MINR: Implicit Neural Representations with Masked Image Modelling

Sua Lee* Joonhun Lee* Myungjoo Kang†
Seoul National University
{susan900,niceguy718,mkang}@snu.ac.kr

Abstract

Self-supervised learning methods like masked autoencoders (MAE) have shown significant promise in learning robust feature representations, particularly in image reconstruction-based pretraining task. However, their performance is often strongly dependent on the masking strategies used during training and can degrade when applied to out-of-distribution data. To address these limitations, we introduce the masked implicit neural representations (MINR) framework that synergizes implicit neural representations with masked image modeling. MINR learns a continuous function to represent images, enabling more robust and generalizable reconstructions irrespective of masking strategies. Our experiments demonstrate that MINR not only outperforms MAE in in-domain scenarios but also in out-of-distribution settings, while reducing model complexity. The versatility of MINR extends to various self-supervised learning applications, confirming its utility as a robust and efficient alternative to existing frameworks.

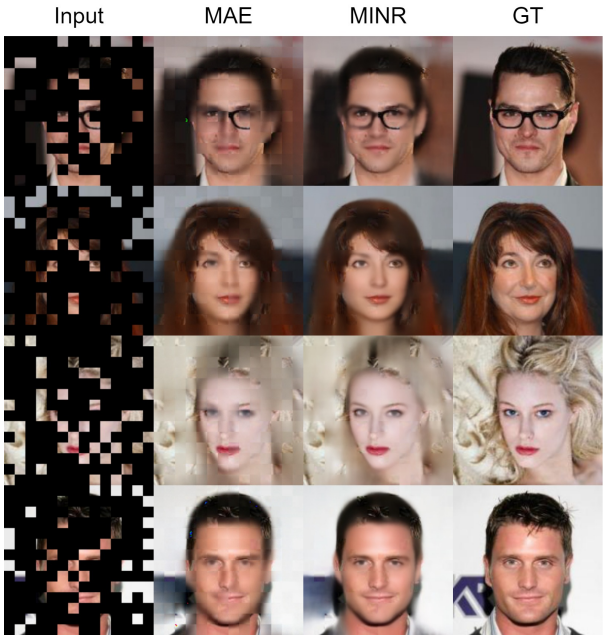

Figure 1: **Qualitative results of mask reconstruction.** For each row, we present the masked image, MAE and MINR reconstructions, and the ground truth, in sequence.

1. Introduction

Deep learning methods have rapidly advanced with supervised learning in computer vision, but it struggles with significant performance degradation when tested on the data distributions not observed during the training phase. This challenge stems from the inherent assumption in supervised learning: the training and test sets are drawn independently and identically from the same underlying data distribution. However, deep learning models often encounter situations requiring adaptation to unseen data distributions; generalizing effectively across such distributions is crucial for ensuring model robustness.

To address this, self-supervised learning (SSL) methods have gained attention. Using pretext tasks, SSL learns robust feature representations without human-annotated labels. Such feature representations have demonstrated increased robustness and performance in varied tasks domain, including domain generalization scenarios [1, 19]. Notable SSL techniques include masked image modelling (MIM) that enhance representation robustness by deliberately masking images and training models to reconstruct the hidden information. The effectiveness of MIM has been demonstrated in various downstream tasks, remaining one of the most powerful pretraining methods.

Masked autoencoder (MAE) has been often highlighted for its versatility and success in various tasks such as ConvNeXt V2 in image recognition [38], DropMAE in video representation [39], and PiMAE in 3D object detection [8]. Although certain frameworks may marginally outperform MAE in some scenarios, MAE's end-to-end approach simplifies the training process by eliminating the need for a separate pretrained model.

*Equal contribution
†Corresponding author

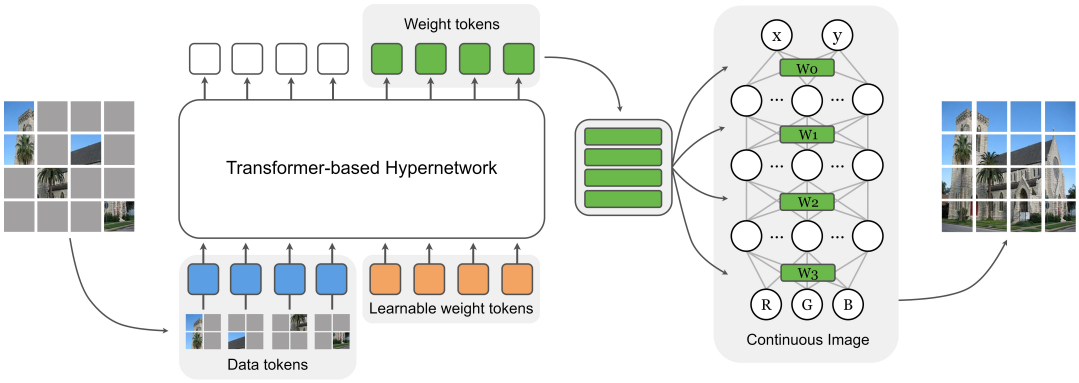

Figure 2: **A schematic illustration of MINR.** During the training phase, a large random subset of image patches are masked out. To ensure robustness, we employ a transformer-based hypernetwork predicting weights for an INR. TransINR directly maps weight sets [12], whereas GINR modulates only the second MLP layer as instance-specific, keeping the rest instance-agnostic [23]. The overall framework constructs weights with the masked-out input image and outputs an interpolated one.

However, a notable limitation of MAE is its dependency on masking strategies, such as mask size and area, as evidenced by several studies [22, 34, 36]. The MAE not only utilizes adjacent patches but also employs explicit information from all visible patches to fill each masked patch [7]. As shown in [24], the hierarchical extraction of explicit information—to fill the masked patches—plays an instrumental role in shaping the learned representation; it is directly influenced by the amount and quality of information in the visible patches. Furthermore, MAE computes the loss only on masked patches during training, thus, in testing, employing a masking strategy unseen during the training phase leads to a drastic decline in performance.

In this work, we introduce the *masked implicit neural representations (MINR)* framework that combines implicit neural representations (INRs) with MIM to address the limitations of MAE. The advantages of MINR include: i) Leveraging INRs to learn a continuous function less affected by variations of information in visible patches, resulting in performance improvements in both in-domain and out-of-distribution settings; ii) Considerably reduced parameters, alleviating the reliance on heavy pretrained model dependencies; and iii) Learning continuous function rather than discrete representations, which provides greater flexibility in creating embeddings for various downstream tasks.

2. Related works

2.1. Masked image modelling

Masked image modelling (MIM) has emerged as a promising approach in the field of self-supervised learning, enabling the derivation of robust representations by reconstructing occluded or masked imagery [4, 6, 14, 18, 43, 37]. The fundamental idea behind MIM is to artificially introduce occlusions in input data, followed by training a neural network to restore the original images from these masked versions. This process encourages the model to extract and focus on meaningful features, thus yielding a more robust and informative representation [31].

MIM can be categorized into two primary frameworks: the teacher-student framework and the MAE framework. In the former, a pretrained "teacher" network guides a "student" network to restore occluded data [3, 28, 37, 41, 45]. BEiT exemplifies this, considering the pretrained encoder as a fixed teacher and incorporating an additional layer mapping the path token to discrete pseudo labels [6]. Conversely, the MAE framework leverages an encoder-decoder architecture, directly predicting the obscured regions [18].

Recent advancements in MIM are geared towards the convergence of both frameworks [5, 25, 44], refinement of cornerstone models like BEiT and MAE [11, 21, 32, 38, 40, 42], and enhancing masking techniques [9, 22, 26, 34, 36]. In contrast, limited work has been done on adapting MIM to different architectures, such as MaskClip and A^2MIM [15, 27].

2.2. Implicit neural representations

Implicit neural representations (INRs) offer a promising alternative to traditional explicit representations, offering an innovative approach to depict complex geometries and continuous data without explicitly defining the underlying function. Rather than directly storing pixel values or clear-cut geometric data, INRs represent the underlying scene implicitly as a continuous function, usually represented by a deep network like a coordinate-based multi-layer perceptron (MLP). This function can project any pixel location

to its associated properties without direct geometric storage [12, 17, 23, 35].

The intrinsic nature of INRs provides a versatile framework, accommodating various input sizes and formats. Prominent developments in INRs research include neural radiance fields (NeRF), which model a volumetric scene's radiance from sparse 2D observations as a function of 3D coordinate function [30]. Additionally, the concept of INRs has been adapted for 2D image representation, enabling decoding for arbitrary output resolution [2, 10].

3. Method

In this section, we present the *masked implicit neural representations (MINR)* framework, designed for masked interpolation of input samples using INRs. Our method efficiently interpolates masked regions and offers robustness in out-of-distribution (OOD) settings. We use the TransINR and GINR architecture as the backbone of our approach, enabling seamless generalization across diverse dataset instances [12, 23]. The overview of proposed framework is visualized in Figure 2.

3.1. Integrating INRs with MIM

Given a dataset $\mathcal{O} = \{o^{(n)}\}_{n=1}^{N}$ containing N observations $o^{(n)} = (x_i^{(n)}, y_i^{(n)})$, we introduce random masks to produce the masked dataset $\mathcal{M}$:

$$\mathcal{M} = \{m^{(n)} \mid m^{(n)} \coloneqq \text{Mask}^{(n)}(o^{(n)})\}_{n=1}^{N}. \tag{1}$$

The primary goal of INRs is to estimate a continuous function $f_\theta : \mathbb{R}^2 \to \mathbb{R}^3$, which maps input coordinates to corresponding properties. Traditional INRs optimize an MLP for each instance, minimizing the L2 loss:

$$\mathcal{L}_n(\theta^{(n)}; m^{(n)}) = \frac{1}{H \times W} \sum_{i=1}^{H \times W} \|y_i^{(n)} - f_{\theta^{(n)}}(x_i^{(n)})\|_2^2. \tag{2}$$

Our INR approach only leverages implicit information present in the masked images, ensuring adaptability across different masking strategies.

3.2. Utilizing hypernetwork for generalizability

For INRs to be effective in OOD settings, it's crucial to train the model on individual instances. However, this basic approach is computationally challenging with large-scale datasets, and such independent training approach restricts the generalizability to unseen instances. Hence, we utilizes transformer-based hypernetwork architectures to modulate the MLP weights efficiently:

$$\theta = \{W_l\}_{l=1}^{L} \subset \mathbb{R}^{\text{in}_l \times \text{out}_l}, \tag{3}$$

with W_l denoting the weight of the l-th layer of the L-layer MLP.

TransINR-based approach. In [12], the hypernetwork predicts the entire set of INR weights $\theta^{(n)} = \{W_l\}_{l=1}^{L}$ simultaneously using the encoded information from the masked image $m^{(n)}$. Given the masked observation $\mathcal{M}$, generalized INR are optimized according to Equation 2 once the predicted weights have been modulated into:

$$\mathcal{L}(\Theta; \mathcal{M}) = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_n(\theta^{(n)}; m^{(n)}), \tag{4}$$

where $\Theta = \{\theta^{(n)}\}_{n=1}^{N}$ represents INR weights for entire instances in the dataset. However, this approach lacks robustness across instances, as weights are independently constructed.

GINR-based approach. Empirically found in [23], partitioning the MLP hidden layers into instance-specific and instance-agnostic layers is effective in learning commonalities across instances. Using the most performant configuration, we denote the second layer as instance-specific, with the other layers being instance-agnostic. The instance-agnostic parameters ϕ are shared across all instances:

$$\mathcal{L}(\Theta, \phi; \mathcal{M}) = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_n(\theta^{(n)}, \phi; m^{(n)}), \tag{5}$$

This design enables the MLP to learn patterns both common across a dataset and unique per instance, making it ideal for OOD settings.

4. Experiments

In this section, we assess the efficacy of our proposed MINR method against the MAE approach for mask reconstruction tasks. We conduct evaluations under both in-domain (ID) and out-of-domain (OOD) settings.

4.1. Datasets

We employ three diverse datasets for our experiments, namely `CelebA` [20], `Imagenette` [29], and `MIT Indoor67` [33]:

- `CelebA`: A comprehensive facial dataset encompassing 202K images.
- `Imagenette`: A subset of ImageNet, containing 10 distinct classes with a total of 7K images.
- `MIT Indoor67`: Specifically curated for indoor scene recognition, this dataset houses 15K images.

Notably, these datasets were selected due to their varied nature, which allows us to evaluate the robustness of our method across different data distributions.

Method		CEL	IMG	IND	# Param.
MAE	Large	15.018	14.693	15.181	313.6M
	Base	15.401	14.452	14.370	106.2M
MINR	TransINR	**21.865**	18.737	17.756	44.5M
	GINR	21.680	**19.358**	**18.622**	**43.7M**

Table 1: **Comparison of PSNR performances in ID mask reconstruction.** Columns represent the CelebA, Imagenette, and MIT Indoor67 datasets, respectively, with the last column indicating the number of parameters.

Method		CEL →		IMG →		IND →	
		IMG	IND	CEL	IND	CEL	IMG
MAE	Large	14.262	14.300	14.853	14.779	14.858	14.949
	Base	14.508	14.464	14.499	14.558	13.831	14.069
MINR	TransINR	**18.058**	**17.361**	19.929	17.920	18.992	18.103
	GINR	18.041	17.336	**19.994**	**18.045**	**19.509**	**18.573**

Table 2: **Comparison of PSNR performances in OOD mask reconstruction.** The arrow (→) indicates the source to target domain transfer.

4.2. Experimental setup

For our experiments, we maintain a consistent input image resolution of 182×182. We employ a 5-layer MLP to define f_θ. Images are segmented into patches of size 14×14, of which a majority, 75%, are subsequently masked out at random. The configuration of our transformer-based hypernetwork is in alignment with the vision transformer architecture detailed in [16].

The peak signal-to-noise ratio (PSNR) serves as our primary evaluation metric, a standard choice for reconstruction tasks [13]. To ensure the reliability of our results, we maintain consistent experimental settings, using models and hyperparameters as per their official implementations.

For ID evaluations, the test data originates from the same distribution as the training set. Conversely, the OOD setting involves evaluation using data from a different distribution, aiming to gauge the model's generalization capability.

4.3. Results

The results of ID and OOD mask reconstruction experiments are summarized in Table 1 and 2, respectively. Our results highlight that MINR consistently outperforms in both settings, achieving superior mask reconstruction with fewer parameters. For the CelebA dataset, commonly used in mask reconstruction evaluations, there was about 6.4dB improvement in ID despite having more than half the parameters reduced. Also, when evaluated on different data distribution from the training set, there was an improvement of more than 3dB for most cases. This is further depicted in Figure 1, where MINR exhibits enhanced clarity in reconstructing masked patches. Considering that MAE computes the loss exclusively on masked patches, we visualize the results with the pasted unmasked patches onto the reconstruction results for fair comparison.

5. Conclusion

In this work, we propose the MINR framework, a method that synergistically combines the principles of MIM with INRs to robustly tackle mask reconstruction tasks. MINR leverages the continuous functional approximation capacity of INRs to improve both ID and OOD performance. Our experimental evaluations against existing MAE approaches demonstrated the superiority of MINR in terms of reconstruction quality and robustness to diverse masking strategies, as substantiated by higher PSNR values across different datasets. Moreover, our proposed framework significantly reduces the model parameters, thereby alleviating the need for heavy pretrained dependencies. Finally, the adaptability of MINR's continuous function provides a flexible pathway for deriving feature embeddings across various downstream tasks. In the future, we plan to leverage the flexibility of MINR, derived from its ability to learn a continuous function and agnosticism to input image sizes, to showcase its performance and applicability across various downstream tasks.

6. Acknowledgement

Myungjoo Kang was supported by the NRF grant [2012R1A2C3010887] and the MSIT/IITP ([1711117093], [NO.2021-0-00077], [No. 2021-0-01343, Artificial Intelligence Graduate School Program (SNU)]).

References

[1] Isabela Albuquerque, Nikhil Naik, Junnan Li, Nitish Keskar, and Richard Socher. Improving out-of-distribution generalization via multi-task self-supervised pretraining. *arXiv preprint arXiv:2003.13525*, 2020. 1

[2] Ivan Anokhin, Kirill Demochkin, Taras Khakhulin, Gleb Sterkin, Victor Lempitsky, and Denis Korzhenkov. Image generators with conditionally-independent pixel synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14278–14287, 2021. 3

[3] Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *International Conference on Machine Learning*, pages 1416–1429. PMLR, 2023. 2

[4] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022. 2

[5] Yutong Bai, Zeyu Wang, Junfei Xiao, Chen Wei, Huiyu Wang, Alan L Yuille, Yuyin Zhou, and Cihang Xie. Masked autoencoders enable efficient knowledge distillers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24256–24265, 2023. 2

[6] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2

[7] Shuhao Cao, Peng Xu, and David A Clifton. How to understand masked autoencoders. *arXiv preprint arXiv:2202.03670*, 2022. 2

[8] Anthony Chen, Kevin Zhang, Renrui Zhang, Zihan Wang, Yuheng Lu, Yandong Guo, and Shanghang Zhang. Pimae: Point cloud and image interactive masked autoencoders for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5291–5301, 2023. 1

[9] Haijian Chen, Wendong Zhang, Yunbo Wang, and Xiaokang Yang. Improving masked autoencoders by learning where to mask. *arXiv preprint arXiv:2303.06583*, 2023. 2

[10] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8628–8638, 2021. 3

[11] Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Sdae: Self-distillated masked autoencoder. In *European Conference on Computer Vision*, pages 108–124. Springer, 2022. 2

[12] Yinbo Chen and Xiaolong Wang. Transformers as meta-learners for implicit neural representations. In *European Conference on Computer Vision*, pages 170–187. Springer, 2022. 2, 3

[13] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 4

[14] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Bootstrapped masked autoencoders for vision bert pretraining. In *European Conference on Computer Vision*, pages 247–264. Springer, 2022. 2

[15] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10995–11005, 2023. 2

[16] 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. 4

[17] Emilien Dupont, Hyunjik Kim, SM Eslami, Danilo Rezende, and Dan Rosenbaum. From data to functa: Your data point is a function and you can treat it like one. *arXiv preprint arXiv:2201.12204*, 2022. 3

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[19] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019. 1

[20] Jeremy Howard and Sylvain Gugger. Fastai: A layered api for deep learning. *Information*, 11(2):108, 2020. 3

[21] Lang Huang, Shan You, Mingkai Zheng, Fei Wang, Chen Qian, and Toshihiko Yamasaki. Green hierarchical vision transformer for masked image modeling. *Advances in Neural Information Processing Systems*, 35:19997–20010, 2022. 2

[22] Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. In *European Conference on Computer Vision*, pages 300–318. Springer, 2022. 2

[23] Chiheon Kim, Doyup Lee, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Generalizable implicit neural representations via instance pattern composers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11808–11817, 2023. 2, 3

[24] Lingjing Kong, Martin Q Ma, Guangyi Chen, Eric P Xing, Yuejie Chi, Louis-Philippe Morency, and Kun Zhang. Understanding masked autoencoders via hierarchical latent variable models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7918–7928, 2023. 2

[25] Youngwan Lee, Jeffrey Willette, Jonghee Kim, Juho Lee, and Sung Ju Hwang. Exploring the role of mean teach-

ers in self-supervised masked auto-encoders. *arXiv preprint arXiv:2210.02077*, 2022. 2

[26] Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. Semmae: Semantic-guided masking for learning masked autoencoders. *Advances in Neural Information Processing Systems*, 35:14290–14302, 2022. 2

[27] Siyuan Li, Di Wu, Fang Wu, Zelin Zang, Baigui Sun, Hao Li, Xuansong Xie, Stan Li, et al. Architecture-agnostic masked image modeling–from vit back to cnn. *arXiv preprint arXiv:2205.13943*, 2022. 2

[28] Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, et al. Mst: Masked self-supervised transformer for visual representation. *Advances in Neural Information Processing Systems*, 34:13165–13176, 2021. 2

[29] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 3

[30] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[31] Jiachun Pan, Pan Zhou, and YAN Shuicheng. Towards understanding why mask reconstruction pretraining helps in downstream tasks. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[32] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 2

[33] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *2009 IEEE conference on computer vision and pattern recognition*, pages 413–420. IEEE, 2009. 3

[34] Yuge Shi, N Siddharth, Philip Torr, and Adam R Kosiorek. Adversarial masking for self-supervised learning. In *International Conference on Machine Learning*, pages 20026–20040. PMLR, 2022. 2

[35] Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P Srinivasan, Jonathan T Barron, and Ren Ng. Learned initializations for optimizing coordinate-based neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2846–2855, 2021. 3

[36] Haochen Wang, Kaiyou Song, Junsong Fan, Yuxi Wang, Jin Xie, and Zhaoxiang Zhang. Hard patches mining for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10375–10385, 2023. 2

[37] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022. 2

[38] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142, 2023. 1, 2

[39] Qiangqiang Wu, Tianyu Yang, Ziquan Liu, Baoyuan Wu, Ying Shan, and Antoni B Chan. Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14561–14571, 2023. 1

[40] Quanlin Wu, Hang Ye, Yuntian Gu, Huishuai Zhang, Liwei Wang, and Di He. Denoising masked autoencoders are certifiable robust vision learners. *arXiv preprint arXiv:2210.06983*, 2022. 2

[41] Zhirong Wu, Zihang Lai, Xiao Sun, and Stephen Lin. Extreme masking for learning instance and distributed visual representations. *arXiv preprint arXiv:2206.04667*, 2022. 2

[42] Jiahao Xie, Wei Li, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Masked frequency modeling for self-supervised visual pre-training. *arXiv preprint arXiv:2206.07706*, 2022. 2

[43] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 2

[44] Yuchong Yao, Nandakishor Desai, and Marimuthu Palaniswami. Moma: Distill from self-supervised teachers. *arXiv preprint arXiv:2302.02089*, 2023. 2

[45] Qiang Zhou, Chaohui Yu, Hao Luo, Zhibin Wang, and Hao Li. Mimco: Masked image modeling pre-training with contrastive teacher. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4487–4495, 2022. 2