT처럼 "간단한 마스킹 → 복원"이 가능할까?**

- 이미지는 dense한 신호이고, 의미 단위가 모호
- 그럼에도 불구하고:

→ **"이미지 일부를 가리고 복원"하는 구조가 잘 작동한다는 것을 MAE가 입증**

🔷 **MAE의 핵심 출발점**

"간단한 픽셀 복원만으로도
시각적 의미 표현을 학습할 수 있다면?
→ 계산 효율적이고 확장 가능한 비전 표현 학습의 새로운 길"

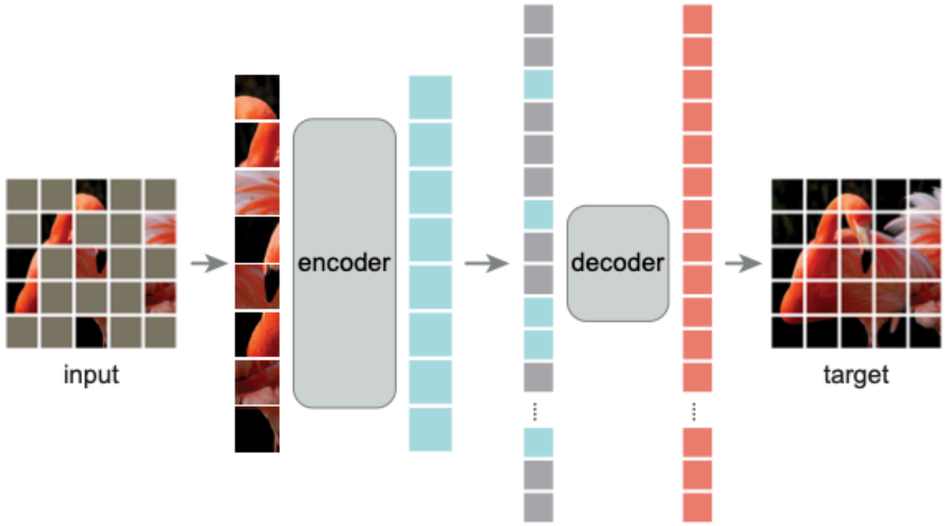


Figure 1. **Our MAE architecture**. During pre-training, a large random subset of image patches (*e.g.*, 75%) is masked out. The encoder is applied to the small subset of *visible patches*. Mask tokens are introduced *after* the encoder, and the full set of encoded patches and mask tokens is processed by a small decoder that reconstructs the original image in pixels. After pre-training, the decoder is discarded and the encoder is applied to uncorrupted images (full sets of patches) for recognition tasks.

Related Work

EWHA EURON

2-1 기존 연구와의 차별점

🔷 기존 Vision 분야 Self-supervised 방법들

방법	주요 아이디어	한계
Denoising Autoencoder	입력에 노이즈 추가 → 원본 복원	이미지 전체를 인코딩함 → 계산량 큼
Context Encoder	중앙 영역 마스킹 후 복원	사각형만 마스킹 → 과도하게 구조에 의존
MoCo / SimCLR	이미지 증강 쌍 간의 representation 유사도 학습 (contrastive)	large batch / memory bank 필요, 학습 어려움
BEiT	discrete visual token 예측 (dVAE로 사전 생성된 토큰 복원)	토큰 품질에 의존, 복잡한 사전처리 단계
MAE (제안 방법)	랜덤 패치 마스킹 + 픽셀 복원 (encoder는 보이는 부분만)	간결한 구조, 효율적인 학습, 고성능

🔷 MAE의 구조적 차별점 요약

- 기존 autoencoder 방식: **전체 이미지 인코딩 → 계산량 높음**
- MAE: **75% 마스킹 + 비대칭 인코더/디코더 구조 → 연산량 절감**
- 기존 contrastive 방식: **positive/negative 쌍 필요 → 복잡한 훈련**
- MAE: **단일 이미지로 충분 → 간단한 self-supervised 구성**

#2-2 MAE는 기존 방식과 무엇이 다른가?

✅ MAE의 차별점 핵심 3가지

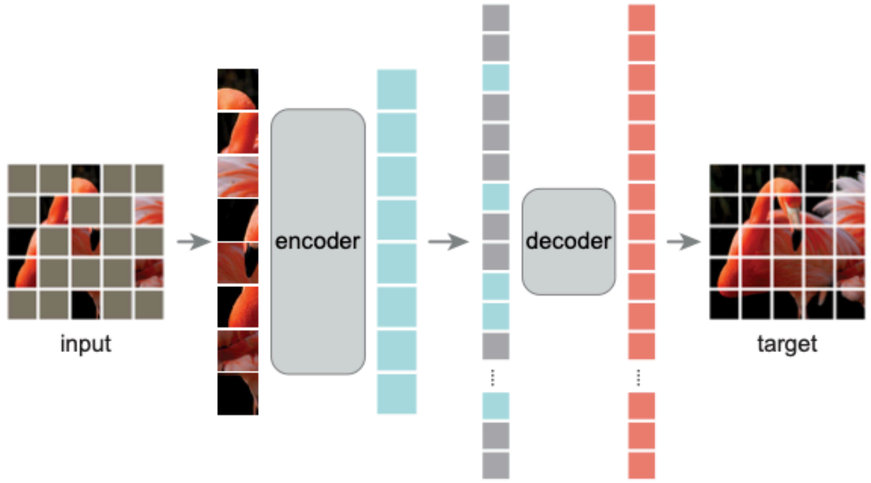


Figure 1. **Our MAE architecture**. During pre-training, a large random subset of image patches (*e.g.*, 75%) is masked out. The encoder is applied to the small subset of *visible patches*. Mask tokens are introduced *after* the encoder, and the full set of encoded patches and mask tokens is processed by a small decoder that reconstructs the original image in pixels. After pre-training, the decoder is discarded and the encoder is applied to uncorrupted images (full sets of patches) for recognition tasks.

🔷 1. 간결한 학습 구조 (Simple Objective)

- contrastive처럼 쌍 만들 필요 없음
- BEiT처럼 discrete token 사전 필요 없음
- 단일 이미지 입력 → **픽셀 복원만으로 학습**

기존 방식: 복잡한 파이프라인
MAE: 이미지 마스킹 → 인코더 → 디코더 → 픽셀 복원 (끝)

🔷 2. 효율적인 비대칭 구조 (Asymmetric Architecture)

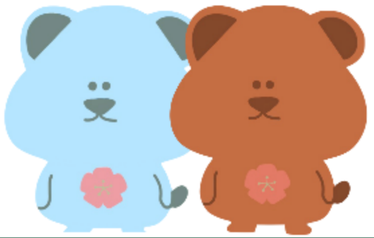
- 인코더: 보이는 패치만 처리 (작고 빠름)
- 디코더: 복원만 담당 (단순한 shallow 구조 가능)

전체 이미지 처리보다 **연산량 약 1/4 이하**
→ **대규모 학습에 유리**

🔷 3. 픽셀 복원만으로도 강력한 표현 학습 (Effective Learning)

- token 예측 없이도 semantic 정보 추출 가능
- linear probing, fine-tuning, transfer task 모두 **SOTA에 근접**

제안 방법론



EWHA EURON

#3 MAE의 핵심 아이디어 (Key Idea)

🔷 핵심 질문에서 출발

"이미지의 일부를 가리고, 그걸 복원하는 것만으로
시각적 표현(visual representation)을 학습할 수 있을까?"

🔷 MAE의 핵심 구성요소 3가지

1. **Random Masking of Input Image**
   - 입력 이미지의 **약 75% 패치**를 무작위로 제거
   - 남은 25%만 인코더에 입력
   - 마스킹 위치는 매번 다르게 설정 (data augmentation 역할도)
2. **Asymmetric Encoder-Decoder Architecture**
   - **인코더**: 마스킹되지 않은 패치만 인코딩 (연산량 절감)
   - **디코더**: 마스크 토큰을 추가한 후 전체 이미지 복원 시도
   - 디코더는 shallow 구조로 설계 (학습 부담 없음)
3. **Pixel-level Reconstruction Objective**
   - BEiT와 달리 discrete token이 아닌 **픽셀 자체 복원**
   - 즉, RGB 패치(target patch)의 **mean squared error(MSE)**를 최소화하는 것이 목적

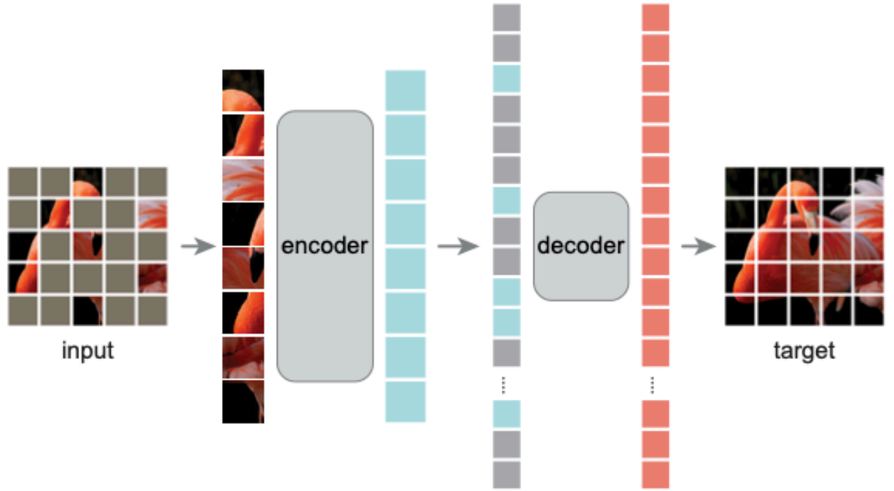


Figure 1. **Our MAE architecture**. During pre-training, a large random subset of image patches (*e.g.*, 75%) is masked out. The encoder is applied to the small subset of *visible patches*. Mask tokens are introduced *after* the encoder, and the full set of encoded patches and mask tokens is processed by a small decoder that reconstructs the original image in pixels. After pre-training, the decoder is discarded and the encoder is applied to uncorrupted images (full sets of patches) for recognition tasks.

#4-1 Architecture - 입력 이미지 처리 및 마스킹 전략

🔷 이미지 패치 분할 (Patchifying)

- 입력 이미지 크기: 예) 224×224
- 패치 단위로 분할: 16×16 또는 32×32 (ViT 방식)
- 총 패치 수 예시 (16×16 패치 기준):

→ $(224 / 16)^2$ = **196개의 토큰**

🔷 무작위 마스킹 (Random Masking)

- 전체 패치 중 **75%를 무작위로 제거**
- 남은 **25%만 인코더에 입력**
- 고정 마스킹 아님 → 훈련 중 **매 배치마다 마스킹 위치 변경**

🔷 왜 높은 비율(75%)로 마스킹할까?

- 학습 난이도를 인위적으로 높임 → 더 강력한 표현 유도
- 과적합 방지 + 인코더가 본질적 구조 학습하도록 유도
- 실험적으로도 75%가 가장 효율적 (논문 Fig. 4 참조)

🔷 마스킹 방식의 특징

특징	설명
랜덤	고정된 위치가 아닌, **패치 단위 무작위 선택**
불균형	일정한 블록 마스킹(X), 다양한 위치 제거
반복성	매 배치마다 다른 마스킹 → augmentation 역할

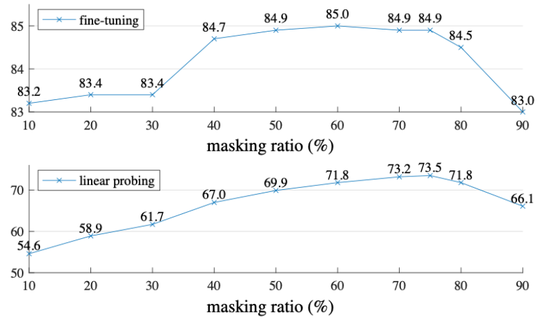


Figure 5. **Masking ratio**. A high masking ratio (75%) works well for both fine-tuning (top) and linear probing (bottom). The y-axes are ImageNet-1K validation accuracy (%) in all plots in this paper.

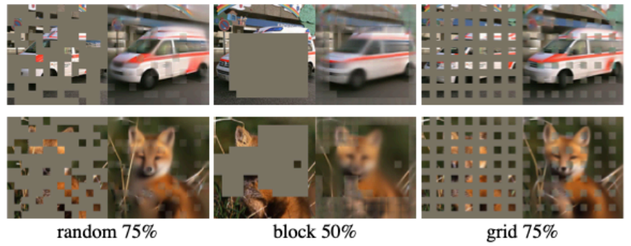


Figure 6. **Mask sampling strategies** determine the pretext task difficulty, influencing reconstruction quality and representations (Table 1f). Here each output is from an MAE trained with the specified masking strategy. Left: random sampling (our default). Middle: block-wise sampling [2] that removes large random blocks. Right: grid-wise sampling that keeps one of every four patches. Images are from the validation set.

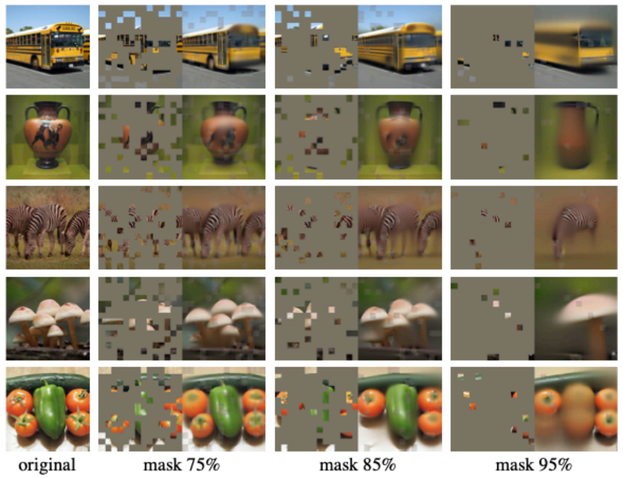


Figure 4. Reconstructions of ImageNet *validation* images using an MAE pre-trained with a masking ratio of 75% but applied on inputs with higher masking ratios. The predictions differ plausibly from the original images, showing that the method can generalize.

#4-2 Architecture - Encoder: Efficient Representation Learning

🔷 Encoder의 입력: 마스킹되지 않은 패치만 처리

- 입력 이미지 → 패치 분할 → 75% 무작위 마스킹
- **남은 25% 패치만 인코더에 전달됨**
- 즉, 전체 이미지의 **일부만** 인식하고 **전부를 이해하려 함**

🔷 Encoder의 역할: 의미 있는 시각 표현 학습

- 복원 목표는 디코더가 맡지만,
- **복원 가능한 표현을 추출하는 주체는 인코더**
- 따라서 MAE에서 중요한 학습은 대부분 인코더에 집중됨

🔷 Transformer Encoder 구조 (ViT 기반)

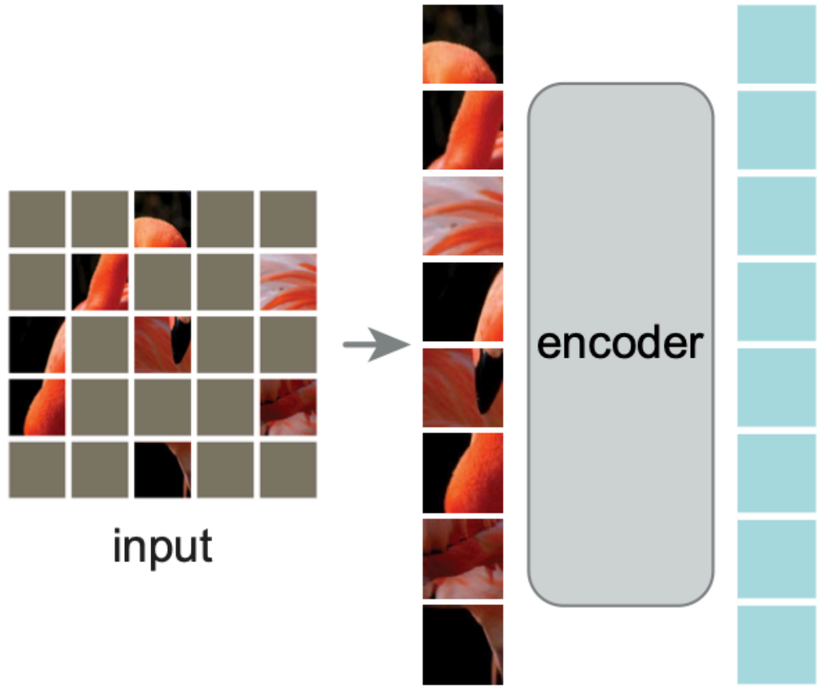
- 구조: ViT와 동일한 multi-head self-attention 기반
- positional embedding 사용 (absolute positional embedding)
- 학습 목표는 복원이지만, 표현 학습은 인코더가 수행

🔷 Fine-tuning 단계에서는 디코더 제거

- downstream task (ex: classification)에서는

→ **encoder만 활용**

→ 간결하고 성능 좋은 transfer 구조



🔷 효율적인 이유: “Sparse Computation”

비교 항목	기존 방식 (ViT)	MAE Encoder
입력 패치 수	전체 (100%)	일부 (25%)
연산량	많음	**약 1/4 이하**
표현 학습	전 패치 대상	**정보 밀도 높은 부분 집중**

→ **연산량은 줄고, 표현 품질은 유지됨**

→ 특히 large-scale training에서 큰 이점

#4-3 Architecture - Decoder: Lightweight Reconstruction Head

🔷 디코더의 목적: 복원 전용 (not representation)

- **인코더가 표현 학습의 핵심**
- 디코더는 단지 **마스킹된 패치의 픽셀 값 복원**만 담당
- 학습 중에만 사용되며, **inference에는 사용하지 않음**

🔷 입력: 전체 시퀀스 구성

- 인코더 출력 (25% 토큰) + 마스크 토큰 (75%)
- 마스크 토큰: 학습 가능한 벡터

→ masking된 위치에 채워 넣음

- positional embedding은 full-length로 다시 적용

🔷 디코더 구조: shallow & simple

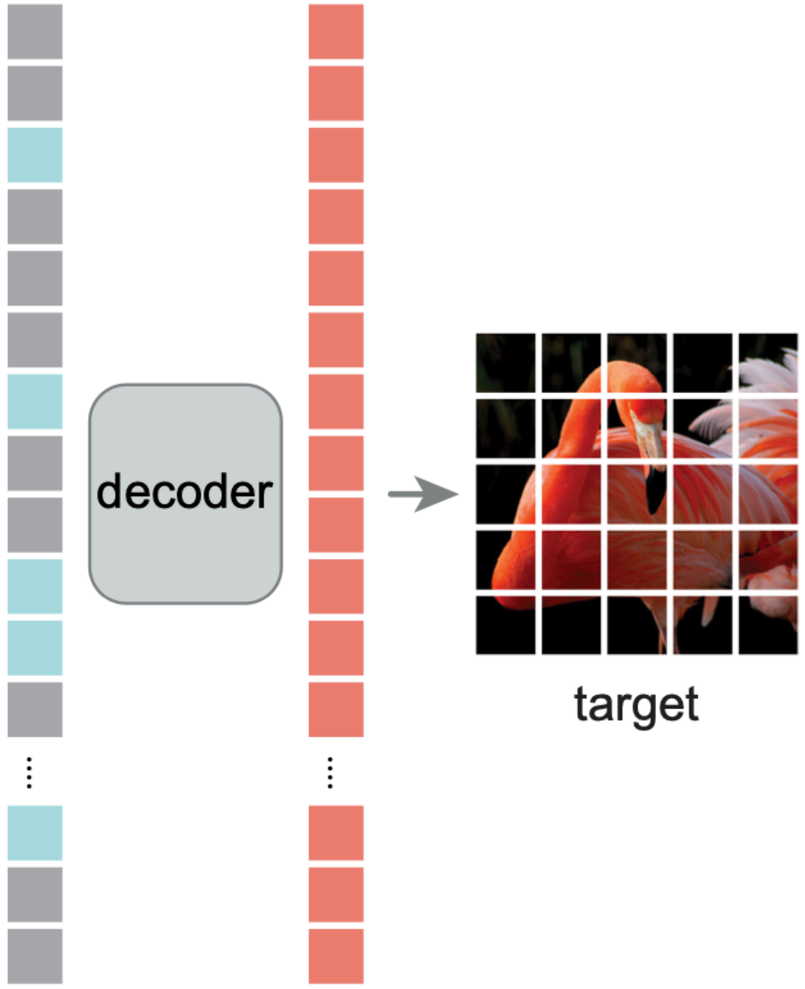
- Transformer 구조이긴 하지만 **4~8 layer 정도의 얕은 구조**
- 전체 패치를 한꺼번에 받아 **픽셀 복원 수행**
- 디코더는 표현이 아닌 **복원 품질만** 최적화함

핵심: **디코더는 학습 중에만 존재하고, 다운스트림 전이에는 쓰이지 않음**

🔷 왜 lightweight한가?

- 표현 학습은 인코더에 집중 → 디코더는 최소 구조로 충분
- 복잡한 디코더는 **오히려 overfitting을 유발**
- **shallow decoder가 실험적으로 가장 좋은 성능**

논문 ablation: 디코더 depth가 깊을수록 linear probing 정확도 감소



#4-4 Architecture - 학습 목표: 픽셀 복원 (Pixel Reconstruction Objective)

🔷 MAE의 학습 목표는 “픽셀 복원”

- 마스킹된 패치의 **원래 픽셀 값(RGB)**을 복원하는 것이 학습 목표
- 예측 대상: 마스킹된 영역의 **mean-normalized patch (e.g., 16×16×3)**
- 사용된 손실 함수: **Mean Squared Error (MSE)**

🔷 학습 중 디코더에만 손실 부여

- MSE 손실은 디코더의 출력과 원래 이미지 간의 차이
- 인코더는 간접적으로 학습

→ **복원이 잘 되도록 representation을 생성**

🔷 왜 픽셀 복원이 효과적인가?

- Token-level 예측 (BEiT) 없이도 **semantic한 표현**을 학습할 수 있음
- 복원 자체가 **context-aware representation**을 만들도록 유도
- 패치 간 관계를 이해해야 정확한 복원이 가능함

→ “복원을 잘 하려면 이미지를 **이해**해야 한다”

🔷 다른 방식들과 비교

방식	예측 대상	특징
SimCLR / MoCo	patch 간 contrastive 비교	augment pair 필요
BEiT	discrete token prediction	pre-learned tokenizer 필요
MAE	픽셀 자체 복원	**가장 단순하고 직접적**

MAE는 학습 파이프라인이 매우 간단하면서도,
표현 품질은 SOTA에 근접하거나 초과함

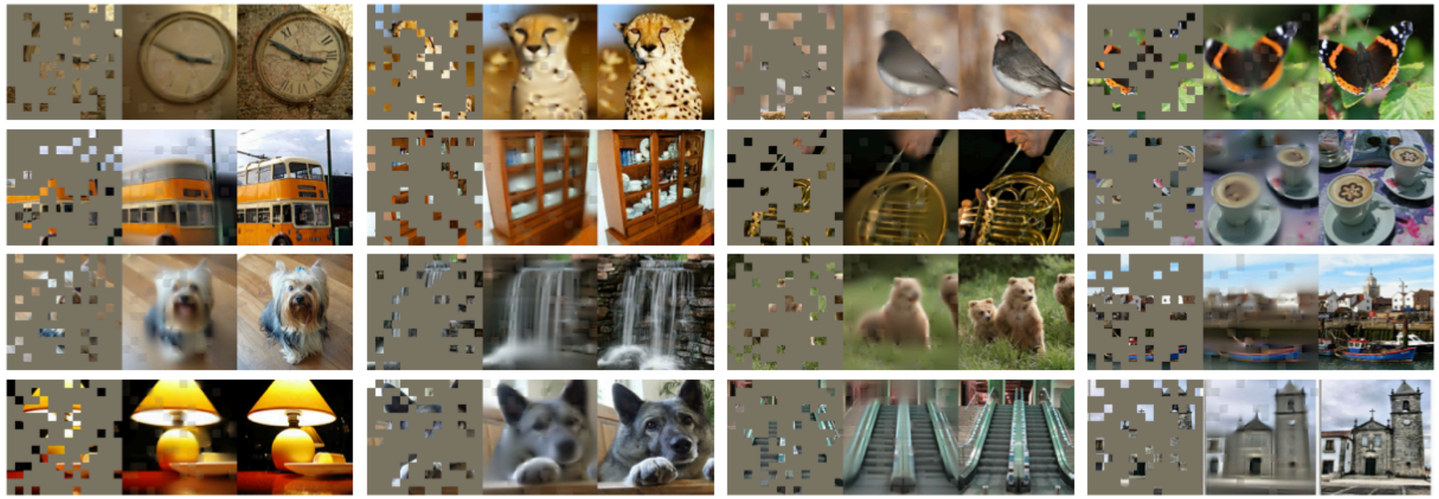
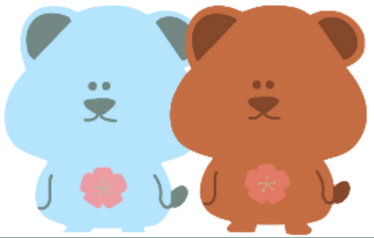


Figure 2. Example results on ImageNet *validation* images. For each triplet, we show the masked image (left), our MAE reconstruction† (middle), and the ground-truth (right). The masking ratio is 80%, leaving only 39 out of 196 patches. More examples are in the appendix.
†*As no loss is computed on visible patches, the model output on visible patches is qualitatively worse. One can simply overlay the output with the visible patches to improve visual quality. We intentionally opt not to do this, so we can more comprehensively demonstrate the method's behavior.*

실험 및 결과



EWHA EURON

#5-0 Architecture - 디자인 선택에 대한 실험적 검증

🔷 마스크 비율(mask ratio) 실험
- 실험 범위: 50%, 60%, 75%, 90%
- 결과: 75%에서 가장 높은 전이 성능
- 너무 적으면 과잉 정보 → 학습 난이도↓
- 너무 많으면 정보 부족 → 복원 실패↑

🔷 디코더 깊이 실험
- 실험 대상: 4, 6, 8, 12-layer decoder
- 결과: shallow decoder(4 layers)가 가장 우수
- deep decoder는 표현 학습에 불필요한 burden을 줌

→ 디코더는 간단할수록 좋음

🔷 Patch size 실험
- 16×16 vs 32×32 비교
- patch가 커질수록 복원 난이도 ↑
- 16×16이 가장 안정적이고 성능 좋음 (default)

🔷 Positional embedding 실험
- Absolute vs relative vs none
- absolute embedding이 가장 일관적 성능

🔷 정리

요소	선택	실험 결과
Mask ratio	75%	최고 성능
Decoder depth	4-layer	가장 안정적
Patch size16×16.	성능 우수	
Positional embedding.	Absolute.	가장 적합

blocks	ft	lin
1	84.8	65.5
2	**84.9**	70.0
4	**84.9**	71.9
8	**84.9**	**73.5**
12	84.4	73.3

(a) **Decoder depth**. A deep decoder can improve linear probing accuracy.

dim	ft	lin
128	**84.9**	69.1
256	84.8	71.3
512	**84.9**	**73.5**
768	84.4	73.1
1024	84.3	73.1

(b) **Decoder width**. The decoder can be narrower than the encoder (1024-d).

case	ft	lin	FLOPs
encoder w/ [M]	84.2	59.6	3.3×
encoder w/o [M]	**84.9**	**73.5**	**1**×

(c) **Mask token**. An encoder without mask tokens is more accurate and faster (Table 2).

case	ft	lin
pixel (w/o norm)	84.9	73.5
pixel (w/ norm)	**85.4**	**73.9**
PCA	84.6	72.3
dVAE token	85.3	71.6

(d) **Reconstruction target**. Pixels as reconstruction targets are effective.

case	ft	lin
none	84.0	65.7
crop, fixed size	84.7	73.1
crop, rand size	**84.9**	**73.5**
crop + color jit	84.3	71.9

(e) **Data augmentation**. Our MAE works with minimal or no augmentation.

case	ratio	ft	lin
random	75	**84.9**	**73.5**
block	50	83.9	72.3
block	75	82.8	63.9
grid	75	84.0	66.0

(f) **Mask sampling**. Random sampling works the best. See Figure 6 for visualizations.

Table 1. **MAE ablation experiments** with ViT-L/16 on ImageNet-1K. We report fine-tuning (ft) and linear probing (lin) accuracy (%). If not specified, the default is: the decoder has depth 8 and width 512, the reconstruction target is unnormalized pixels, the data augmentation is random resized cropping, the masking ratio is 75%, and the pre-training length is 800 epochs. Default settings are marked in gray.

#5-1. 마스킹 비율 실험: 75%가 최적

🔷 실험 목적

- 마스킹 비율(mask ratio)이 **표현 학습 성능에 미치는 영향 분석**
- 너무 낮으면 → 정보 과잉 → 학습 쉬움, 일반화↓
- 너무 높으면 → 정보 부족 → 복원 실패

목표: 가장 효율적인 마스킹 비율 찾기

🔷 실험 설정

- MAE pretraining 후, linear probing으로 평가
- 사용된 비율: 0%, 25%, 50%, 60%, 75%, 90%
- dataset: **ImageNet-1K**

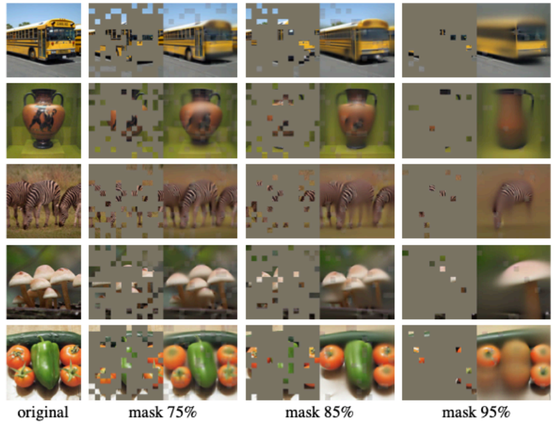


Figure 4. Reconstructions of ImageNet *validation* images using an MAE pre-trained with a masking ratio of 75% but applied on

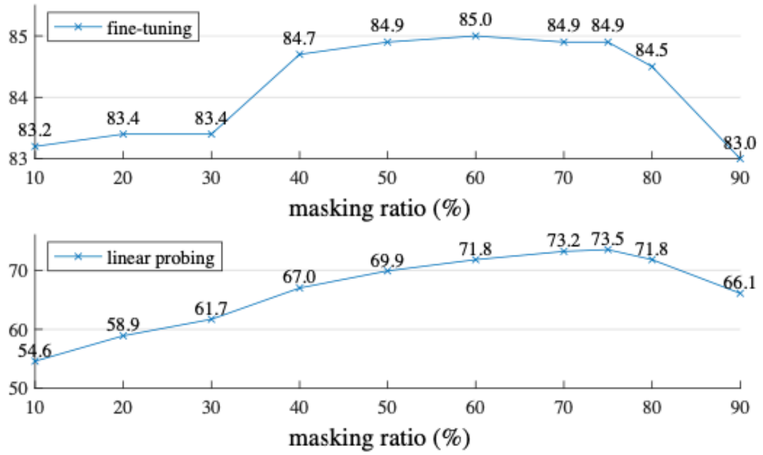


Figure 5. **Masking ratio**. A high masking ratio (75%) works well for both fine-tuning (top) and linear probing (bottom). The y-axes are ImageNet-1K validation accuracy (%) in all plots in this paper.

🔷 주요 결과

- **75% 마스킹에서 최고 성능 기록**

→ linear probing accuracy: **67.8% (ViT-B 기준)**

- 성능 곡선은 **U자형 곡선**

→ 너무 적거나 너무 많아도 성능 떨어짐

🔷 해석 및 인사이트

- **75% 마스킹**은:
  - 복원이 어려워지며 인코더가 더 **의미 있는 표현**을 학습하게 만듦
  - 하지만 90% 이상은 정보가 부족해서 복원 성능 하락
- 높은 마스킹은 데이터 효율성도 높임

→ 적은 패치로도 강력한 표현 학습 가능

🔷 정리

- MAE는 매우 **높은 비율의 마스킹**도 견딜 수 있는 구조
- **75% 마스크**는 representation과 efficiency 모두에서 최적 지점
- contrastive learning은 일반적으로 2-view만 사용하지만,

MAE는 하나의 view에서 **정보 일부만 보고 학습**

5-2. 디코더 구조, 증강, 마스킹 전략 비교 실험

- MAE의 구성 요소들(디코더 깊이, 너비, 증강, 마스크 방식)이 학습 성능에 얼마나 영향을 미치는지 분석

→ **복잡한 구조나 heavy augmentation 없이도 성능이 유지되는가?**

🔷 디코더 깊이(depth) 실험

- 실험: 4 / 6 / 8 / 12 layer transformer decoder
- 결과: **4-layer 디코더에서 가장 우수한 성능**
- 깊어질수록 성능 하락 → 표현 학습을 방해

🔷 디코더 너비(width) 실험

- 실험: hidden dimension 256 / 512 / 768
- 결과: **너비가 커질수록 오히려 복원 성능 하락**
- 이유: 표현에 과도한 capacity 부여 → overfitting

🔷 Data Augmentation 사용 여부

- 기존 contrastive learning (SimCLR, MoCo 등)은 heavy augmentation 사용
- MAE는 **crop 외에는 augmentation 없음**
- 실험 결과: augmentation을 추가해도 성능 개선 **미미**
- → 구조 자체가 **self-supervision에 강함**

🔷 Mask Sampling 전략

- 실험: random / block-wise / fixed / attention-guided 등
- 결과: **random sampling이 가장 성능 우수**
- 복잡한 마스킹 전략 불필요 → simple is strong

🔷 핵심 인사이트

- MAE는 **shallow & narrow decoder**, **minimal augmentation**, **simple masking**만으로도 **강력한 성능 달성**
- 구조 설계의 **단순함이 오히려 일반화에 도움**이 됨

EWHA EURON

5-3. Linear Probing vs Fine-tuning: MAE 표현의 전이 성능

method	pre-train data	ViT-B	ViT-L
supervised	IN1K w/ labels	47.4	49.9
MoCo v3	IN1K	47.3	49.1
BEiT	IN1K+DALLE	47.1	53.3
MAE	IN1K	**48.1**	**53.6**

Table 5. **ADE20K semantic segmentation** (mIoU) using UperNet. BEiT results are reproduced using the official code. Other entries are based on our implementation. Self-supervised entries use IN1K data *without* labels.

🔷 평가 방식 설명: Linear Probing vs Fine-tuning

항목	Linear Probing	Fine-tuning
정의	인코더 freeze + 최종 분류기만 학습	전체 네트워크 파라미터 재학습
목적	학습된 **표현력만 평가**	downstream task에 맞게 **전체 조정**
난이도	빠르고 간단	시간/리소스 많이 소모됨

→ linear probing 성능이 좋다는 건 **표현 그 자체가 강력하다는 것**

🔷 실험 결과: ImageNet-1K 기준

- **MAE (ViT-H/14 + 300 epoch pretrain)** 기준
- Linear probing 정확도: **67.8% (ViT-B), 77.6% (ViT-L), 80.1% (ViT-H)**
- Fine-tuning 정확도: **83.6% (ViT-B), 85.9% (ViT-L), 87.8% (ViT-H)**

🔶 MAE는 **linear probing만으로도 supervised pretraining과 유사한 수준 달성**

🔶 fine-tuning 시엔 SOTA 성능 경신 (특히 ViT-H)

🔷 비교 대상들과의 차이점

방법	Pretrain	Linear Probing	Fine-tune
MoCo-v3	contrastive	65.4%	83.2%
BEiT	token prediction	66.8%	85.2%
MAE	pixel reconstruction	**67.8%**	**87.8%**

→ MAE가 **두 방식 모두에서 최고 성능 기록**

🔷 핵심 인사이트

- Linear probing만으로도 높은 정확도 → **표현의 일반화 능력 탁월**
- Fine-tuning에선 SOTA 성능 → **표현 + 적응력 모두 뛰어남**
- MAE는 **적은 compute 비용으로 높은 효율**을 달성

#5-4. Partial Fine-tuning: 적은 파라미터로 강력한 성능

🔷 실험 목적

- 전체 모델이 아닌 **일부 파라미터만 학습시켜도 성능이 유지되는가?**
- 모델 경량화, 리소스 절약을 위한 실용적 전이 전략 탐색
- 특히 **LayerNorm + 일부 layer만 fine-tune**

🔷 실험 설정

- 전체 인코더 중 특정 층만 **fine-tune**
  - 예: 마지막 블록만 학습, 나머지는 freeze
- 또는 **LayerNorm만 fine-tune** (1~2% 파라미터만 학습)

🔷 결과 요약

- ViT-L 기준:
  - Full fine-tuning: **85.9%**
  - Last block only: **83.4%**
  - LayerNorm only: **83.1%**
- 파라미터 수를 1~10%만 사용하고도 **선형 probing보다 높은 성능 달성**

🔷 실용적 의미

- **모든 파라미터를 학습하지 않아도 고성능 가능**
- 특히 대규모 모델(ViT-H 이상)에서는 **적은 fine-tune cost**로 높은 효율
- 개인화, edge 환경, 리소스 제한 상황에 유리

결론 & Discussion

EWHA EURON

#6-1 결론

“MAE는 단순한 self-supervised 구조로도, 비전에서 확장 가능하고 효율적인 표현 학습이 가능하다는 것을 증명했다.”

🔷 핵심 요소 설명

✅ Masked Autoencoder 구조
이미지의 일부만 보고 복원하는 self-supervised 학습

✅ 비대칭 구조
인코더는 마스킹된 입력만 처리, 디코더는 lightweight

✅ 75% 마스크 비율
적은 정보로 강력한 표현을 학습

✅ 전이 성능
Linear probing, fine-tuning 모두 SOTA 성능

✅ 효율성
Partial fine-tuning, minimal augmentation으로도 고성능

🔷 MAE의 차별점

- **픽셀 복원 기반 self-supervised 학습**이라는 새로운 패러다임
- NLP의 BERT-style masking을 **비전 분야에 성공적으로 적용**
- 학습 구조는 단순하지만, 성능은 강력함
- Augmentation이나 contrastive view 없이도 SOTA 달성

🔷 실용적 장점

- 고성능 ViT 모델과 쉽게 결합 가능 (e.g., ViT-L, ViT-H)
- 경량화 가능한 구조 → **partial fine-tuning**에 최적
- compute 자원이 제한된 환경에서도 학습 및 전이 효율이 높음

🔷 한계 및 향후 연구 방향

❗ 시각적 의미 이해 부족
단순 픽셀 복원은 **고차원 의미 학습이 약할 수 있음**

🔬 semantic pretext 부족
BEiT처럼 token-level semantic prediction 아님

🔄 확장 가능성
multimodal MAE (예: CLIP+MAE), temporal MAE (video), hierarchical MAE 등

🔶 최근에는 MAE를 활용한 **비디오 복원**, **3D reconstruction**, **multimodal MAE** 연구가 활발해지고 있음

EWHA EURON

#6-2 discussion

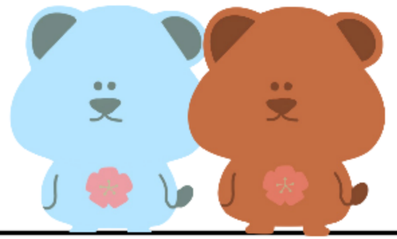
🔷 **75% 마스킹 비율이 다른 task에도 최적인가?**

- ImageNet에서는 75%가 best였지만, 자연 영상 외에:
- 의료 이미지, 위성 사진, 은하 이미지 등에서는 어떨까?
- 복잡하거나 중요한 feature가 국소적으로 몰린 경우, adaptive mask가 더 나은 선택일까?

🔷 **MAE를 multimodal 학습에 확장하면 어떤 점이 어려울까?**

- 예: CLIP-style 텍스트-이미지 MAE
- 어떤 modality에 masking을 적용하고 어떤 modality로 복원할 것인가?
- multimodal에서는 단순한 masking이 정보 손실을 초래할 수 있다는 우려에 대한 논의

THANK YOU



EWHA EURON