ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10448091

J-MAE: JIGSAW MEETS MASKED AUTOENCODERS IN X-RAY SECURITY INSPECTION

Weichen Xu[1,†], *Jian Cao*[1,*,†], *Tianhao Fu*[1], *Awen Bai*[2], *Ruilong Ren*[1], *Zicong Hu*[1], *Xixin Cao*[1], *Xing Zhang*[1,3,*]

[1]School of Software and Microelectronics, Peking University, Beijing, China
[2]Beihang University, Beijing, China
[3]Shenzhen Graduate School, Peking University, China

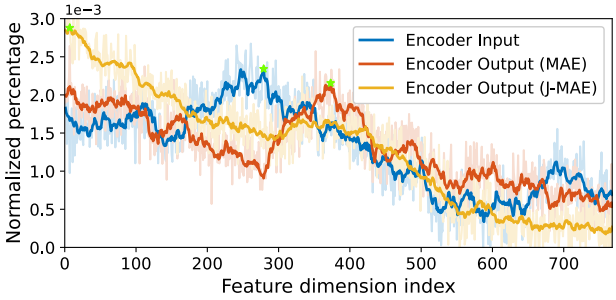
ABSTRACT

The X-ray security inspection aims to identify any restricted items to protect public safety. Due to the lack of focus on unsupervised learning in this field, using pre-trained models on natural images leads to suboptimal results in downstream tasks. Previous works would lose the relative positional relationships during the pre-training process, which is detrimental for X-ray images that lack texture and rely on shape. In this paper, we propose the jigsaw style MAE (J-MAE) to preserve the relative position information by shuffling the position encoding of visible patches. This forces the network to perform semantic reasoning to understand the shape and composition of X-ray objects. Meanwhile, we propose the Incremental Shuffling Module (ISM) and Permute Predicting Module (PPM) to make the training process more stable and accelerate convergence. Our proposed method has consistently outperformed other methods on three downstream X-ray security inspection datasets.

Index Terms— X-ray security inspection, Unsupervised learning, Masked image modeling, Jigsaw puzzles

1. INTRODUCTION

With the rising crowd density in public transportation hubs, security inspection has become important for ensuring public safety. Typically, X-ray scanners are employed to identify restricted items. However, pinpointing prohibited items amidst many X-ray images makes it challenging for security inspectors to consistently and accurately detect all potential threats. Recent works [1, 2, 3, 4] have achieved remarkable progress in recognizing prohibited items from X-ray images. Utilizing the ImageNet pre-training weights through representation learning [5, 6] as a starting point has become the consensus among these exceptional works. However, the notable disparity in imaging principles between natural images and X-ray images [7] hinders the convergence of the prohibited items detection network. Unsupervised learning for X-ray images in security inspection remains an unexplored territory.



(a)

(b)

Fig. 1. (a) Frequency histogram indexed by feature maxima. (b) Example images and relative position distribution map of encoder input and features obtained after MAE and J-MAE pre-trained encoder.

Masked Image Modeling (MIM) [8, 6, 9, 10, 11] has gained significant attention in visual unsupervised representation learning due to its exceptional fine-tuning performance in downstream tasks. It involves masking random portions of an image and compelling the model to reconstruct these

* Corresponding author. † Equal contribution. This work was supported by Peking University-Delta Electronics Joint Industrial IoT and Intelligent Systems Laboratory Innovation Research Program.

masked areas. The 768-dimensional features of the downstream dataset OPIXray [2], HiXray [3], and EDS [4] are extracted based on the ViT-Base/16 [12] model pre-trained with the MAE [6] objective using SiXray [1] training data. The frequency histogram, sourced from 20173888 patches and indexed by feature maxima, is shown in Fig. 1(a). We perform statistical analysis on the encoder's input and output features. To reduce dimensionality, we utilize the features of the two largest dimensions (indicated by green pentagrams) to characterize the 768 dimensions. For each example image in HiXray, we divide it into 196 patches. We then extract the features of the encoder input and output for each patch. We perform dimensionality reduction and obtain low-dimensional features with corresponding position information for each patch by utilizing the statistical information in Fig. 1(a). By taking the central patch (indicated by the red square) as a reference, we calculate the feature difference between the remaining patches and the reference patch. The normalized relative position distribution map is shown in Fig. 1(b). Due to positional encoding, the input features of the encoder exhibit clear relative positional relationships. However, for the features obtained after MAE pre-trained encoder, the relative positional relationships are disturbed. This flaw in MAE significantly negatively impacts X-ray recognition tasks because X-ray images typically contain less texture information, and the shape and composition are crucial for discrimination [1].

To preserve the relative position information of the image and further infer the shape and composition of X-ray objects, we consider optimizing MAE using jigsaw puzzles [13, 14], which aims at recovering an original image from its shuffled patches. Specifically, after masking a high proportion of the image, we shuffle the preserved image patches. This is accomplished by adding a scrambled positional encoding to the preserved image patches. The image reconstruction process avoids duplicating close pixels and instead relies on semantic reasoning to understand the shape and composition of X-ray objects. To stabilize the training processes in the early stages, we propose the Incremental Shuffling Module (ISM), which gradually increases the degree of shuffling through a curriculum learning schedule [15]. On the other hand, the permute matrix is used to explicitly supervise the reconstruction process in the Permute Predicting Module (PPM) to accelerate convergence. Specifically, all the output image patches are passed through the Convolutional Block Attention Module (CBAM) [16] to further get the position vector, which serves as the key for the attention module. Subsequently, the position query initialized with Gaussian sampling is summed with the pre-defined positional encoding to form the position query. In attention module, each position query performs a query operation on all the keys, resulting in an attention weight matrix. This matrix is then subjected to operations such as scale, Gumbel softmax, and Sinkhorn [17], ultimately transforming into doubly stochastic matrices [17]. The permute matrix is used as ground truth to alleviate the pressure on the reconstruction process. As shown in Fig. 1(b), for the features obtained after the jigsaw style MAE (J-MAE) pre-trained encoder, the relative positional relationships are partially preserved, which is beneficial for X-ray security inspection.

The main contributions of our approach are:

- We propose jigsaw style MAE (J-MAE) to preserve more relative positional relationships in the representation learning process. We are the first to investigate unsupervised learning for X-ray images in security inspection.
- We propose the Incremental Shuffling Module (ISM) and Permute Predicting Module (PPM) to stabilize the pre-training process and converge rapidly.
- Comprehensive experiments performed on the downstream dataset demonstrate the effectiveness of our method. Our approach achieves state-of-the-art results by introducing jigsaw in generative unsupervised learning.

2. METHODOLOGY

2.1. Overall Pipeline

An overview of our proposed J-MAE is presented in Fig. 2. First, the image is randomly masked. The position encodings of preserved patches are shuffled using the permute matrix generated by the Incremental Shuffling Module. The visible patches are then fed into the Mask-Guided Image Modeling for reconstruction. On the other hand, the reconstructed patches are indexed by the position query to approximate the permute matrix and explicitly predict the shuffling pattern.

2.2. Incremental Shuffling Module

After random masking, we get the masking matrix $\boldsymbol{\mathcal{M}}$, where 1 represents the masked patch, and 0 represents the preserved patch. We plan to shuffle the positional encoding of a portion of the preserved patches. We define that all 1s in $\boldsymbol{\mathcal{M}}$ belong to $\mathbf{1}$, and all 0s in $\mathbf{0}$. To make the training process more stable in the early stages, we introduce curriculum learning [15], where the shuffling ratio $\tau(t)$ increases with the epochs, starting from 0.5, then 0.75, and finally reaching 1.0. The preserved patches are divided into $\mathbf{0_F}$ and $\mathbf{0_E}$ using $\tau(t)$:

$$\mathbf{0_F}, \mathbf{0_E} = RandSelect(\mathbf{0}, \tau(t)), \tag{1}$$

where $\mathbf{0_F}$ and $\mathbf{0_E}$ represent the fixed and the exchangeable patches, respectively. We define the original patch index sequence as $\boldsymbol{\mathcal{I}} = [0, 1, 2, ..., n-1]$, where n represents the number of patches. The shuffled patch index sequence is denoted as $\widetilde{\boldsymbol{\mathcal{I}}}$. We define the shuffling rules as follows:

$$\begin{cases} \widetilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{i}} = \boldsymbol{\mathcal{I}}_{\boldsymbol{i}} & \text{if } \boldsymbol{\mathcal{M}}_{\boldsymbol{i}} \in \{\mathbf{1}, \mathbf{0_F}\} \\ \widetilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{j}} = \boldsymbol{\mathcal{I}}_{\boldsymbol{i}} & \text{if } i \neq j \text{ and } \boldsymbol{\mathcal{M}}_{\boldsymbol{i}}, \boldsymbol{\mathcal{M}}_{\boldsymbol{j}} \in \mathbf{0_E} \\ \widetilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{j}} \neq \widetilde{\boldsymbol{\mathcal{I}}}_{\boldsymbol{k}} & \text{if } k \neq j \end{cases}. \tag{2}$$

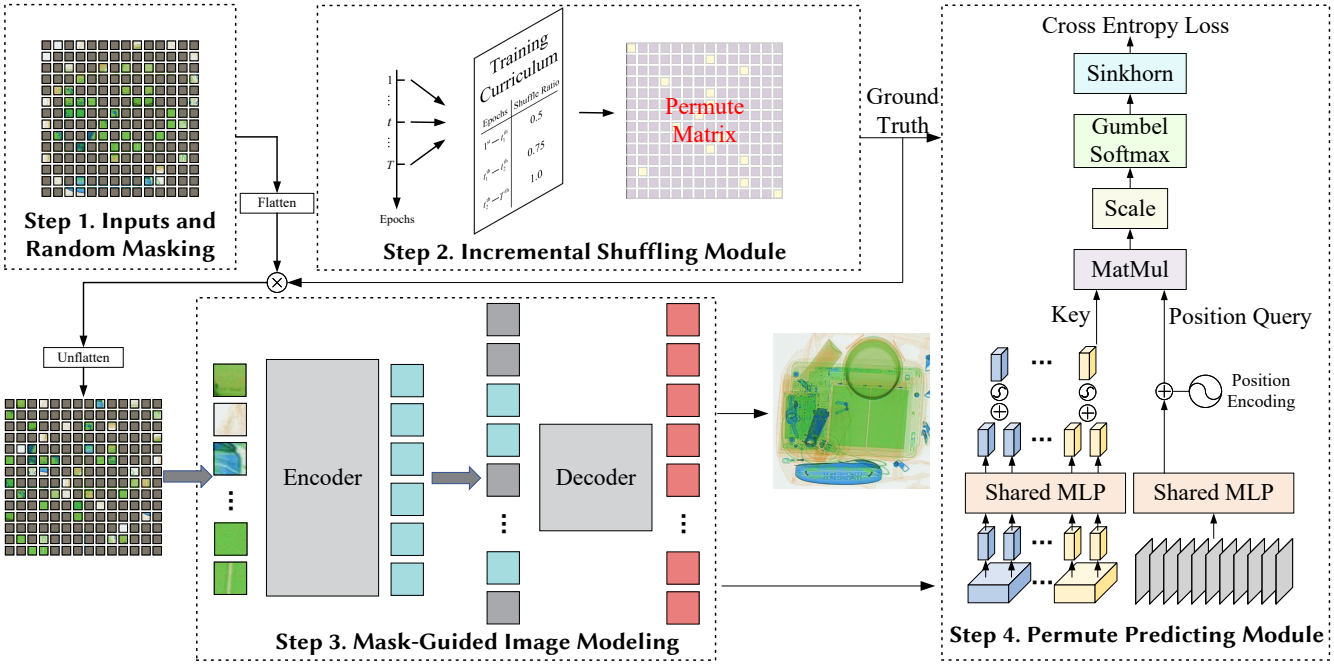


Fig. 2. **The proposed J-MAE framework** consists of two main modules: Incremental Shuffling Module and Permute Predicting Module. Permute matrix is exploited to shuffle the masked image in jigsaw style while serving as ground truth for Permute Predicting Module. The gray features represent the position query initialized with Gaussian sampling.

Then, the shuffled patch index sequence $\widetilde{\mathcal{I}}$ is constructed into a permutation matrix $\mathcal{R}$. We flatten the masked image to obtain $\mathcal{P}$, multiply $\mathcal{P}$ by the permutation matrix $\mathcal{R}$ to obtain the shuffled image $\widetilde{\mathcal{P}}$, and add sequential positional encoding to it. Since the image patches have been shuffled, the positional encoding is scrambled.

2.3. Permute Predicting Module

To alleviate the burden of reconstruction, in addition to the MAE reconstruction loss, we employ the permute matrix $\mathcal{R}$ to supervise the shuffling process explicitly. Specifically, on one hand, the output patch features $\mathcal{F}_i (i = 0, 1, 2, \dots n-1)$ are fed into CBAM [16] to extract position vectors $\mathcal{V}_i$:

$$\mathcal{V}_i = \sigma(\mathrm{MLP}(\mathrm{P_m}(\mathcal{F}_i)) + \mathrm{MLP}(\mathrm{P_a}(\mathcal{F}_i)), \tag{3}$$

where $\mathrm{P_m}$ and $\mathrm{P_a}$ represent 1-D max pooling and 1-D average pooling, respectively. The position vectors $\mathcal{V}$ are further employed as the keys for the attention module. On the other hand, we use position encoding as the position query. However, to avoid the illogical situation of exploding attention weights caused by fixed position encoding, we follow the approach used in variational autoencoders [18] to set the mean and variance for each channel and initialize with a Gaussian distribution. This results in the initialization of the position query $\mathcal{Q}_I$. After passing through the MLP, $\mathcal{Q}_I$ is added to the fixed position encoding to obtain the final position query $\mathcal{Q}$. Subsequently, the position query $\mathcal{Q}$ is employed to query attention weights from the position vectors $\mathcal{V}$, ideally with the ground truth being the permute matrix $\mathcal{R}$. Considering that the permute matrix $\mathcal{R}$ belongs to the class of doubly stochastic matrices, where the sum of each row and column is 1, we apply Gumbel softmax and Sinkhorn [17] operations to the attention weights. This helps reduce the difficulty of predicting the permute matrix by matrix transformation. The permute prediction loss can be represented as follows:

$$L_{\mathrm{P}} = \mathrm{CrossEntropy}(\mathcal{R}, \mathrm{S}(\mathrm{G}(\frac{\mathcal{Q} \cdot \mathcal{V}^{\mathbf{T}}}{\sqrt{d_q}}))), \tag{4}$$

where G, S represent Gumbel softmax and Sinkhorn, respectively, and d_q is the dimension of $\mathcal{Q}$, which is 768 by default.

3. EXPERIMENTS AND DISCUSSIONS

3.1. Experimental Setup

We do unsupervised pre-training on large-scale SIXray [1], consisting of 1,059,231 X-ray images. Then we do supervised training on three prohibited items detection datasets OPIXray [2], HiXray [3], and EDS [4].

In pre-training, we adopt ViT-Base/16 [12] as the backbone. AdamW optimizer with cosine learning rate scheduler is employed, and the training is conducted for 800 epochs. The hyperparameters for training include a batch size of 512, an image resolution of 224^2, and a base learning rate of 1e-4. Our data augmentation strategy is consistent with MAE.

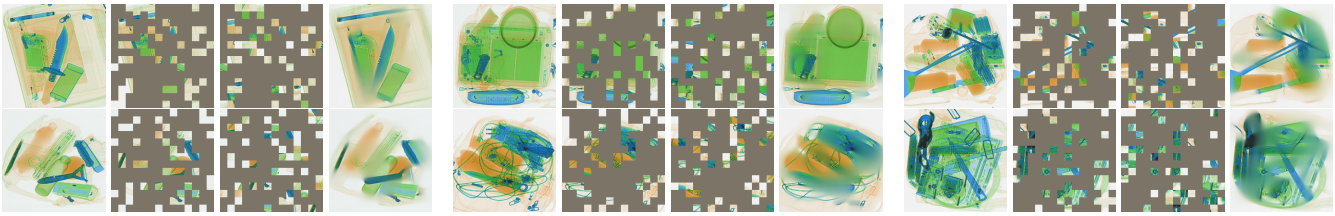


Fig. 3. Example visualization results on HiXray testing split images from ViT-Base/16 model pre-trained with the J-MAE objective using SiXray training data. For each image quadruplet, we show the original input image (1st column), the masked image (2nd column), the jigsaw style masked image (3rd column), and the reconstructed image (4th column).

In downstream tasks, We finetune the advanced YOLOX [19] with a ViT-based FPN backbone. We train the network using AdamW optimizer with 100 epochs.

3.2. Comparison to State-of-the-Art

Table 1. Comparison to state-of-the-art generative unsupervised learning methods on OPIXray, HiXray, EDS testing set. MU, etc., are abbreviations of category names.

method	**OPI**	MU	FO	**Hi**	MP	PO1	**EDS**	SE	UM
BEiT [8]	68.4	74.6	70.2	75.4	95.1	85.3	48.5	44.8	83.5
MAE [6]	73.5	83.2	77.6	77.7	95.3	87.7	52.8	49.6	85.8
SimMIM [9]	73.2	83.3	76.5	78.2	95.8	88.8	52.6	52.3	85.6
MaskFeat [10]	74.4	84.2	77.9	77.5	96.4	89.3	52.1	51.9	86.6
CAE [20]	76.5	84.4	79.5	78.9	96.4	89.6	54.5	53.9	88.7
LoMaR [21]	75.3	**84.9**	79.8	79.5	95.9	90.4	53.9	53.3	**89.4**
MILAN [11]	**77.8**	83.9	80.6	78.6	97.0	89.8	54.6	54.6	88.7
J-MAE (Ours)	77.2	84.6	**81.2**	**80.1**	**97.3**	**90.7**	**56.2**	**55.4**	89.2
Improvements	+3.7	+1.4	+3.6	+2.4	+2.0	+3.0	+3.4	+5.8	+3.4

We compare with other advanced generative unsupervised learning methods on downstream OPIXray, HiXray, and EDS testing sets, as shown in Table 1. Here, OPI, Hi, and EDS represent the average precision across all categories. Besides, we also include precision for the top two categories with the most instances. Our method achieves the best precision on three downstream datasets and outperforms the baseline MAE. This demonstrates the importance of preserving relative positional relationships for X-ray security inspection.

3.3. Ablation Study

Compared to MAE, J-MAE has made three improvements, namely: jigsaw style shuffling, ISM, and PPM. The results in Table 2 demonstrate jigsaw, ISM, and PPM all play essential roles. Jigsaw style shuffling can improve the average precision from 77.7% to 78.5%. This indicates that emphasizing the shape and composition of X-ray objects during pre-training is beneficial for X-ray security inspection. ISM and PPM can stabilize the training process and alleviate the reconstruction pressure, leading to a 1.0% and 0.6% improvement.

To demonstrate the advantage of J-MAE in preserving relative positional relationships, we introduced deep PE supervision in MAE, which explicitly adds position encoding before each transformer layer in the encoder. Table 3 compares the performance with deep PE supervision, showing that deep PE supervision can improve the performance of MAE but is inferior to J-MAE. This is because the position encoding partially gets lost when transferred to downstream tasks.

Table 2. Ablation study on HiXray testing set. PO1, etc., are abbreviations of category names, as mentioned in [3].

jigsaw	ISM	PPM	**AVG**	PO1	PO2	WA	LA	MP	TA	CO	NL
			77.7	87.7	87.5	85.4	95.1	95.3	89.7	61.0	19.8
✓			78.5	89.3	**92.4**	87.6	91.5	95.3	91.1	59.4	21.2
✓	✓		79.5	90.2	88.7	89.4	97.7	94.8	**91.9**	60.4	**22.9**
✓	✓	✓	**80.1**	**90.7**	91.2	**90.3**	**97.9**	**97.3**	90.2	**61.2**	21.6

Table 3. Influence of J-MAE compared to deep PE supervision on HiXray testing set. PO1, etc., are the same as Table 2.

method	**AVG**	PO1	PO2	WA	LA	MP	TA	CO	NL
MAE [6]	77.7	87.7	87.5	85.4	95.1	95.3	89.7	61.0	19.8
J-MAE (Ours)	**80.1**	**90.7**	**91.2**	**90.3**	**97.9**	**97.3**	**90.2**	61.2	**21.6**
MAE + deep PE supervision	78.8	89.1	88.0	87.2	97.0	95.8	90.2	**62.2**	21.3

3.4. Visualization

In Fig. 3, the successful performance of the reconstruction results indicates that the pre-training process attempts to infer and understand the shape and composition of X-ray objects.

4. CONCLUSIONS

In this paper, we note that existing masked image modeling methods would lose the relative positional relationships during the pre-training process, which is detrimental for downstream X-ray security inspection tasks that rely on shape information. To alleviate this issue, we propose J-MAE, an advanced unsupervised learning approach that jointly addresses masked image modeling and jigsaw puzzles. The process of reordering shuffled patches and reconstructing the image forces the network to understand the shape and composition of X-ray objects. In addition, we propose the Incremental Shuffling Module (ISM) and Permute Predicting Module (PPM) to stabilize the training process and alleviate the complexity of reconstruction. Equipped with these components, J-MAE achieves state-of-the-art performance on three downstream X-ray security inspection datasets, offering new insights to the security screening community.

5. REFERENCES

[1] Caijing Miao, Lingxi Xie, Fang Wan, chi Su, Hongye Liu, jianbin Jiao, and Qixiang Ye, "Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[2] Yanlu Wei, Renshuai Tao, Zhangjie Wu, Yuqing Ma, Libo Zhang, and Xianglong Liu, "Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[3] Renshuai Tao, Yanlu Wei, Xiangjian Jiang, Hainan Li, Haotong Qin, Jiakai Wang, Yuqing Ma, Libo Zhang, and Xianglong Liu, "Towards real-world x-ray security inspection: A high-quality benchmark and lateral inhibition module for prohibited items detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[4] Renshuai Tao, Hainan Li, Tianbo Wang, Yanlu Wei, Yifu Ding, Hongping Zhi Bowei Jin and, Xianglong Liu, and Aishan Liu, "Exploring endogenous shift for cross-domain detection: A large-scale benchmark and perturbation suppression network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[6] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[7] J Anthony Seibert, "X-ray imaging physics for nuclear medicine technologists. part 1: Basic principles of x-ray production," *Journal of nuclear medicine technology*, vol. 32, no. 3, pp. 139–147, 2004.

[8] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[9] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu, "Simmim: A simple framework for masked image modeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[10] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[11] Zejiang Hou, Fei Sun, Yen-Kuang Chen, Yuan Xie, and Sun-Yuan Kung, "Milan: Masked image pretraining on language assisted representation," *arXiv preprint arXiv:2208.06049*, 2022.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[13] Mehdi Noroozi and Paolo Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European Conference on Computer Vision*, 2016.

[14] Chen Wei, Lingxi Xie, Xutong Ren, Yingda Xia, Chi Su, Jiaying Liu, Qi Tian, and Alan L Yuille, "Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[15] Xin Wang, Yudong Chen, and Wenwu Zhu, "A survey on curriculum learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[16] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *ECCV*, 2018.

[17] Richard Sinkhorn, "A relationship between arbitrary positive matrices and doubly stochastic matrices," *The annals of mathematical statistics*, 1964.

[18] Diederik P Kingma, Max Welling, et al., "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, 2019.

[19] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[20] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang, "Context autoencoder for self-supervised representation learning," *arXiv preprint arXiv:2202.03026*, 2022.

[21] Jun Chen, Ming Hu, Boyang Li, and Mohamed Elhoseiny, "Efficient self-supervised vision pretraining with local masked reconstruction," *arXiv preprint arXiv:2206.00790*, 2022.