

Supplementary Materials: Global Patch-wise Attention is Masterful Facilitator for Masked Image Modeling

Anonymous Authors

In the supplementary material, we first provide our implementation details to enable readers to fully reproduce our results in Section A. Next, in Section B, we provide the detail of our soft-to-hard mask generation. Finally, in Section C we visualize GPA and discuss its potential in further work.

A IMPLEMENTATION DETAILS

For all experiments in this paper, we utilized ImageNet-1K as the dataset for mask image modeling pre-training, and the detailed settings can be found in the Table S1 and Table S2. These specifics are also available in the code we provided.

Table S1: Pre-training settings.

setting	value
augmentation	RandomResizedCrop
model	ViT-B/16
training epochs	200
warmup epoch	10
optimizer	AdamW
base learning rate	1.5e-4
learning rate schedule	cosine decay
weight decay	0.05
layer-wise lr decay	1.0
batch size	4096

Table S2: Finetuning settings.

setting	value
augmentation	RandAug(9,0.5)
model	ViT-B/16
training epochs	100
warmup epoch	5
optimizer	AdamW
base learning rate	5e-4
learning rate schedule	cosine decay
weight decay	0.05
layer-wise lr decay	0.8
batch size	1024
label smoothing	0.1
mixup	0.8
cutmix	1.0
drop path	0.1

Algorithm S1 Soft-to-Hard Mask Generation

Input: Number of patches N ; Number of epochs E ; Global Attention Degree G ; Mask ratio α ; Soft mask ratio α_0^s and α_T^s ; Hard mask ratio α_0^h and α_T^h .

Output: Binary Mask $M = \{M_1; M_2; \dots; M_E\}$.

```
1:  $R = \text{random}(N)$ .
2: for  $epoch = 1$  to  $E$  do
3:   Training Step  $Tstep = epoch/E$ 
4:    $nM = \alpha * N$ 
5:    $nM_{soft} = \alpha_0^s + Tstep * (\alpha_T^s - \alpha_0^s)$ 
6:    $nM_{hard} = \alpha_0^h + Tstep * (\alpha_T^h - \alpha_0^h)$ 
7:   if  $nM_{soft} + nM_{hard} > nM$  then
8:      $nM_{soft} = nM - nM_{hard}$ 
9:   end if
10:   $nM_{rand} = nM - (nM_{soft} + nM_{hard})$ 
11:  if  $nM_{soft} > 0$  then
12:     $pM_{soft} = \text{sample}(prob = G, n = nM_{soft})$ 
13:     $R_{soft} = [2.0] * nM_{soft}$ 
14:     $R.\text{scatter}(pM_{soft}, R_{soft})$ 
15:  end if
16:  if  $nM_{hard} > 0$  then
17:     $pM_{hard} = \text{argsort}(x = G, n = nM_{hard})$ 
18:     $R_{hard} = [3.0] * nM_{hard}$ 
19:     $R.\text{scatter}(pM_{hard}, R_{hard})$ 
20:  end if
21:   $idSort = \text{argsort}(R)$ 
22:   $M_{epoch} = [1] * nM + [0] * (N - nM)$ 
23:   $M_{epoch} = M_{epoch}.\text{gather}(idSort)$ 
24: end for
25: return  $M$ 
```

B IMPLEMENTATION OF SOFT-TO-HARD MASKING

Algorithm S1 illustrates the details of our soft-to-hard masking. We implement soft masks through GPA-weighted sampling and hard masks through argsorting. The combination of soft, hard, and random masks is achieved by assigning values to a random variable. Specifically, we generate a random variable of length equal to the number of patches with values ranging from 0 to 1. After GPA-weighted sampling and argsorting, we assign the value 2 to all positions corresponding to the sampled locations and 3 to those corresponding to the argsorted locations. In this way, 2 and 3 always rank at the forefront of the random variable, and the value can effectively distinguishing different types of masks.



Figure S1: GPA overly focus on a small region.



Figure S2: GPA dispersed focus over a large region.

C DISCUSSION OF GLOBAL PATCH-WISE ATTENTION

We further visualized GPA on the validation sets of different datasets (ImageNet, COCO, ADE20K), as shown in Figure S3, S4 and S5. The model used for this visualization is pre-trained for 200 epochs on

ImageNet-1K. It can be observed that our GPA map accurately identifies high-semantic patches. This also demonstrates that GPA has strong generalization capabilities, as it has never seen any data from COCO and ADE-20k during the model's pre-training phase. However, due to the high-semantic patches being determined by the votes from all patches (sum of attention from all patches to a specific patch), GPA might overly focus or neglect a small region of a high-semantic patches (as shown in Figure S1). It may also be dispersed over large high-semantic patches (as shown in Figure S2). Exploring algorithms to enhance the ability of GPA in locating all high-semantic regions is a direction for further research. Additionally, given precise grasp of semantics, we can investigate more possibilities for its use as a feature in future work.

D MORE RESULTS

To demonstrate the universality of GPA on objects of different sizes, we conducted more in-depth experiments on the COCO dataset. The results from the Table Table S3 show that GPA outperforms MAE for objects of all sizes.

Model	AP_s	AP_m	AP_l	AP_{50}	AP_{75}
MAE	11.7	31.4	47.1	47.3	31.8
GPA	12.3	33.2	48.9	49.5	33.3

Table S3: More results on COCO.

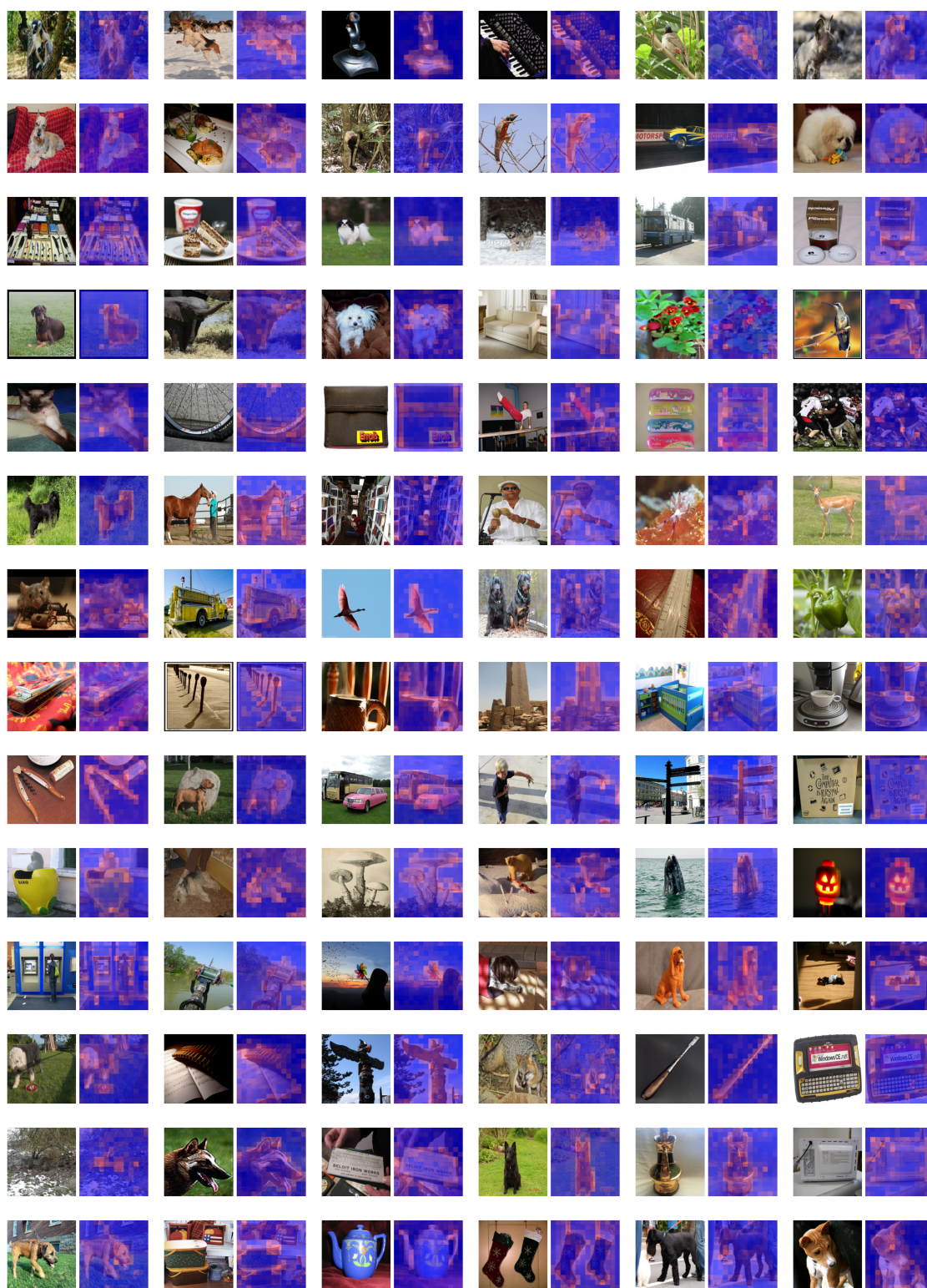


Figure S3: More visualization of GPA on ImageNet-1K validation set. For each tuple, we show the *input image* (left), *Global Patch-wise Attention* (right). Red means high GPA.

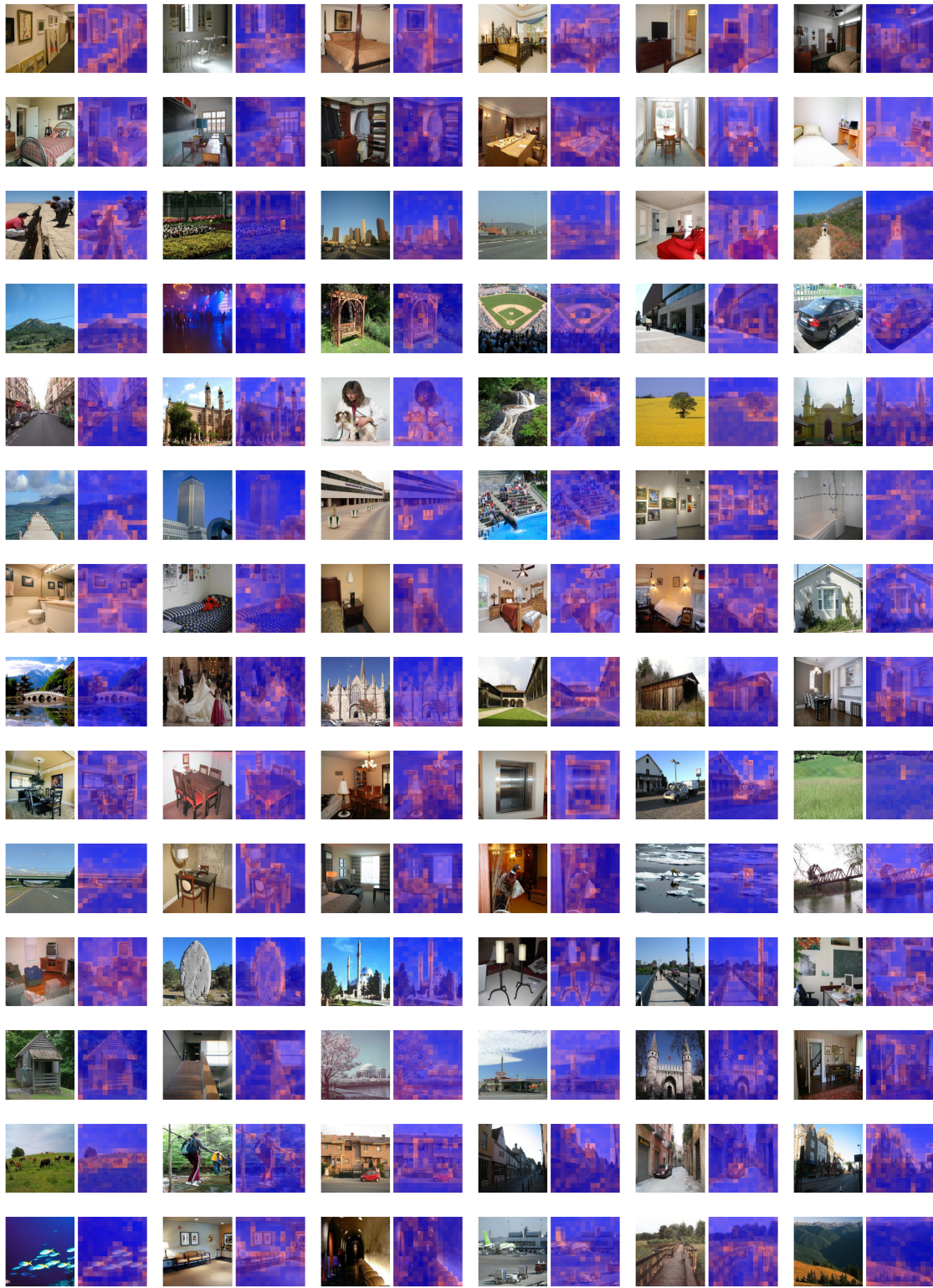


Figure S4: More visualization of GPA on ADE20k validation set. For each tuple, we show the *input image* (left), *Global Patch-wise Attention* (right). Red means high GPA.

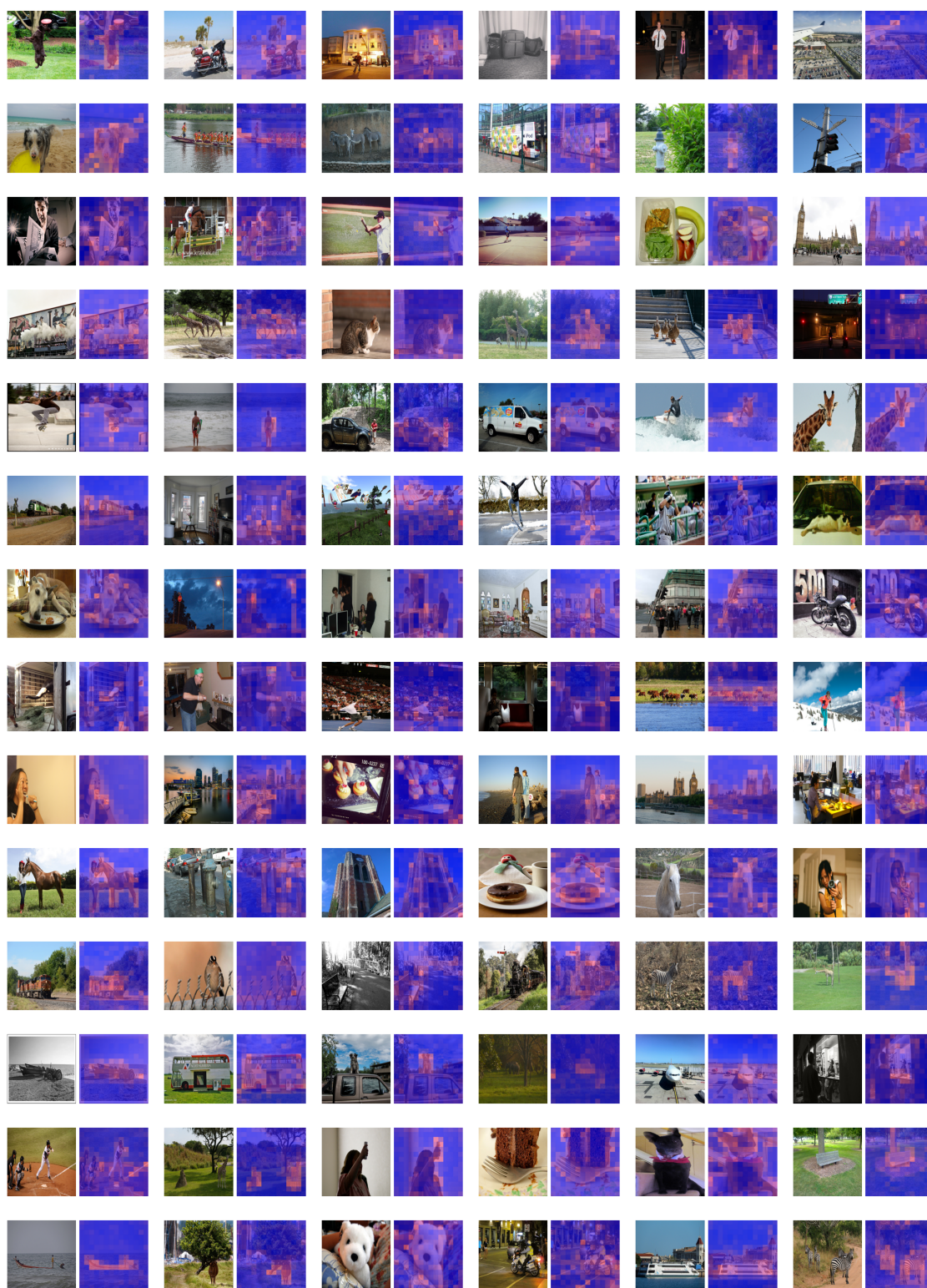


Figure S5: More visualization of GPA on COCO validation set. For each tuple, we show the *input image* (left), *Global Patch-wise Attention* (right). Red means high GPA.