LSNet: See Large, Focus Small

Ao Wang[1] Hui Chen[2*] Zijia Lin[1] Jungong Han[3] Guiguang Ding[1]
[1]School of Software, Tsinghua University [2]BNRist, Tsinghua University
[3]Department of Automation, Tsinghua University
wanga24@mails.tsinghua.edu.cn jichenhui2012@gmail.com linzijia07@tsinghua.org.cn
jungonghan77@gmail.com dinggg@tsinghua.edu.cn

A. Implementation and Architectural Details

A.1. Implementation Details

For image classification on ImageNet-1K [2], we adopt the same training recipe as [9, 15, 21]. Specifically, we employ the standard image size of 224×224 for both training and testing. All models are trained from scratch for 300 epochs. We use the AdamW optimizer [17] with a cosine learning rate scheduler. The initial learning rate is set to 4×10^{-3}, and the total batch size is set to 2048. For data augmentation, we leverage mixup [26], RandAugment [1], CutMix [25], and random erasing [27], *etc*. Tab. 1 provides the training details of LSNet.

For object detection and instance segmentation on COCO-2017 [14], we employ the same training setting as [12, 15, 21]. Specifically, we utilize the AdamW optimizer and train the model for 12 epochs with a batch size of 16. The training resolution is 1333×800 and the initial learning rate is set to 2×10^{-4}. The learning rate decays with a rate of 0.1 at the 8-th and 11-th epochs. We initialize the backbones with the pretrained ImageNet-1K weights.

For semantic segmentation on ADE20K [28], following [13, 21], all models are trained for 40K iterations by the AdamW [17] optimizer with a batch size of 32. We adopt the poly learning rate schedule with the power of 0.9 and the initial learning rate of 2×10^{-4}, like [13, 21]. We employ the training resolution of 512×512 and report the single scale testing results on the ADE20K validation set, as in [20, 24]. The backbone models are initialized with the pretrained weights on ImageNet-1K.

For robustness evaluation, following [16, 19, 21], we employ the ImageNet-C [6], ImageNet-A [8], ImageNet-R [7], and ImageNet-Sketch [22] benchmarks. Specifically, ImageNet-C consists of algorithmically generated corruptions that are applied to the ImageNet test set. ImageNet-A contains naturally occurring examples misclassified by ResNets [4]. ImageNet-R comprises natural renditions of object classes in ImageNet, incorporating various textures and image statistics. ImageNet-Sketch includes white and black sketches of all ImageNet classes, gathered through google image queries.

Table 1. Training details on ImageNet-1K.

Model	LSNet-T/S/B
optimizer	AdamW
batch size	2048
training epochs	300
LR schedule	cosine
learning rate	0.004
warmup epochs	5
weight decay	0.025/0.025/0.05
augmentation	RandAug(9, 0.5)
random erase	0.25
color jitter	0.4
mixup	0.8
cutmix	1.0
gradient clip	0.02
label smooth	0.1

Table 2. Architectural details of LSNet variants.

Stage	Resolution	Type	Config	LSNet		
				T	S	B
stem	$\frac{H}{2}\times\frac{W}{2}$	Convolution	channels	16	24	32
	$\frac{H}{4}\times\frac{W}{4}$	Convolution	channels	32	48	64
	$\frac{H}{8}\times\frac{W}{8}$	Convolution	channels	64	96	128
1	$\frac{H}{8}\times\frac{W}{8}$	LS Block	channels	64	96	128
			blocks	0	1	4
2	$\frac{H}{16}\times\frac{W}{16}$	LS Block	channels	128	192	256
			blocks	2	2	6
3	$\frac{H}{32}\times\frac{W}{32}$	LS Block	channels	256	320	384
			blocks	8	8	8
4	$\frac{H}{64}\times\frac{W}{64}$	MSA Block	channels	384	448	512
			blocks	10	10	10

*Corresponding author.

A.2. Architectural Details

Tab. 2 presents the architectural details of LSNet variants, which are distinguished by the number of blocks and the number of channels within each stage.

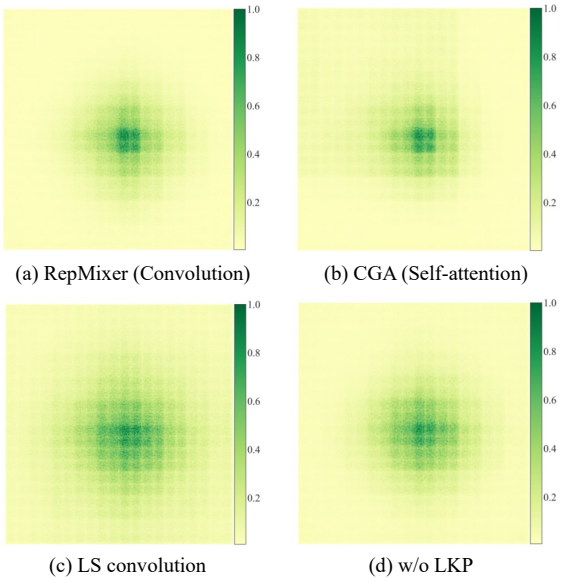
B. More Comparisons

We present more comparisons between LS convolution and others from mathematical perspectives. Specifically, for simply combining large-kernel with small-kernel convolutions, it follows the similar perception $\mathcal{P}_{conv}$ and aggregation $\mathcal{A}_{conv}$ processes as the standard convolution, *i.e.*, leveraging relative positions for relationship modeling and static kernel weights for feature integration. However, compared with LS convolution, it suffers from the limited modeling capability due to the lack of adaptability for different contexts. In other dynamic ways, Involution [11] leverages MLP for perception $\mathcal{P}_{inv}$ to derive the aggregation weights conditioned on x_i. Its aggregation $\mathcal{A}_{inv}$ then use the weights to convolve the features in $\mathcal{N}_K(x_i)$ with the process of $y_i = \mathcal{A}_{inv}(\mathcal{P}_{inv}(x_i), \mathcal{N}_K(x_i)) = \text{MLP}(x_i) \circledast \mathcal{N}_K(x_i)$. Although the aggregation process is dynamic, its perception process is confined to x_i, which leads to inadequate neighborhood relationship modeling compared with LS convolution. Additionally, CondConv [23] proposes per-example routing with global average pooling and MLP to linearly combining multiple convolution kernels for the aggregation weights in its perception $\mathcal{P}_{cond}$. Its aggregation $\mathcal{A}_{cond}$ then convolves the features in $\mathcal{N}_K(x_i)$ with the weights. Its process $y_i = \mathcal{A}_{cond}(\mathcal{P}_{cond}(X), \mathcal{N}_K(x_i))$ can be formulated as $y_i = (\sum \text{MLP}(\text{GAP}(X)) \cdot W_{cond}) \circledast \mathcal{N}_K(x_i)$. However, unlike LS convolution, CondConv leverages example-dependent perception, which prevents distinct tokens to adapt to diverse contexts.

C. Qualitative Analyses

C.1. Analyses for LS Convolution

We present the visualization analyses to qualitatively show the effectiveness of LS convolution. Specifically, we employ the effective receptive field [3, 18] method to compare LS convolution with convolution and self-attention, based on LSNet-T. We introduce the state-of-the-art RepMixer [21] and CGA [15] as the representatives of convolution and self-attention, respectively. Besides, we simply replace all LS convolutions in the model with others. As shown in Fig. 1, RepMixer and CGA suffer from the unnatural patterns, caused by static convolution kernels and window-based self-attention, respectively. In contrast, LS convolution enjoys both central area focusing and extensive peripheral viewing, showing smooth visual processing. Meanwhile, compared with "w/o LKP" where the large-kernel depth-wise convolution in the LKP is removed, LS convolution exhibits an enlarged effective receptive field. It is attributed to the ability of LKP to efficiently capture broad contextual information.



(a) RepMixer (Convolution) (b) CGA (Self-attention)

(c) LS convolution (d) w/o LKP

Figure 1. Visualization of the effective receptive field. Best viewed when zoomed in. (a) and (b) show that RepMixer and CGA exhibit unnatural patterns in the effective receptive field. (c) illustrates that LS convolution enables broad peripheral perception and central view focusing simultaneously. (d) shows that without LKP, LS convolution presents a smaller receptive field compared with (c), indicating the effectiveness of LKP.

Furthermore, we conduct visualization for the aggregation weights in LS convolution. Specifically, we obtain the cumulative value of the aggregation coefficients corresponding to each token in all aggregation processes it is involved in. We then visualize the average of the absolute values of all channels in the last layer at the third stage and perform upsampling for display. As shown in Fig. 2, the aggregation weights of SKA enjoy favorable interpretability. They effectively strengthen semantically relevant vision regions and accurately capture discriminative patterns in images. Besides, compared with "w/o LKP", LS convolution exhibits more precise emphasis on important visual areas, showcasing the improved modeling of spatial relationships facilitated by LKP. Based on LKP and SKA, LS convolution can thus help the model to grasp the critical visual information under limited computational costs, enhancing both efficiency and effectiveness.

Besides, we also visualize the feature maps generated by the LKP and SKA for more inspection. Specifically, we use the features after the large-kernel depth-wise convolution and the small-kernel dynamic convolution in the first stage for demonstration. As shown in Fig. 3, the feature maps produced by LKP exhibit a broad receptive field, capturing a wide range of contextual information in the scene. This

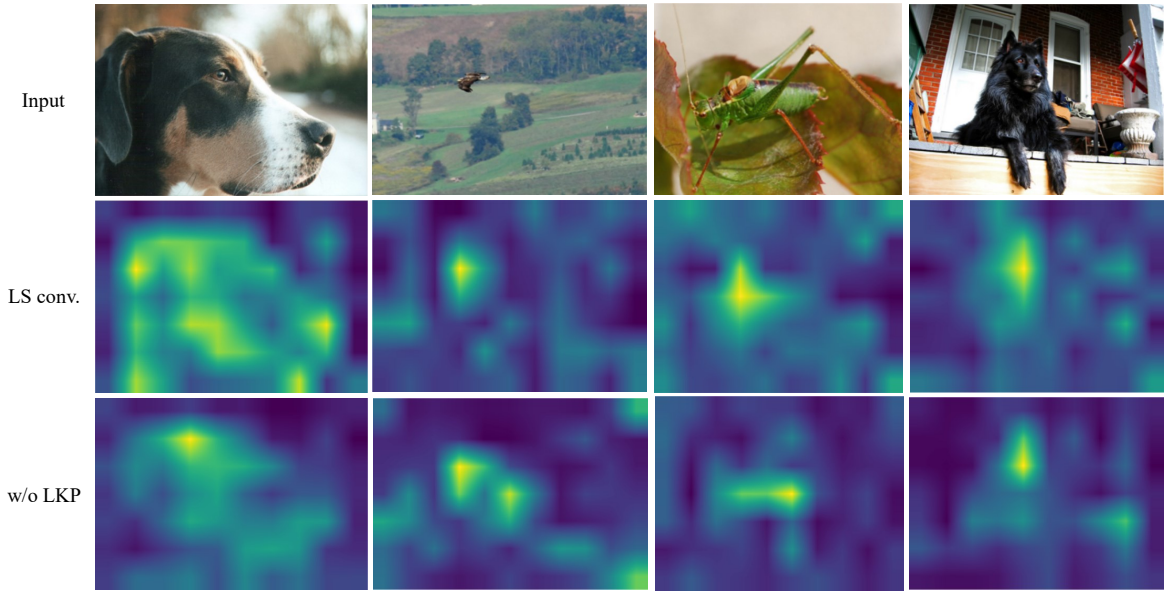


Figure 2. Visualization of the aggregation weights in LS convolution. The second row shows that the aggregation weights are well correlated with semantic relevant areas. The third row indicates that integrating LKP enables LS convolution to capture more precise visual patterns with improved contextual information.

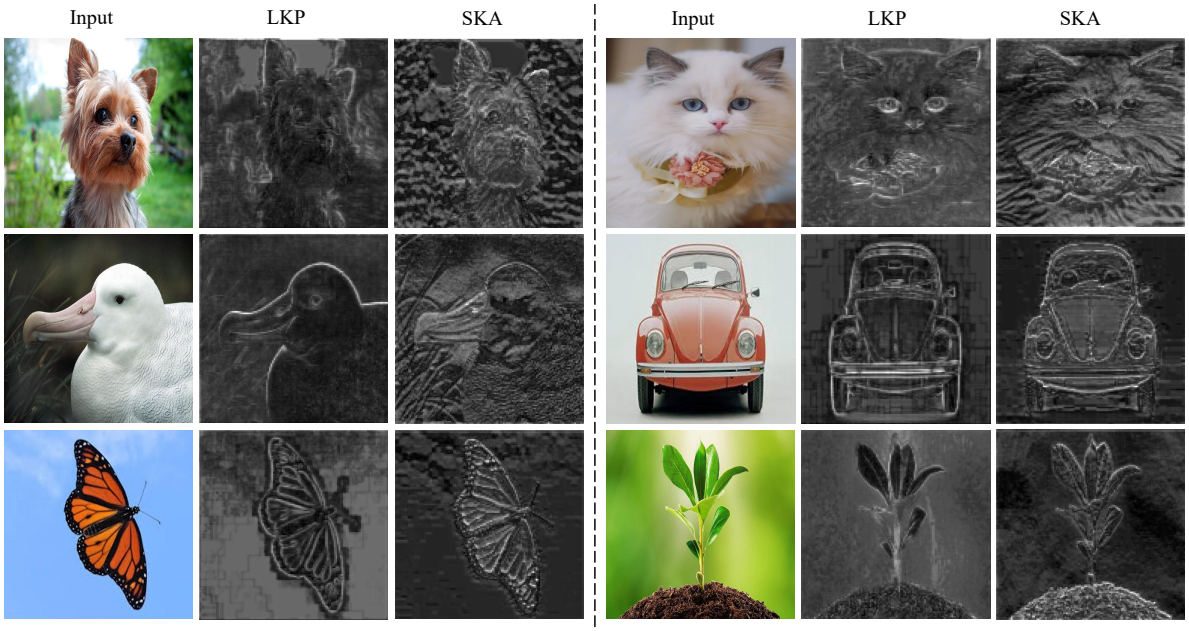


Figure 3. Visualization of the feature maps of LKP and SKA. The second column in each part shows that LKP can encompass a broad view of the scene. The third column in each part indicates that based on LKP, SKA can further grasp more subtle features and detailed patterns.

characteristic is reminiscent of the human peripheral vision system, adept at sensing the general surroundings. On the other hand, based on LKP, SKA further demonstrates the ability to grasp finer details within the image. It can result in more subtle features like gradients of hairs and clear outlines. This behavior is analogous to the human central vision system, which excels at discerning fine details and high-resolution information. Thanks to them, LS convolution can well help the model achieve the effective and efficient perception and aggregation processes.

Figure 4. Qualitative results for object detection and instance segmentation on COCO-2017 [14].

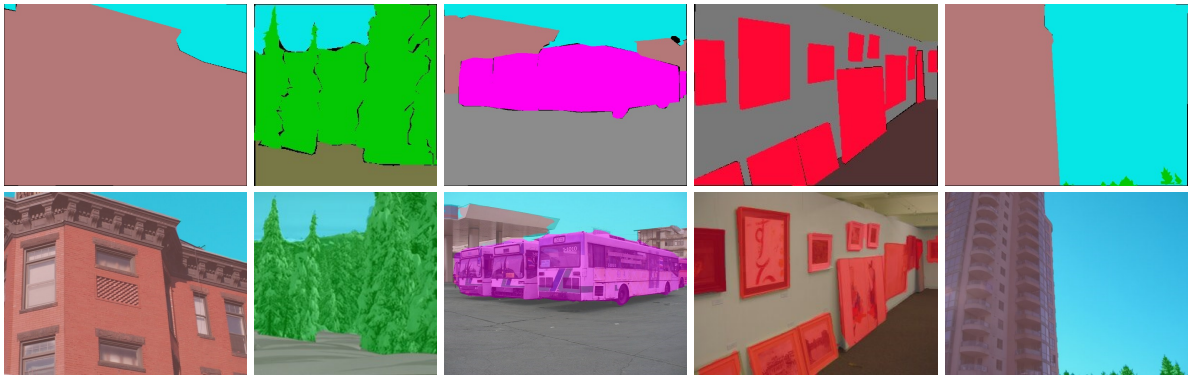


Figure 5. Qualitative results for semantic segmentation on ADE20K [28]. The upper row shows the ground truth masks, and the lower row presents the predicted masks.

C.2. Analyses for Downstream Tasks

We present the qualitative results when integrating LSNet into the Mask-RCNN framework [5] for object detection and instance segmentation tasks, and into the Semantic FPN framework [10] for the semantic segmentation task. As illustrated in Fig. 4, the model can achieve precise detection and segmentation of instances in diverse images. Besides, as shown in Fig. 5, the model demonstrates the ability to generate high-quality semantic segmentation masks.

D. Contribution, Limitation, and Impact

Contribution. In summary, our contributions are threefold, as follows:

1. We advocate a new strategy "See Large, Focus Small", inspired by the human vision system, for lightweight and efficient network design. By encompassing a broad perceptual range with enriched contextual information, it facilitates focused feature aggregation, fostering detailed visual understanding.
2. We propose LS convolution as a novel operation for modeling visual features in lightweight models. LS convolution integrates large-kernel perception and small-kernel aggregation, enabling proficient processing of visual information through both effective and efficient perception and aggregation processes.
3. We present a new family of lightweight vision networks, namely LSNet, which is built on LS convolution. Extensive experiments demonstrate that LSNet achieves the state-of-the-art performance and efficiency trade-offs compared with other lightweight networks across a broad range of vision tasks.

Limitation. Due to the limited computational resources, we do not extend the application of our LSNet to other scenarios, such as visual-language tasks or unsupervised learning. We do not investigate the pretraining of LSNet

on large-scale datasets, *e.g.*, ImageNet-21K [2], due to the same reason. However, we are enthusiastic about exploring more applications for LSNet in the future.

Societal Impact. We observe that this study is purely academic, and we have not identified any direct negative social impact resulting from our work. Nevertheless, we acknowledge the potential for malicious use of our models, which is a concern that affects the field. While we believe that it should be mitigated, discussions concerning this matter are beyond the scope of this paper.

References

[1] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 5

[3] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11963–11975, 2022. 2

[4] 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*, pages 770–778, 2016. 1

[5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4

[6] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 1

[7] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 1

[8] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 1

[9] Zhipeng Huang, Zhizheng Zhang, Cuiling Lan, Zheng-Jun Zha, Yan Lu, and Baining Guo. Adaptive frequency filters as efficient global token mixers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6049–6059, 2023. 1

[10] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019. 4

[11] Duo Li, Jie Hu, Changhu Wang, Xiangtai Li, Qi She, Lei Zhu, Tong Zhang, and Qifeng Chen. Involution: Inverting the inherence of convolution for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12321–12330, 2021. 2

[12] Yanyu Li, Geng Yuan, Yang Wen, Ju Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. Efficientformer: Vision transformers at mobilenet speed. *Advances in Neural Information Processing Systems*, 35: 12934–12949, 2022. 1

[13] Yanyu Li, Ju Hu, Yang Wen, Georgios Evangelidis, Kamyar Salahi, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Rethinking vision transformers for mobilenet size and speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16889–16900, 2023. 1

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 4

[15] Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan. Efficientvit: Memory efficient vision transformer with cascaded group attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14430, 2023. 1, 2

[16] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 1

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[18] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29, 2016. 2

[19] Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 12042–12051, 2022. 1

[20] Junting Pan, Adrian Bulat, Fuwen Tan, Xiatian Zhu, Lukasz Dudziak, Hongsheng Li, Georgios Tzimiropoulos, and Brais Martinez. Edgevits: Competing light-weight cnns on mobile devices with vision transformers. In *European Conference on Computer Vision*, pages 294–311. Springer, 2022. 1

[21] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. Fastvit: A fast hybrid vision transformer using structural reparameterization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5785–5795, 2023. 1, 2

[22] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[23] Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolu-

tions for efficient inference. *Advances in neural information processing systems*, 32, 2019. 2

[24] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10819–10829, 2022. 1

[25] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 1

[26] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 1

[27] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13001–13008, 2020. 1

[28] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 1, 4