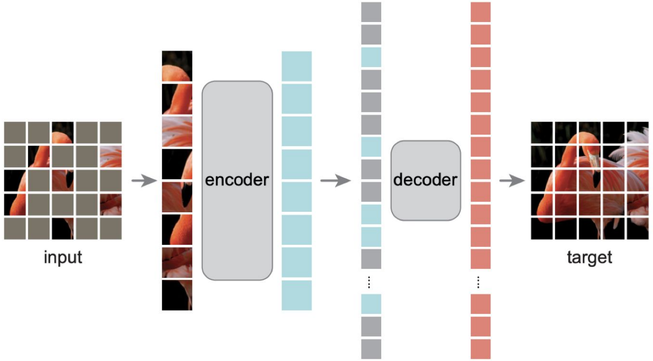
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

# Understanding Masked Autoencoders via Hierarchical Latent Variable Models

Lingjing Kong*, Martin Q. Ma*, Guangyi Chen, Eric P. Xing, Yuejie Chi,
Louis-Philippe Morency, Kun Zhang

CARNEGIE MELLON UNIVERSITY
PITTSBURGH PENNSYLVANIA 1900

# Masked Autoencoders: SOTA Self-supervised Learning Paradigm



Courtesy: He et al., 2022

Mask sampling: random masks are determined by *masking ratio* and *patch size.*

Encoding: the encoder maps the unmasked input to a representation.

Decoding: the decoder reconstructs the masked input from the representation and the positional information.

MAE attains state-of-the-art fine-tuning performance on various vision tasks, including classification, detection, segmentation, and more.

Great! But in principle,

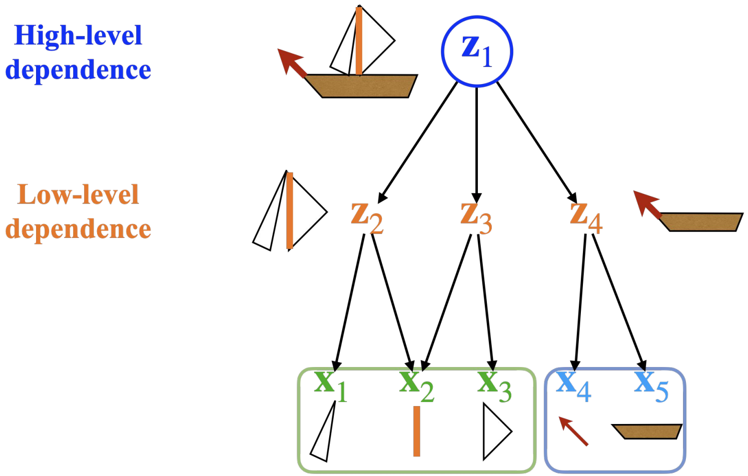
a. Why can MAE learn meaningful representation?
b. How do key hyperparameters determine the representation properties?

We offer insights from a latent-variable identification perspective!

# A hierarchical data-generating process for vision data

“Hierarchical” to represent various levels of dependence among pixels:

- Low-level dependence within a single object.
- High-level dependence between distinct objects.



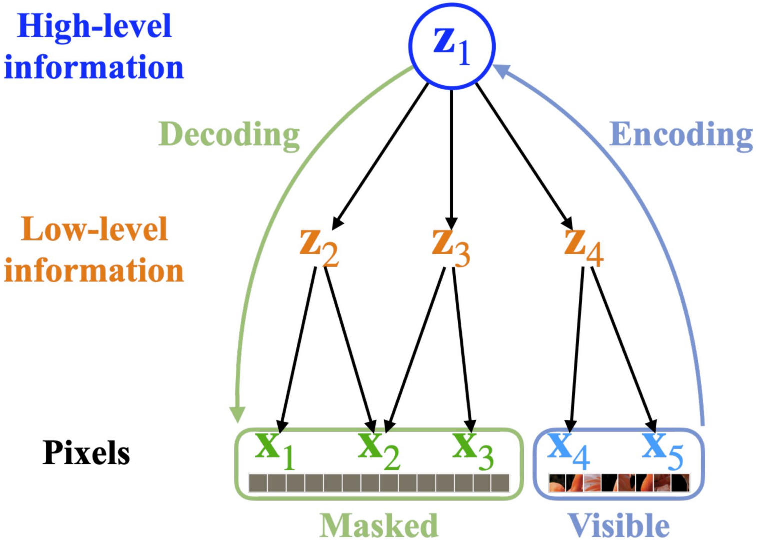
Assumptions:

- No directed edge among observed variables (i.e., pixels).
- Generating processes are invertible.

# MAE works by identifying latent variables in the generating process!

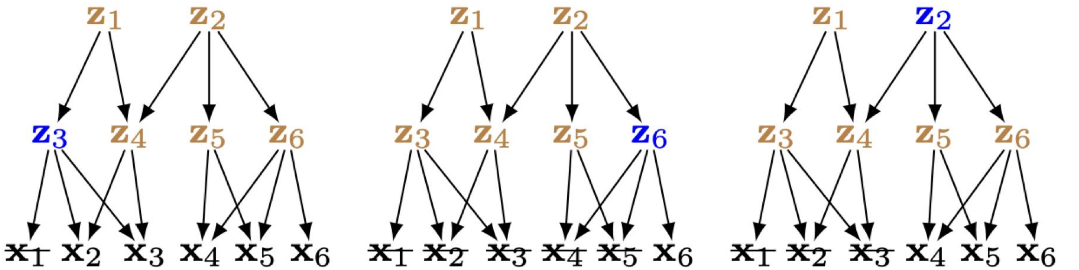
Each specific mask corresponds to a specific set of latent variables (Theorem 2).

MAE can provably recover the true latent variables specified by masking (Theorem 1, 2).



# How do hyperparameters determine representation quality?

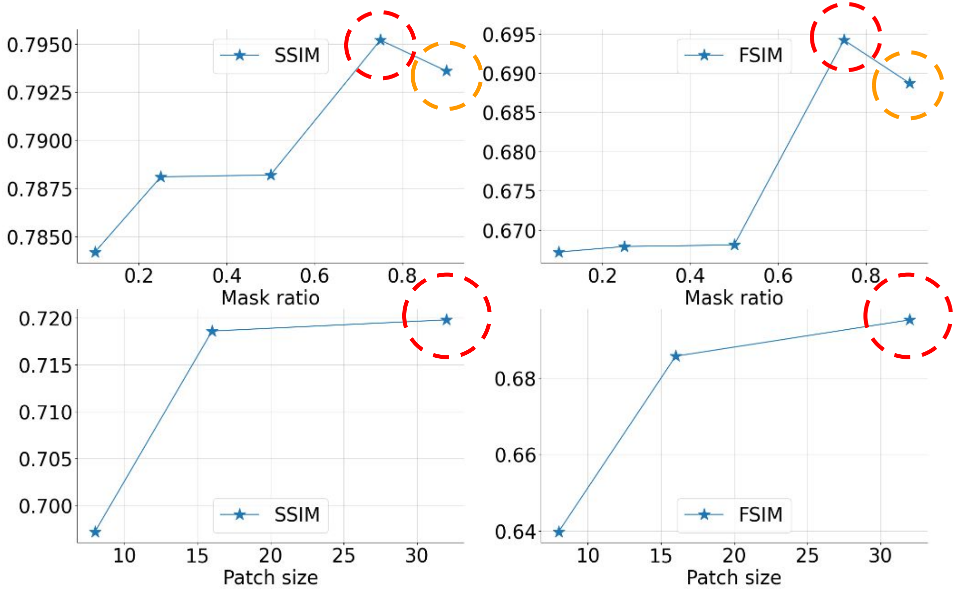
- Masking ratios and sizes induce the model to capture low- or high-level information.
- Learning high-level representations is generally hard with random masking.

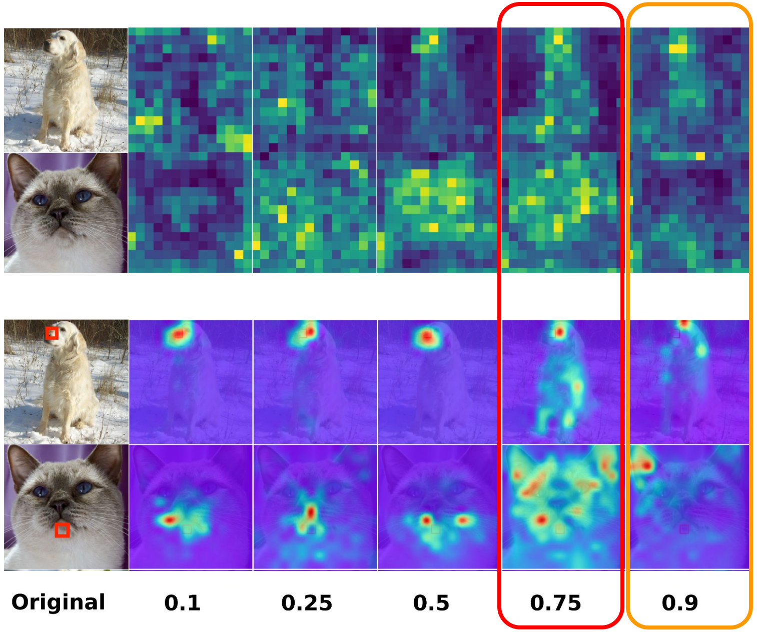


$\mathbf{z}_1$ $\mathbf{z}_2$
$\mathbf{z}_3$ $\mathbf{z}_4$ $\mathbf{z}_5$ $\mathbf{z}_6$
~~$\mathbf{x}_1$~~ $\mathbf{x}_2$ $\mathbf{x}_3$ $\mathbf{x}_4$ $\mathbf{x}_5$ $\mathbf{x}_6$
$\mathbf{z}_1$ $\mathbf{z}_2$
$\mathbf{z}_3$ $\mathbf{z}_4$ $\mathbf{z}_5$ $\mathbf{z}_6$
~~$\mathbf{x}_1$~~ ~~$\mathbf{x}_2$~~ ~~$\mathbf{x}_3$~~ ~~$\mathbf{x}_4$~~ ~~$\mathbf{x}_5$~~ $\mathbf{x}_6$
$\mathbf{z}_1$ $\mathbf{z}_2$
$\mathbf{z}_3$ $\mathbf{z}_4$ $\mathbf{z}_5$ $\mathbf{z}_6$
~~$\mathbf{x}_1$~~ ~~$\mathbf{x}_2$~~ ~~$\mathbf{x}_3$~~ $\mathbf{x}_4$ $\mathbf{x}_5$ $\mathbf{x}_6$

(a) **Conservative mask** (b) **Aggressive mask** (c) **Ideal mask**

# Experiments: appropriate masking ratios capture high-level information





<span style="color:red">Higher</span> masking ratios and sizes are structurally more similar to the original image and capture more high-level semantic information, but extreme masking induces model to capture low-level information.

# Conclusion

- Why MAE can learn meaningful representation: MAE provably recovers high-level representations by identifying latent variables.
- Higher masking ratios and patch sizes induce the model to learn higher-level image representations.
- Learning high-level representations is generally hard with random masking.

## Formulation of Masking

- Mask samples: random masks are sampled from a distribution determined by *masking ratio* and *patch size*
- MAE encoder maps the unmasked input to a representation
- MAE decoder reconstructs the masked input from the representation and the positional information

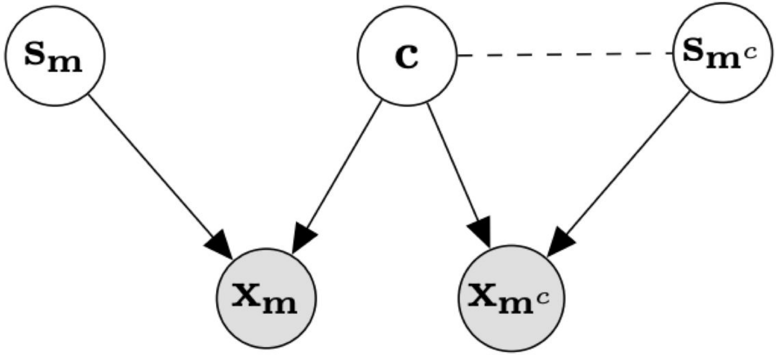
# Contribution: Understanding with latent variable models

- In this work, we establish a framework to understand MAE via identifiability guarantees.
- We first formulate the underlying data-generating process as a hierarchical latent variable model; and
- Then, under reasonable assumptions, MAE can recover a subset of true latent variables in the generating process
- The level of latent variables in the hierarchical model depends on how masking performs (masking ratio and patch size)
- We show that a moderate-to-aggressive masking ratio captures high-level information, while extremely aggressive or conservative masking captures low-level information

# Assumption: latent variable

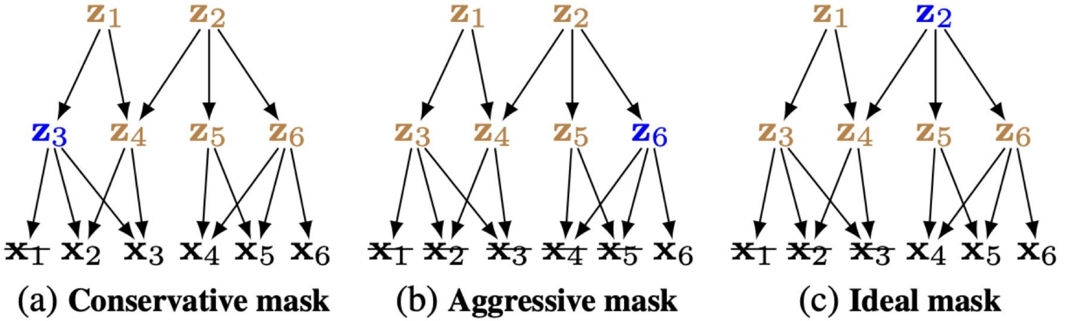
Latent variable c: a minimal set of variables to satisfy the following:

- $\mathbf{s_m}$
- TODO
- c is minimal



# Results: Identifiability

# Interpretations

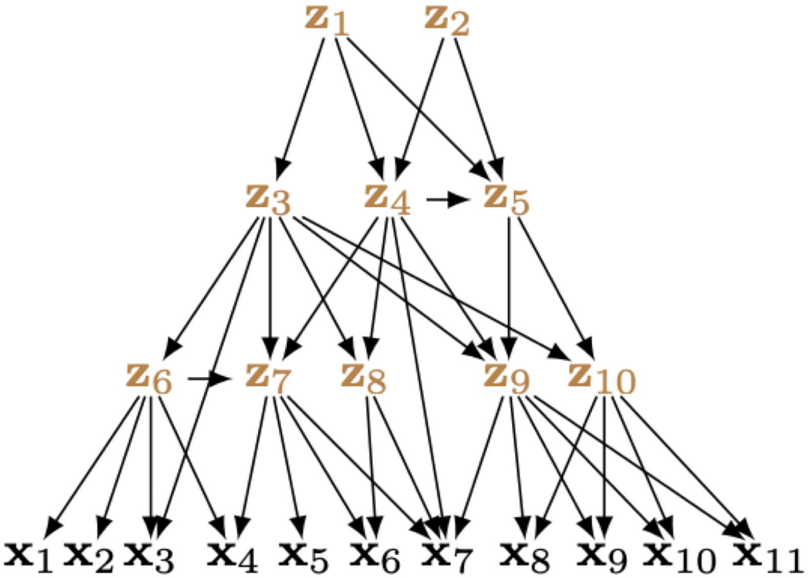


(a) **Conservative mask** (b) **Aggressive mask** (c) **Ideal mask**

- Conservative masking: undesirable, as the recovered latent variables are still at a low level
- (Too) aggressive masking: undesirable, as the recovered latent variables are also at a low level
- Ideal masking: moderate masking recovered latent variables at a high level

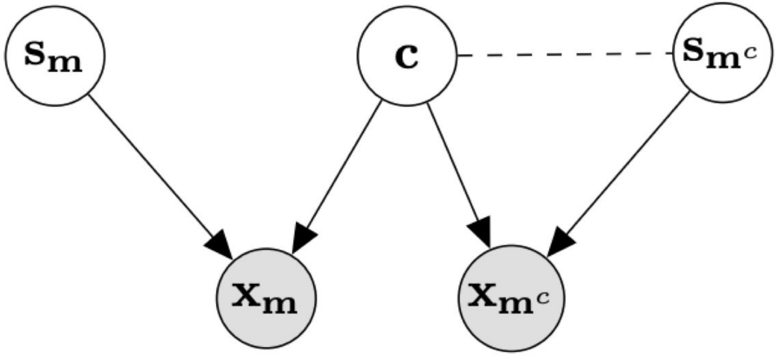
# Assumption:

- There is no direct edge between any two observables Xs; and
- Each variable is generated by their parents from a higher layer in a directed acyclic graph (DAG), combined with the exogenous variable in each layer.



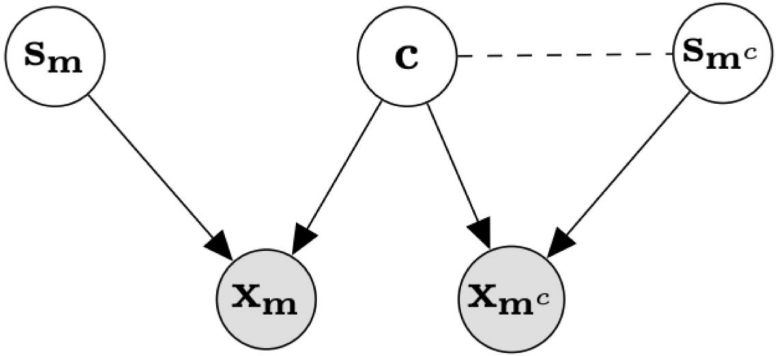
# Identifiability assumptions

- In a hierarchical latent variable structure, for any specific mask, there exists a minimal set of latent variables   such that the generating process can be expressed using the figure.
- Essentially, we want to locate a subset of true latent variables that fully captures the statistical dependency between the masked and visible parts.
- The transformations from a higher layer to a lower layer in the data-generating process are invertible.
- , where   or   refers to the information specific to the masked or unmasked part.
- The content variable   is minimal in terms of dimensions.



# Identifiability results

- Result 1: for each mask m, there exists a unique that contains sufficient high-level information to reconstruct the masked and the unmasked .
- Result 2: For any mask, the MAE encoder can recover all the information of the minimal set .



$\mathbf{s}_{\mathbf{m}}$
$\mathbf{c}$
$\mathbf{s}_{\mathbf{m}^c}$
$\mathbf{x}_{\mathbf{m}}$
$\mathbf{x}_{\mathbf{m}^c}$