Enhanced Super-Resolution Using Cross Attention: Refining HMA for Local Image Restoration

Yuya Masuda
Department of Computer Science and Communications Engineering
Waseda University
Tokyo, Japan
msd_8172@fuji.waseda.jp

Hiroshi Ishikawa
Department of Computer Science and Communications Engineering
Waseda University
Tokyo, Japan
hfs@waseda.jp

***Abstract*—In this paper, we propose a novel method that integrates Cross Attention into the existing Hybrid Multi-Axis Aggregation Network for Image Super-Resolution(HMANet) to improve local super-resolution accuracy. While previous HMANet methods primarily focused on enhancing the resolution of the entire image, our approach emphasizes local image regions for more detailed restoration. By leveraging Cross Attention for context interaction, we achieve localized super-resolution with a focus on specific parts of the image. Our experiments demonstrate that the proposed method outperforms existing approaches in terms of accuracy, and shows promising results when evaluated with different loss functions. Our source code is available at https://github.com/msdsm/hmaca-for-local-image-restoration.**

I. Introduction

Image super-resolution is a crucial technique for restoring high-resolution images from low-resolution inputs, with widespread applications in fields such as surveillance, medical imaging, and computer vision. In the early stages, convolutional neural network (CNN)-based models like SRCNN [7] laid the foundation for super-resolution tasks by demonstrating the effectiveness of deep learning in enhancing image quality. Subsequently, generative adversarial network (GAN)-based models, such as SRGAN [8], introduced perceptual loss functions and adversarial training, resulting in more visually appealing and realistic images. Recently, Vision Transformer [9] based models have gained prominence due to their superior performance in image super-resolution tasks. Notable examples include Swin Transformer for Image Restoration (SwinIR) [1], Hybrid Attention Transformer(HAT) [2], Deep Residual Connected Transformer(DRCT) [3], and Hybrid Multi-Axis Aggregation Network for Image Super-Resolution(HMANet) [4]. These models primarily focus on enhancing the resolution of the entire image, aiming to improve overall image quality.

However, the tasks addressed by these models, such as those in HMANet, are originally designed for global resolution enhancement. Thus, they do not capable for fine-grained restoration of specific local regions in an image. In practice, scenarios, such as in surveillance footage or medical imaging, in which detailed restoration of local regions of an image, are becoming more common. Therefore, addressing this challenge is increasingly important.

In this paper, our contributions can be summarized as following:

1) We propose a task, localized image super-resolution, which aims to restore detailed parts of the image, and create a dataset for this purpose.
2) We introduce Cross Attention into HMANet to improve the model's ability to restore localized image regions in greater detail.
3) We design and explore various loss functions to test and improve the performance of the model in restoring localized areas of an image.

II. Related Works

A. HMANet

HMANet is a network composed of three main modules: Shallow Feature Extraction, Deep Feature Extraction, and Image Reconstruction, as shown in Fig. 1. Given a low-resolution input image $I_{LR} \in \mathbb{R}^{H \times W \times C_{in}}$, the high-resolution output image I_{HR} is obtained as follows:

$$\begin{aligned} F_0 &= H_{conv}\left(I_{LR}\right), \\ I_{HR} &= H_{REC}\left(H_{DF}\left(F_0\right) + F_0\right). \end{aligned} \tag{1}$$

Here, $H_{conv}(\cdot)$, $H_{DF}(\cdot)$, and $H_{REC}(\cdot)$ denote the Shallow Feature Extraction, Deep Feature Extraction, and Image Reconstruction, respectively.

B. Shallow Feature Extraction

The Shallow Feature Extraction module handles the initial phase of feature extraction in image processing and is constructed with a single convolutional layer. This module extracts fundamental low-level features, such as edges and textures, from the input image. As an essential preprocessing step, it ensures the efficient operation of the subsequent deep feature extraction, despite its relatively low computational cost.

C. Deep Feature Extraction

The Deep Feature Extraction stage is critical for extracting high-level features, enhancing the network's representational capacity. It is structured using multiple Residual Hybrid Transformer Blocks (RHTBs). Each RHTB integrates two distinct attention mechanisms: the Fused Attention Block (FAB) and

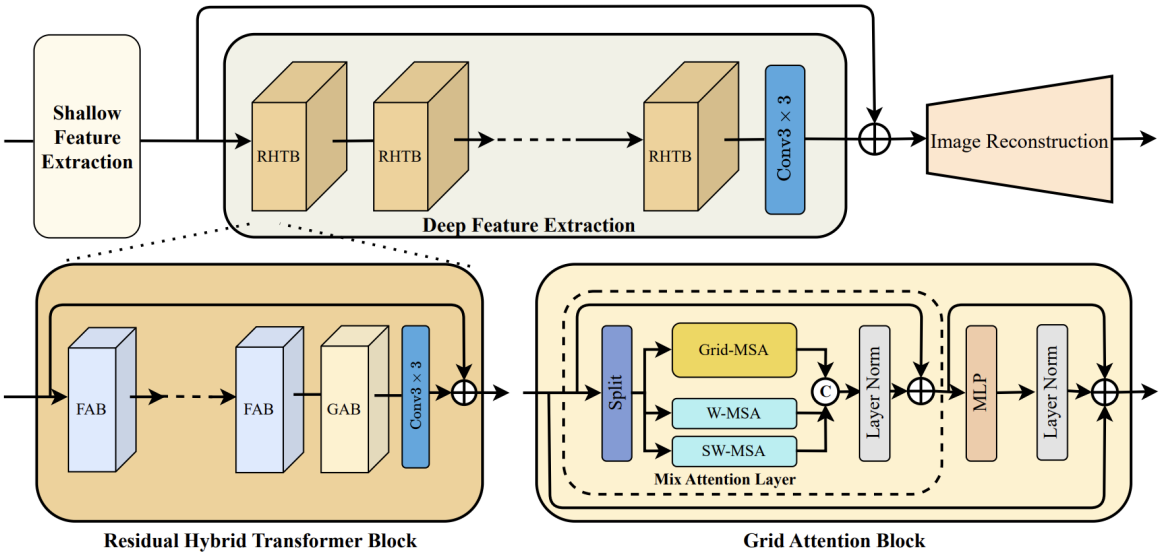

Fig. 1. Overview of HMANet, reproduced from the original paper [1].

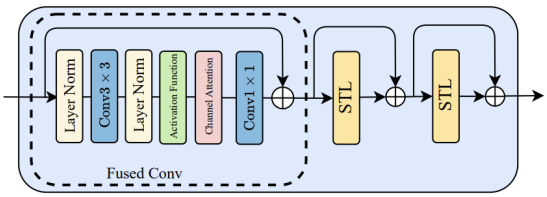

Fig. 2. The architecture of FAB, reproduced from the original paper [1].

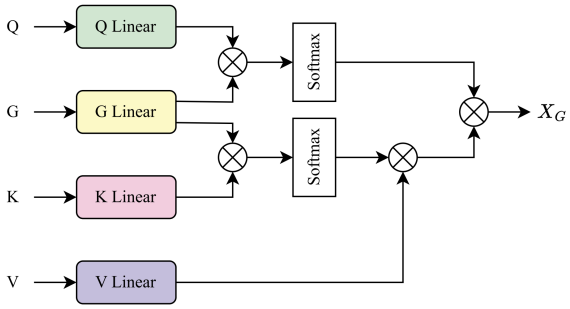

Fig. 3. Illustration of Grid Attention. Quoted from HMANet [1].

the Grid Attention Block (GAB), balancing the capacity of feature representation and efficiency in information flow.

1) Fused Attention Block (FAB): The Fused Attention Block (FAB) aims to capture both local and global features effectively by integrating Channel Attention with a vanilla attention mechanism, as illustrated in Fig. 2. Channel Attention emphasizes inter-channel correlations, focusing on significant channels to enhance learning capacity. Additionally, the Swin Transformer Layer (STL) processes local patch information efficiently, improving both computational efficiency and accuracy. This leads to a significant enhancement in the network's performance by extracting intricate details from the image.

2) Grid Attention: Given an input X, Q, G, K, and B are computed as:

$$\begin{aligned} Q &= XW^Q, \\ G &= XW^G, \\ K &= XW^K, \\ V &= XW^V. \end{aligned} \tag{2}$$

Grid Attention can then be expressed as:

$$\begin{aligned} \hat{X} &= \mathrm{SoftMax}\left(\frac{GK^\top}{d} + B\right) V, \\ \mathrm{Attention}\left(Q, G, \hat{X}\right) &= \mathrm{SoftMax}\left(\frac{QG^\top}{d} + B\right) \hat{X}. \end{aligned} \tag{3}$$

The illustration of Grid Attention is shown in Fig. 3.

3) Grid Attention Block (GAB): The input to the GAB $F_{in} \in \mathbb{R}^{H\times W\times C}$ is split into $F_G \in \mathbb{R}^{H\times W\times \frac{C}{2}}$, $F_{W_1} \in \mathbb{R}^{H\times W\times \frac{C}{4}}$, and $F_{W_2} \in \mathbb{R}^{H\times W\times \frac{C}{4}}$. The output of GAB F_{out} is described as follows:

$$\begin{aligned} X_{W_1} &= \mathrm{W-MSA}\left(F_{W_1}\right), \\ X_{W_2} &= \mathrm{SW-MSA}\left(F_{W_2}\right), \\ X_G &= \mathrm{Grid-MSA}\left(F_G\right) \\ X_{\mathrm{MAL}} &= \mathrm{LN}\left(\mathrm{Cat}\left(X_{W_1}, X_{W_2}, X_G\right)\right) + F_{in}, \\ F_{out} &= \mathrm{LN}\left(\mathrm{MLP}\left(X_{\mathrm{MAL}}\right)\right) + X_{\mathrm{MAL}} + F_{in}. \end{aligned} \tag{4}$$

Grid Attention Block (GAB) computes grid attention, shift window attention, and window attention along the channel

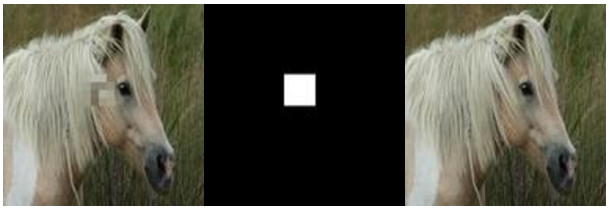

Fig. 4. Example of Horse-Pixelization. From left: degraded image, mask image, ground truth.

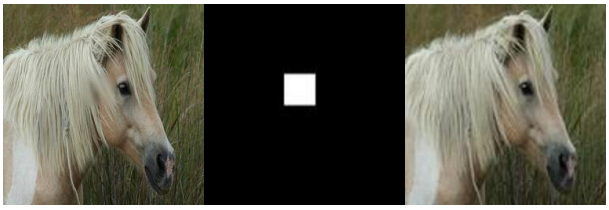

Fig. 5. Example of Horse-GaussianBlur. From left: degraded image, mask image, ground truth.

axis, subsequently concatenating them. This multi-scale feature fusion allows the network to capture both fine local details and broader contextual information. The channel division ratio of 2:2:1 ensures a balanced focus across different regions, enhancing the model's feature extraction capability while reducing redundancy.

D. Image Reconstruction

Image Reconstruction module generates a high-resolution output image from the extracted high-dimensional features. This phase employs sub-pixel convolution, a computationally efficient technique that reconstructs fine-grained pixel-level details, outperforming traditional upsampling methods. With its efficiency, sub-pixel convolution plays a pivotal role in generating visually superior high-resolution images.

III. Proposed Method

A. Dataset Creation and Degradation Techniques for Local Image Restoration

To improve the accuracy of local image restoration tasks, it is essential to have datasets that accurately replicate various types of image degradation. In this study, we investigated three degradation methods and constructed datasets based on these methods.

1) Pixelization Process: Pixelization is a technique that reproduces a state where image details are lost by coarsening specific regions. In this study, the target regions were divided into fixed-size blocks, and the pixel values within each block were averaged using a mean pooling technique. This process retains the contours of the image while losing fine details, thus generating degraded images with missing details.

2) Gaussian Blur Process: Blurring reduces visual sharpness and is a common method for image degradation. In this study, we adopted a Gaussian as blur kernel. By applying a Gaussian kernel, smooth blurring was applied either globally or locally, generating realistic blurred images.

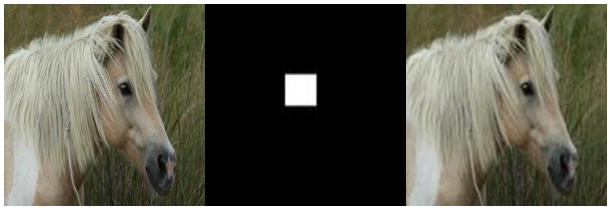

Fig. 6. Example of Horse-GaussianNoise. From left: degraded image, mask image, ground truth.

3) Gaussian Noise Addition Process: Noise addition involves the introduction of random noise to degrade image quality. We injected Gaussian noise in this study. By adjusting the noise variance, we simulated sensor errors and transmission noise introduced in the image capturing process.

Using these three degradation techniques, we processed both the horse dataset used in CycleGAN [5] and the ImageNet dataset, creating a total of six new datasets:

1) Horse-Pixelization
2) Horse-GaussianBlur
3) Horse-GaussianNoise
4) ImageNet-Pixelization
5) ImageNet-GaussianBlur
6) ImageNet-GaussianNoise

Figures 4, 5, and 6 illustrate examples of degraded images, mask images, and ground truth images for Horse-Pixelization, Horse-GaussianBlur, and Horse-GaussianNoise, respectively.

B. HMANet with Cross Attention

To address the task of generating high-quality images from degraded images and mask images, we propose HMANet with Cross Attention (HMACA). This architecture integrates Cross Attention conditioned on mask images into HMANet, as illustrated in Figure 7.

Let the output of the i-th Residual Hybrid Transformer Block (RHTB) be denoted as X_i $(i = 1, \ldots, L)$, and let M represents the feature map obtained by applying a convolutional layer to the mask image. The input F_{i+1} of the $(i+1)$-th RHTB is defined as follows:

$$\begin{aligned} Q_i &= X_i W^Q, \\ K &= M W^K, \\ V &= M W^V, \\ F_{i+1} &= \text{Attention}\left(Q_i, K, V\right), \end{aligned} \tag{5}$$

where W^Q, W^K, and W^V are the learned weight matrices for the query, key, and value projections, respectively. This Cross Attention mechanism allows the network to effectively utilize spatial and contextual information from the mask image, enhancing its ability to restore the degraded image.

Let the above calculation be denoted as $CA(X_i, M)$ to represent the Cross Attention mechanism. Given a low-resolution input image I_{LR} and a corresponding mask image I_M, the overall structure of HMACA can be formulated as follows:

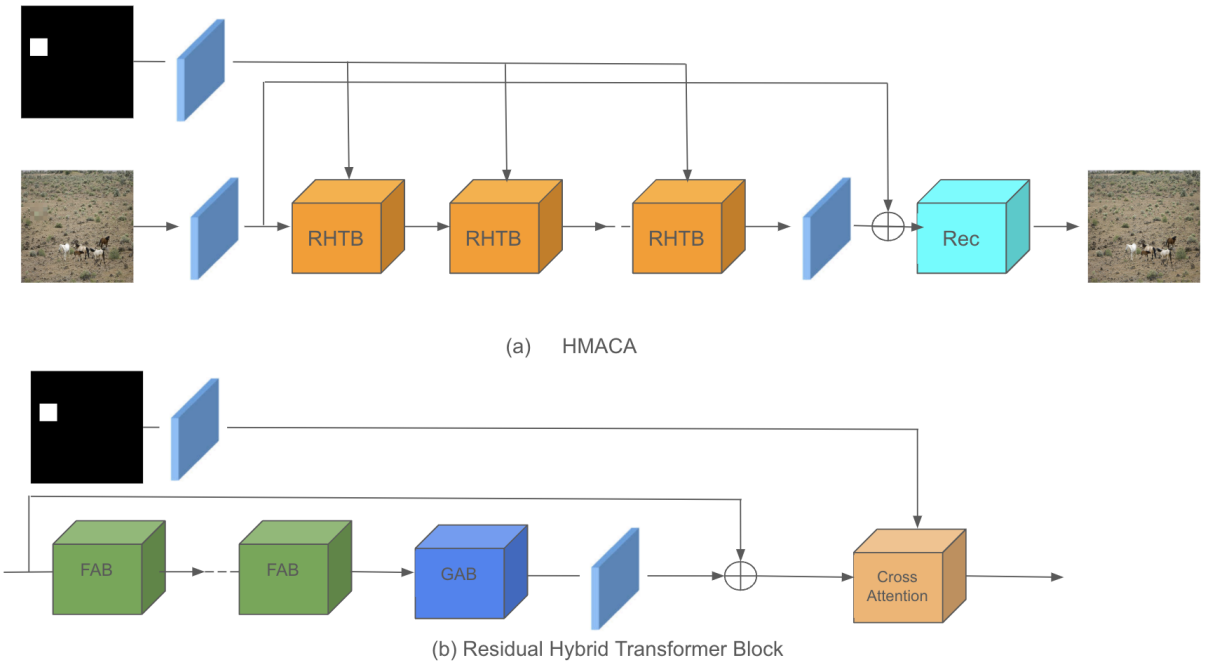

Fig. 7. Architecture of HMANet with Cross Attention (HMACA)

TABLE I
COMPARISON OF HMANET AND HMACA ON HORSE-PIXELIZATION

	HMA1	HMA2	HMACA1	HMACA2
PSNR	39.30	38.99	**39.32**	39.23

$$
\begin{aligned}
F_0 &= H_{conv}\left(I_{LR}\right), \\
M &= H_{conv}\left(I_M\right), \\
F_{i+1} &= CA\left(F_i, M\right) \quad (i = 0, \ldots, L-1), \\
F_{DF} &= H_{conv}\left(F_L\right), \\
I_{HR} &= H_{rec}\left(F_0 + F_{DF}\right),
\end{aligned} \tag{6}
$$

where H_{conv} denotes a convolutional operation and H_{rec} represents the reconstruction module. By leveraging the Cross Attention mechanism $CA(\cdot)$, the network efficiently integrates features from both the input image and the mask, enabling a more precise restoration of the high-resolution image I_{HR}. This formulation highlights the sequential refinement of feature maps through each Residual Hybrid Transformer Block (RHTB) and demonstrates the importance of integrating spatial context from the mask image throughout the process.

IV. EXPERIMENTS

A. HMANet vs HMACA: Ours

The quantitative results of HMANet and HMACA on the Horse-Pixelization dataset are shown in Table I, where the image size is set to 128×128. The model parameters for HMANet, denoted as upscale, are set to 1 and 2, referred to as HMA1 and HMA2, respectively. In HMA2, a convolutional layer is applied at the output to restore the image resolution from 2x back to the original size. Similarly, for HMACA, the upscale is also set to 1 and 2, referred to as HMACA1 and HMACA2. HMACA2 follows the same procedure as HMA2, applying a convolutional layer at the output to match the input image size. The evaluation metric adopted is Peak Signal-to-Noise Ratio (PSNR) [6]. To evaluate the quality of localized image super-resolution, the values of the non-mask regions in the output are aligned with the corresponding values from the input image before measuring.

B. Proposed Loss Function and Its Results

Next, we modified the loss function from L1 loss to a custom loss. Let M represent the pixel region corresponding to the mask, and N represent the non-mask region of the image. The model's output and the ground truth image are denoted as I_{HR} and I_{gt}, respectively. The loss function $\mathcal{L}$ is then defined as follows:

$$
\mathcal{L} = \alpha L_1\left(I_{HR}^{(M)}, I_{gt}^{(M)}\right) + (1-\alpha)\, L_1\left(I_{HR}^{(N)}, I_{gt}^{(N)}\right). \tag{7}
$$

Using this loss function, we conducted experiments with HMACA1 on the aforementioned 6 datasets. The resolution of the Horse dataset is 128×128, and the resolution of the ImageNet dataset is 64×64. The experimental results are shown in Tables II and III, where the effect of changing the value of α on performance is observed. Table I presents the results for the Horse dataset, while Table II shows the results for the ImageNet dataset.

The qualitative results of the best performing models on the Horse-pixelization, Horse-GaussianBlur, and Horse-

TABLE II
EXPERIMENTAL RESULTS OF VARYING α IN THE LOSS FUNCTION OF HMACA1 ON THE HORSE DATASET

	$\alpha = 0.0$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$	$\alpha = 1.0$
Horse-pixelization	39.33	**40.19**	39.99	40.11	39.81	39.98	39.91	39.81	39.64	39.33	38.26
Horse-GaussianBlur	46.91	51.14	51.21	**51.35**	51.06	50.62	51.13	50.76	50.59	50.43	48.00
Horse-GaussianNoise	**51.78**	51.16	51.01	51.14	51.02	50.91	50.57	50.38	50.86	50.50	49.91

TABLE III
EXPERIMENTAL RESULTS OF VARYING α IN THE LOSS FUNCTION OF HMACA1 ON THE IMAGENET

	$\alpha = 0.0$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$	$\alpha = 1.0$
ImageNet-pixelization	39.32	42.31	42.33	42.35	42.43	43.03	42.44	**42.46**	42.11	42.34	41.51
ImageNet-GaussianBlur	42.56	47.83	**48.88**	48.26	48.31	48.44	48.32	48.58	48.31	48.36	47.43
ImageNet-GaussianNoise	48.12	48.43	48.79	48.59	48.45	**49.10**	48.54	49.04	48.56	48.75	48.06

Fig. 8. Output example for Horse-pixelization

Fig. 9. Output example for Horse-GaussianBlur

GaussianNoise datasets are shown in Figures 8, 9, and 10, respectively. The results are presented in the following order from left to right: input image, mask image, output image, and ground truth image.

V. CONCLUSION

In this paper, we proposed the HMANet with Cross Attention (HMACA) for image local super-resolution tasks. By integrating Cross Attention into the HMANet architecture, we were able to condition the model by mask images, which enhanced the performance of the super-resolution process. We conducted experiments on multiple datasets, including Horse-pixelization, Horse-GaussianBlur, and Horse-GaussianNoise, and compared the performance of HMANet and HMACA models. The results demonstrated that HMACA consistently outperformed HMANet, particularly in terms of PSNR.

Furthermore, we introduced a novel loss function that incorporates a weight factor, α, which adjusts the influence of the L1 loss on different pixel regions. The experiments showed that tuning this parameter leads to significant improvements in the performance across various datasets.

Overall, our proposed HMACA model exhibited superior performance in restoring high-quality images from degraded inputs, and the adaptive loss function provided further enhancements in the restoration accuracy.

Acknowledgment. This work was partially supported by JSPS KAKENHI Grant Number JP20H00615.

Fig. 10. Output example for Horse-GaussianNoise

REFERENCES

[1] LIANG, Jingyun, et al. Swinir: Image restoration using swin transformer. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021. p. 1833-1844.

[2] Chen, Xiangyu, et al. Activating more pixels in image super-resolution transformer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[3] HSU, Chih-Chung; LEE, Chia-Ming; CHOU, Yi-Shiuan. DRCT: Saving Image Super-resolution away from Information Bottleneck. *arXiv preprint arXiv:2404.00722*, 2024.

[4] CHU, Shu-Chuan, et al. HMANet: Hybrid Multi-Axis Aggregation Network for Image Super-Resolution. *arXiv preprint arXiv:2405.05001*, 2024.

[5] ZHU, Jun-Yan, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the IEEE international conference on computer vision. 2017*. p. 2223-2232.

[6] HUYNH-THU, Quan; GHANBARI, Mohammed. Scope of validity of PSNR in image/video quality assessment. *Electronics letters*, 2008, 44.13: 800-801.

[7] DONG, Chao, et al. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 2015, 38.2: 295-307.

[8] LEDIG, Christian, et al. Photo-realistic single image super-resolution using a generative adversarial network. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017. p. 4681-4690.

[9] DOSOVITSKIY, Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.