Learning Gabor Texture Features for Fine-Grained Recognition Supplementary Materials

Lanyun Zhu[1] Tianrun Chen[2] Jianxiong Yin[3] Simon See[3] Jun Liu[1*]
Singapore University of Technology and Design [1] Zhejiang University [2] NVIDIA AI Tech Centre [3]
lanyun_zhu@mymail.sutd.edu.sg tianrun.chen@zju.edu.cn
{jianxiongy, ssee}@nvidia.com jun_liu@sutd.edu.sg

The supplementary materials are arranged as follows. In Sec. A, additional details are presented to provide a more comprehensive understanding of our method. In Sec. B, further experimental results are included to validate the effectiveness of our approach. In Sec. C and D, additional visualization results and analysis are provided to offer deeper insights into the functioning of our method.

A. More Details of Method

A.1. Derivation of Parameter Valid Ranges

In Sec 3.3 of the text, we propose a constraint to ensure the stability of model training. This constraint limits the values of Gabor filter parameters into their valid ranges, which can be derived based on the properties of Gabor filters and digital images in different domains. According to the periodicity of angles, the orientation parameter has a valid range of $[0, \pi]$. As for the other parameters, we provide a detailed derivation of their valid ranges in the subsequent sections.

Scale Parameter $[\sigma_x, \sigma_y]$**.** $[\sigma_x, \sigma_y]$ are the scale parameters that determine the filter effective size in both spatial and frequency domains. In spatial domain, a Gaussian function modulates the sinusoidal plane wave, which is defined in the infinite signal space to satisfy its mathematical properties [1]. However, an image is a finite length signal in the spatial domain, with its valid signal zone determined by the image width S. Specifically for an image with size $S \times S$, the valid zone for each axis can be expressed as $[-0.5S, 0.5S]$ with the image center as the coordinate origin. Previous research [1] indicates that directly applying infinite-length-defined filters to the finite-length image zone would cause mathematical deficiency, which could limit the effectiveness of Gabor filters due to waveform distortion. To alleviate the problem, we propose a solution to concentrate most of the Gabor filter energy within the finite signal zone. This ensures that only a small amount of filter energy spills over the finite signal, minimizing the negative effects of using infinite-length-defined filters on finite-length images. Specifically for a Gaussian with mean μ and variance σ, we constrain $[\mu - \alpha\sigma, \mu + \alpha\sigma]$ to fall in the valid signal zone of the image. α is a hyper-parameter to control the energy concentration degree. According to the experimental results shown in Table. 1, we choose 2.5 to be the optimal value for α. By doing so, 98.76% of the Gaussian energy can be subtended, and only 1.24% of filter energy spills out of the image signal, whose negative effect is negligible. Based on the above analysis, we derive the spatial-wise constraints for parameters $[\sigma_x, \sigma_y]$ as follows:

$$\begin{cases} [-0.25\sigma_x, 0.25\sigma_x] \subseteq [-0.5S, 0.5S] \\ [-0.25\sigma_y, 0.25\sigma_y] \subseteq [-0.5S, 0.5S] \end{cases} \tag{1}$$

We further analyze the frequency-wise constraints for $[\sigma_x, \sigma_y]$. We perform a Fourier transform on Eq. 1 of text and get the frequency-wise expression of Gabor filters as follows:

$$G(u, v) = \exp\left[-\frac{1}{2}\left(\left(4\pi^2\sigma_x^2 (u - W)^2\right) + 4\pi^2\sigma_y^2 v^2\right)\right], \tag{2}$$

As can be observed from Eq. 2, in frequency domain, the Gabor filter also contains a Gaussian with mean $\{W, 0\}$ and variance $\{\frac{1}{2\pi\sigma_x}, \frac{1}{2\pi\sigma_y}\}$ to control its effective size. The valid signal zone in frequency domain can be derived according to Nyquist sampling theorem, which indicates that for a given sample rate f_s, perfect reconstruction is guaranteed possible when the frequency $|W| < (f_s/2)$, otherwise signal aliasing would happen. In an image, the sample rate equals 1 pixel, so any frequency component larger than 0.5 is distorted thus being invalid. This means the valid signal zone in frequency domain is $[-0.5, 0.5]$. Following the constraints in spatial domain, we subtend 98.76% of the frequency-wise Gaussian energy into the valid signal zone to avoid distortion, getting the constraints in frequency do-

*Corresponding Author

α	Subtended Energy	Accuracy
1.0	68.27%	85.8
1.5	86.64%	89.0
2.0	95.45%	90.2
2.5	98.76%	90.8
3.0	99.73%	90.3
3.5	99.95%	90.0

Table 1. Ablation results of hyper-parameter α for constraining Gabor filter parameters. When the value of α is too small, the percentage of subtended energy is also low. This leads to a large amount of filter energy spilling out of the effective signal zone, which in turn negatively impacts training stability. Conversely, if α is too large, the filter parameters may be constrained to a small range, leading to a loss of information across certain frequencies. Experimental results show that the optimal value for α is 2.5. This choice strikes a balance between training stability and the availability of sufficient multi-frequency information.

mains as follows:

$$\begin{cases} [W - \frac{2.5}{2\pi\sigma_x}, W + \frac{2.5}{2\pi\sigma_x}] \subseteq [-0.5, 0.5] \\ [-\frac{2.5}{2\pi\sigma_y}, \frac{2.5}{2\pi\sigma_y}] \subseteq [-0.5, 0.5] \end{cases} \tag{3}$$

Solving Eq. 1 and Eq. 3, we get $[\frac{5}{2\pi(1-2W)}, \frac{S}{5}]$ and $[\frac{5}{2\pi}, \frac{S}{5}]$ to be the valid ranges for σ_x and σ_y respectively.

Frequency Parameter W. We further analyze the valid range for the frequency parameter W. Due to the symmetry of image frequencies, any W less than 0 is mirrored with its opposite number $-W$, so frequency components less than 0 are not considered and the lower bound for W is set to 0. The upper bound can be derived from two constraints. First, according to Nyquist sampling theorem, frequency should be lower than 1/2 to avoid aliasing. Second, the upper bound of σ_x should be higher than its lower bound. Mathematically, these constraints are formulated as follows:

$$\begin{cases} W < 0.5 \\ \frac{S}{5} > \frac{5}{2\pi\,(1-2W)} \end{cases} \tag{4}$$

Solving Eq.4, we get $\frac{2\pi S-25}{4\pi S}$ to be the upper bound and $[0, \frac{2\pi S-25}{4\pi S}]$ to be the valid range of W.

A.2. FPN Block in Region Selection Gate

The proposed region selection gate employs a FPN block to generate a feature F using the intermediate features of the CNN-based semantic branch, which is then used to assist in selecting informative regions for texture extraction (see Sec. 3.5 of the main paper for details). In Fig. 1, we show the detailed structure of the FPN block. The output channel number of all 1×1 convolution layers in the block is 128. This block integrates multi-level information from different

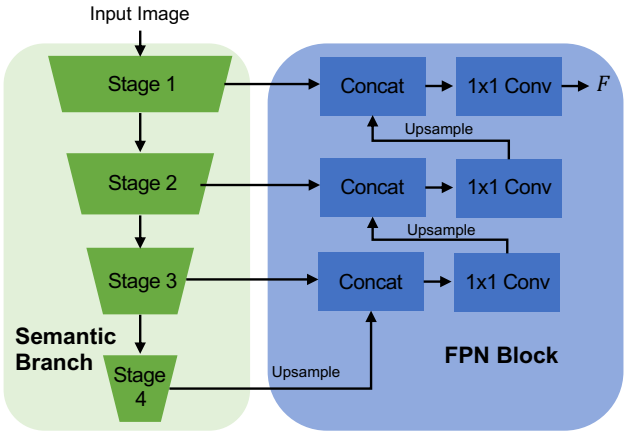


Figure 1. Structure of the FPN block in the proposed Region Selection Gate.

Method	Top-1 (%)
ResNet50	76.1
Ours (with ResNet50 backbone)	77.9

Table 2. Validation results on ImageNet.

layers. As a result, the generated F contains comprehensive information for the effective key part localization.

A.3. Back-Propagation of Improved Semantic Hashing.

In the proposed region selection gate, we employ the improved semantic hashing technique to make the selection operation differentiable. Specifically, for the k-th region proposal, a standard Gaussian noise is first added to its score s^k to produce $\hat{s}^k$. Then two vectors are generated from $\hat{s}^k$, including a binary discrete feature d^k and a continuously differentiable vector c^k (see Eq. 11 of the main paper for details). In forward-propagation, d^k is used to make region selection decisions. In back-propagation, we consider the gradient of c^k with respect to $\hat{s}^k$ an approximation of the gradients for updating the parameters from the discrete gate d^k. This gradient replacement operation could be realized by $d^k = d^k + c^k - c^k.\text{detach}()$ in PyTorch. During inference, we skip the Gaussian noise sampling step and directly use the discrete output from its original score as the selection decision, i.e., $\mathbb{1}\left(s^k > 0\right)$.

B. More Experimental Results

B.1. Experiments on ImageNet

In addition to fine-grained recognition datasets, we also validate our method on ImageNet, which is a widely-used dataset for general image classification. The results are presented in Table. 2. As a baseline, ResNet50 achieves Top-1 accuracy of 76.1%. By using our method with ResNet50 as the semantic branch, we obtain Top-1 of 77.9%, which outperforms ResNet50 by 1.8%. Despite achieving higher

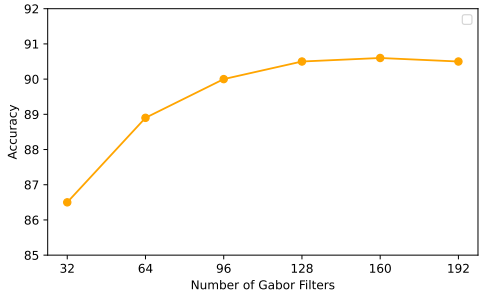


Figure 2. Ablation results of the Gabor filter number.

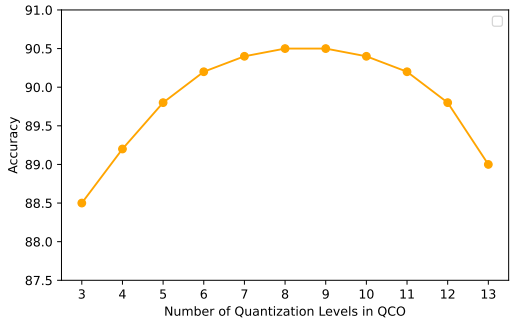


Figure 3. Ablation results for the number of quantization levels in LHO.

accuracy, we observe that our method can bring greater improvement on fine-grained datasets than ImageNet. This can be explained by the different types of features required for different datasets. Specifically, the visual appearances and semantic meanings of different categories in ImageNet are significantly different, allowing us to classify different classes from a global perspective without explicitly exploiting local details. As a result, the high-level semantic information captured by CNN is already sufficient to distinguish different classes, while local detailed textures captured by our method can be less crucial. In contrast, fine-grained recognition datasets often include categories with very similar visual appearances and high-level semantic meanings (e.g., different bird species). These categories are very similar from the global view, only having subtle differences in some local areas. In this scenario, features from deep CNNs are insufficient for classification due to their lack of local detailed features and high-frequency information, as discussed in the Introduction section of our paper. Texture information extracted from our method can serve as an effective supplement to CNN features, significantly improving fine-grained recognition.

B.2. Ablation Study of Hyper-Parameters

In this section, we present ablation results of hyper-parameters used in our method, including the number of Gabor filters, the number of quantization levels in LHO, λ in the loss function, and the size that each region is zoomed into. Experiments in this section are conducted on CUB-200-2011 with ResNet50 as the semantic branch. We report the average results of 5 repeated experiments

λ	Accuracy	Flops(G)
0.01	90.0	25.57
0.05	90.2	23.26
0.1	90.7	21.05
0.2	90.8	20.72
0.3	90.5	20.65
0.5	90.2	20.46
1	90.0	20.20

Table 3. Ablation results of λ in Eq. 13 of the main paper.

Ablation of Gabor Filter Number The texture branch in our approach utilizes N learnable Gabor filters to process input regions. In Fig. 2, we present the validation results of our method using varying numbers of learnable Gabor filters. As shown in Fig. 2, increasing N from 32 to 128 results in an improvement in validation accuracy from 86.5 to 90.5. However, performance improvement becomes insignificant when N exceeds 128. Therefore, we choose $N = 128$ as the optimal number of Gabor filters.

Ablation for the Number of Quantization Levels. The proposed LHO involves a step that quantizes the intensity map into M levels in order to extract statistical information (refer to Eq. 4 and Eq. 5 of the main paper for further details). In Fig. 3, we present the validation results of using different numbers of quantization levels. It is observed that when M is greater than 7 and less than 11, the accuracy remains stable and near 90.5. Conversely, when M is too small, the quantization is coarse, resulting in less effective statistical feature extraction and lower validation accuracy. Furthermore, when M is too large, overfitting may occur to hinder the model's effectiveness. Based on experimental results, we chose 8 as the setting for M. It is worth noting that our proposed method consistently outperforms the baseline ResNet50 significantly when M ranges from 3 to 13, thus demonstrating the high effectiveness of our approach.

Ablation of λ in Loss Function. As shown in Eq. 13 of the main paper, we use a hyper-parameter λ to control the trade-off between the two loss items. In Table. 3, we present the validation accuracy and average flops when λ is set to different values. The fluctuation of accuracy is less than 0.8 when λ varies from 0.01 to 1, showing that our method is non-sensitive to the choice of hyper-parameter λ.

Ablation of Size that Each Region is zoomed in. Each of the selected regions is zoomed into the size $S \times S$ before being feeding into the texture branch for feature extraction. In Fig. 4, we show the validation accuracy when setting S to different values. The accuracy keeps stable when S

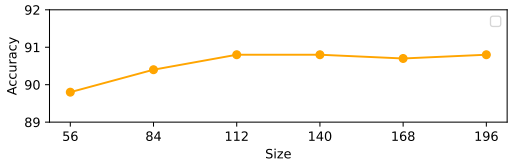


Figure 4. Ablation results for S indicating the size that each region is zoomed into.

Low-frequency Interval	High-frequency Interval	Accuracy
$[0, \frac{u_w}{5}]$	$[\frac{u_w}{5}, u_w]$	90.1
$[0, \frac{u_w}{4}]$	$[\frac{u_w}{4}, u_w]$	90.5
$[0, \frac{u_w}{3}]$	$[\frac{u_w}{3}, u_w]$	90.5
$[0, \frac{u_w}{2}]$	$[\frac{u_w}{2}, u_w]$	90.8
$[0, \frac{2u_w}{3}]$	$[\frac{2u_w}{3}, u_w]$	90.4
$[0, \frac{3u_w}{4}]$	$[\frac{3u_w}{4}, u_w]$	90.0
$[0, \frac{4u_w}{5}]$	$[\frac{4u_w}{5}, u_w]$	89.4

Table 4. Ablation results of different divisions for the low-frequency interval and high-frequency interval. $u_w = \frac{2\pi S - 25}{4\pi S}$ denotes the upper bound of frequency parameter W.

Low-frequency Filters	High-frequency Filters	Accuracy
$\frac{N}{5}$	$\frac{4N}{5}$	89.8
$\frac{N}{4}$	$\frac{3N}{4}$	90.3
$\frac{N}{3}$	$\frac{2N}{3}$	90.8
$\frac{N}{2}$	$\frac{N}{2}$	90.8
$\frac{2N}{3}$	$\frac{N}{3}$	90.7
$\frac{3N}{4}$	$\frac{N}{4}$	90.0
$\frac{4N}{5}$	$\frac{N}{5}$	89.4

Table 5. Ablation results of different amount allocations for the low-frequency filters and high-frequency filters.

is greater than or equal to 112. Setting a smaller value for S results in a reduced input size for the texture branch and lower computational costs. Thus, we choose 112 to be the setting of S.

B.3. Ablation of High Frequency Enhancement Strategies.

To alleviate the frequency-bias problem and enhance the high-frequency texture extraction capability of Gabor filters, we propose a high frequency enhancement strategy by setting two value intervals for frequency parameter W: $[0, \frac{2\pi S-25}{8\pi S}]$ and $[\frac{2\pi S-25}{8\pi S}, \frac{2\pi S-25}{4\pi S}]$, which are equally divided from the valid range $[0, \frac{2\pi S-25}{4\pi S}]$ of W. We then constrain W of $N/2$ Gabor filters to fall between $[0, \frac{2\pi S-25}{8\pi S}]$ and the other $N/2$ filters to fall between $[\frac{2\pi S-25}{8\pi S}, \frac{2\pi S-25}{4\pi S}]$, such that they serve as the low-frequency expert and high-frequency expert respectively. In Table. 4, we present the results obtained from applying different strategies with varying divisions for low-frequency and high-frequency intervals. We denote the upper bound of W as u_w, which equals $\frac{2\pi S-25}{4\pi S}$. In Table. 5, we present the results of varying amounts of allocations for low-frequency and high-frequency filters. From the results in both tables, it can be observed that the performance remains stable when the division ranges from $\frac{u_w}{4}$ to $\frac{2u_w}{3}$ and the amount of low-frequency filters ranges from $\frac{N}{3}$ to $\frac{2N}{3}$. The results demonstrate that our proposed method is not sensitive to these specific settings.

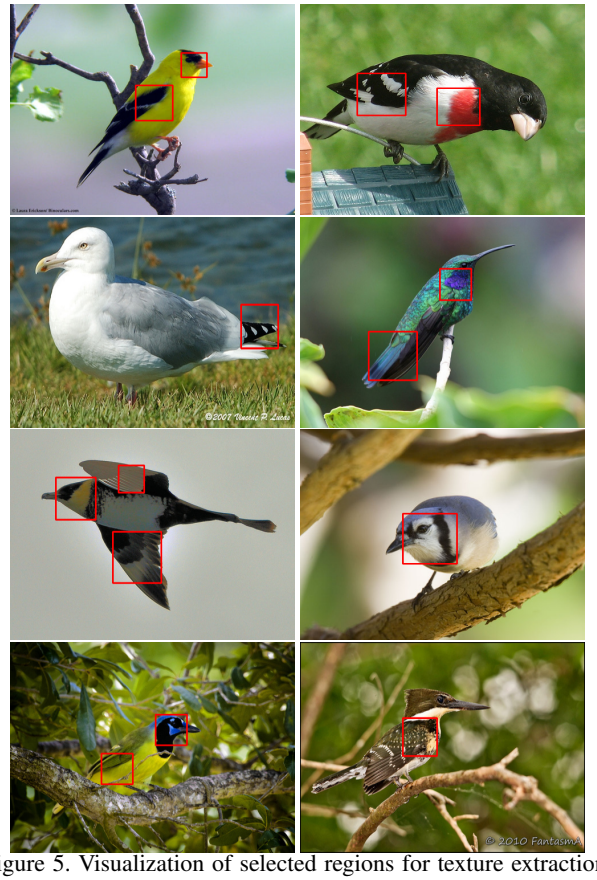


Figure 5. Visualization of selected regions for texture extraction. The selected regions are marked by the red bounding boxes.

C. More Visualization Results

C.1. Visualization of Selected Regions

In Fig. 5, we present some visualization results of the selected regions for texture extraction. These regions are marked by the red bounding boxes. The selected regions contain informative texture features that are difficult to be extracted by the vanilla CNNs. Using the proposed texture branch, we can extract effective texture features from these regions to facilitate fine-grained recognition.

C.2. Visualization for the Output of Gabor Filters

In Fig. 4 of the main paper, we have provided some visualization results of the outputs obtained from learned Gabor filters. In Fig. 8, we present more visualization results.

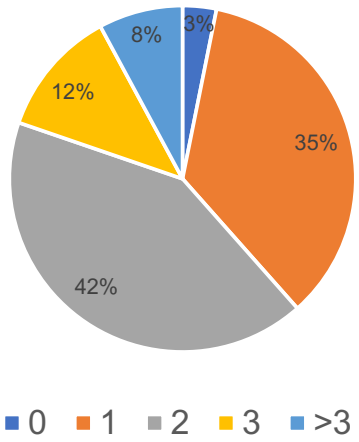


Figure 6. Percentage of images with different numbers of selected regions.

More specifically, Fig. 8 (c) and Fig. 8 (d) show the average output of all high-frequency and low-frequency Gabor filters, respectively. As can be observed, the high-frequency filters primarily capture information of undulating areas such as speckles and ripples, whereas the low-frequency filters primarily capture information related to smooth changing areas. Both kinds of information are critical for recognition. By exploiting sufficient and balanced multi-frequency features through the carefully-designed learnable Gabor filters, our method can leverage comprehensive information for effective fine-grained recognition.

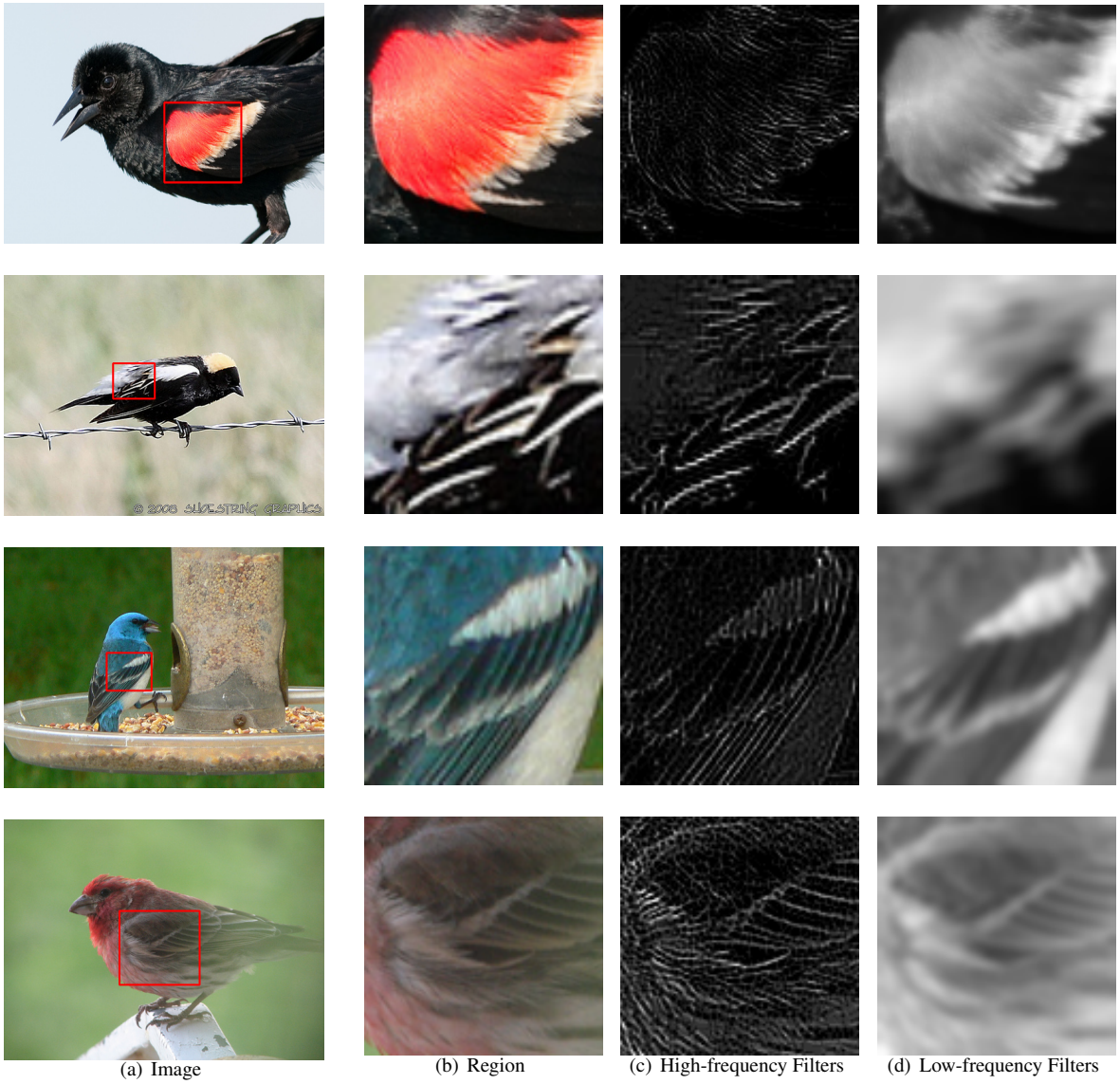
D. Statistical Analysis for the Number of Selected Regions.

Fig. 6 displays the percentage of images with various numbers of selected regions for texture extraction. The results indicate that, in general, only a few regions are selected for most images. This minimizes information redundancy and reduces computation costs. Specifically, 35% of all images have only one region selected, while 42% of all images have two regions selected for feature extraction. It is worth noting that a very small percentage of images have no regions selected for texture extraction. Fig. 7 illustrates two examples of such images. Typically, these images do not contain significant texture information that can facilitate recognition due to the low image quality or the category properties. Therefore, no region is selected for additional texture extraction.

Figure 7. Two image examples that have no region to be selected for texture extraction.

References

[1] Kenneth Steiglitz. *Digital Signal Processing Primer*. Courier Dover Publications, 2020. 1



(a) Image (b) Region (c) High-frequency Filters (d) Low-frequency Filters

Figure 8. Visualization of output from Gabor filters. (a), (b), (c) and (d) present the original images, the selected regions, average output of all high-frequency and low-frequency Gabor filters, respectively.