Variable resolution improves visual question answering under a limited pixel budget

Andrey Gizdov[1], Shimon Ullman[1], and Daniel Harari[1]

Weizmann Institute of Science, Rehovot 7610001, Israel
{andrey.gizdov, shimon.ullman, hararid}@weizmann.ac.il

Abstract. AI-based systems for visual scene understanding benefit from a large field of view (FOV). Multiple camera systems extend the FOV, but larger and higher-quality images strain acquisition, communication, and computing resources. Sub-sampling the FOV effectively addresses this challenge without compromising performance on complex tasks that require fine visual cues and contextual information. We demonstrate that a variable sampling scheme, inspired by human vision, outperforms uniform sampling in several visual question answering (VQA) tasks with a limited sample budget (3% of full resolution). Specifically, we show accuracy gains of 3.7%, 2.0%, and 0.9% on the GQA, VQAv2, and SEED-Bench datasets, respectively. This improvement, achieved without image scanning, holds regardless of the fixation point location, as confirmed by control experiments. The results show the potential of the biologically inspired image representation to improve the design of visual acquisition and processing models in future AI systems.

Keywords: Visual question answering · Efficient vision · Variable resolution

1 Introduction

There is consistent evidence in nature that highly developed species utilize large field of views (FOV) to cope with fundamental visual tasks, including efficient detection of danger, food and social agents (19; 26). Similar to biological systems, an artificial intelligence system, designed to operate autonomously in natural environments, will require a visual system with a large FOV. Indeed, self-driving cars utilize multiple cameras at various viewing directions, to gain a wide FOV of the surroundings (20). On the other hand, many visual tasks also require fine details at high resolution; for example, threading a needle. Integrating those two requirements by utilizing images of increasing size and quality presents a great challenge to the acquisition and compute resources of AI-based systems (22; 27).

A simple solution to these challenges is to sub-sample a sufficiently large FOV, *i.e.* picking pixels at sparse locations such that the total count remains significantly low. Evolution provided an elegant and efficient solution in the form of a variable-resolution visual system, which acquires images at high resolution only in a small region at the center of the visual field (a.k.a. fovea centralis). In

Fig. 1: Visual question answering example. (a) baseline full resolution, **(b)** variable resolution, **(c)** uniform resolution. The model yields the correct answer when applied to the variable resolution image, with a mere 3% sample budget. While the uniform resolution is sufficient to recognize the giraffes (which normally live in Africa), only the variable resolution provides the fine details of the artificial shade, which is critical to answer the question correctly.

the rest of the visual field, resolution decreases with eccentricity (distance from the center). As a result, the human FOV spans over 120°, with peak resolution approaching 0.5 arcminutes (18). The brain provides mechanisms to facilitate visual perception by combining foveal and peripheral vision (2; 7; 21; 29).

In this work, we study the behavior of deep neural network models performing complex visual tasks when applied to variable resolution images, inspired by the human visual system, under an extremely limited pixel budget. In particular, we focus on the task of visual question answering (VQA). The VQA task is highly related to the task of complex visual scene understanding, for which we hypothesize that the contribution of a variable resolution system is significant (see example in Fig. 1). Understanding object relations and interactions in scenes, as well as object attributes, requires the integration of fine visual cues with contextual information, which are both available with the variable resolution scheme. To emphasize the general advantage of the variable sampling scheme, we focus on models' evaluation at a single fixation, without any scanning, where the variable resolution peaks at an arbitrarily chosen fixed image location, not aligned with the objects' layout in the image.

The main contributions of this paper are as follows:

1. Large vision-language foundation models (VLMs) provide state-of-the-art performance on the task of VQA. We apply three VLMs on three VQA datasets, comparing between the variable and uniform sampling schemes. In particular, we evaluate the models ViLT (13), MDETR (12) and BLIP2 (15) on VQAv2 (6), GQA (10) and SEED-Bench (14) datasets, respectively. We show that the models improve significantly for most question types with

Fig. 2: Alternative sampling schemes. (a) Variable resolution with peak sample density at the center of fixation (image center in this paper) and linearly decreasing number of samples with eccentricity. **(b) Uniform resolution** with a constant density of samples. Both schemes distribute an equal number of samples over the entire FOV (using log-polar coordinates). In this work we address the question: which of the two alternative (motorbike) representations improves on complex visual tasks with existing DNN architectures, given that both alternatives consist of an equal number of samples?

the variable resolution images compared to the naive uniform resolution alternative. Furthermore, with a mere 3% pixel budget, models reach about 80% accuracy in comparison with the full resolution baseline.
2. We conduct extensive control on the advantage exhibited by the variable model and show that it holds generally, irrespective of fixation point location.
3. In VQA many questions are around objects, their attributes and relations, hence requiring the capacities of sub-tasks such as object detection. We train models for this task with each of the sampling schemes and reveal a similar performance improvement.

2 Experimental setup

We consider three main sampling configurations of the input images given, and train a different model for each of the input sampling configurations. Importantly, we note that for the sub-sampling techniques (variable and uniform) we utilize a simple bilinear interpolation to form the final image while preserving spatial alignment. This allows us to extensively test existing vision systems made to work with Certasian coordinates without performing architectural changes. We do not attempt computational gains and instead isolate the effects of subsampling strategies on complex visual tasks.

Baseline. We refer to the given original images, without any pre-processing, as "baseline" or "full resolution", utilizing 100% of the available pixel budget. The FOV spanned by the original image pixels is referred to as the "full FOV".

Variable sampling. The sampling approach follows (24) and (31), which modeled the human representation of visual information in the retina and the

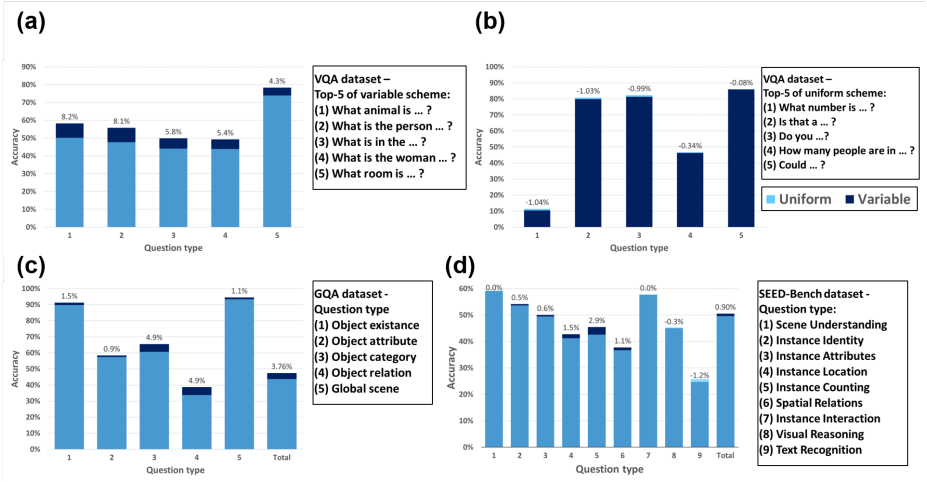

Fig. 3: Evaluation of visual question answering. (a,b) Accuracy on the VQAv2 dataset (ViLT): (a) The top-5 questions where the variable scheme outperforms the uniform, (b) where the uniform outperforms the variable. The Variable advantage is high and statistically significant, while the uniform advantage is not. (c) Accuracy on the GQA dataset (MDETR). (d) Accuracy on the SEED-Bench dataset (BLIP2). Numbers above the bars indicate the marginal accuracy difference.

visual cortex. The *variable resolution* scheme consists of sampling with a receptive field size, which continually increases with eccentricity (Fig. 2a). We apply a Log-Polar transformation to each image, where sample density remains constant $\forall\theta \in [0, 2\pi]$ and decreases linearly with r

$$(r, \theta) = (log(\sqrt{(x - x_f)^2 + (y - y_f)^2}), arctan(\frac{y - y_f}{x - x_f}))$$

where x_f, y_f are the coordinates of the fixation point. The sampling yields a budget of about 10K samples (pixels) within the full FOV. In the context of the COCO dataset (16), with typical image size of 480×640, this amounts to a pixel budget of about 3%. Note that throughout this paper, we arbitrarily picked the center of the image as the location of the highest sample density, regardless of the task or object location in the scene. We introduce multiple mechanisms to control for the central image bias (see Section 4).

Uniform sampling. In this sampling approach the budget of samples is uniformly distributed across the entire FOV. We employ a concentric grid to comply with the variable sampling. This *uniform resolution* schema is akin to simply down-scaling the FOV. See Figure 2b.

Training paradigm. We trained all object detection and backbone models (Mask-RCNN, DETR, ResNet101 (3; 8; 9)) from scratch using their original hyperparameters and training methods. Each model was trained on one of three versions of its dataset, matching the sampling configuration used in testing. As such, we have three models for every experiment. Due to computational

resource limitations we fine-tuned only the MDETR model among the VLMs, which yielded consistent findings with the non-fine-tuned version, mitigating the need to re-train the other VLMs.

3 Visual question answering (VQA)

Visual question answering is a complex yet one of the most fundamental visual tasks for an intelligent agent. This task requires the perception of subtle cues related to object relations, interactions and causality in a scene. The VQA task combines multiple mechanisms to answer questions about specific elements of the image, the general layout of the scene or both. In our experiments, we first applied a pretrained ViLT model, on the VQAv2 dataset consisting of 65 question types. We evaluated the model on three sampling schemes: baseline-full resolution, variable and uniform. The results clearly indicate a significant advantage for the variable sampling scheme, with an overall accuracy gain of 1.96% on the validation set (1.44% on the test set) compared to the uniform scheme ($M_{var} = 64.9 \pm 19.8\%$; $M_{uni} = 62.9 \pm 19.9\%$; $t[64] = 9.16$; $p < 1 \times 10^{-6}$, on the validation set). The mean accuracy for the baseline with full resolution images on this dataset is 81.1%.

Next, we applied the MDETR model on the GQA dataset, which demonstrated similar trends (Table 1). The model evaluated on the variable sampling scheme achieved a total accuracy of 47.4% compared to 43.7% for the uniform scheme ($t[4] = 2.32$; $p = 0.04$). The GQA 'testdev' set consists of 12,578 questions and 398 images. Both sampling schemes improved with fine-tuning (Table 1). Lastly, we tested the BLIP2 model on the new SEED-Bench dataset,

Table 1: MDETR model evaluation on GQA testdev dataset.

Model	Question type	Accuracy Baseline (full)	Accuracy Variable	Accuracy Uniform
Pretrained	(1) Object existence	95.6%	**91.3%**	89.8%
	(2) Object attribute	71.2%	**58.3%**	57.4%
	(3) Object category	76.0%	**65.5%**	60.6%
	(4) Object relation	53.1%	**38.6%**	33.7%
	(5) Global scene	95.8%	**94.5%**	93.3%
	Total	61.7%	**47.4%**	43.7%
Fine-tuned	(1) Object existence	95.6%	**93.8%**	93.4%
	(2) Object attribute	71.2%	**62.6%**	62.6%
	(3) Object category	76.0%	**71.1%**	70.3%
	(4) Object relation	53.1%	**46.5%**	44.9%
	(5) Global scene	95.8%	95.2%	**95.7%**
	Total	61.7%	**54.3%**	53.3%

where the variable scheme outperforms the uniform by 0.9% (50.5% vs. 49.6% total accuracy; $t[14232] = -6.00$; $p < 1 \times 10^{-6}$; see Supplementary).

As shown in Figure 3, most question types benefit from the variable sampling scheme, which provides finer details in the center of the image, over the uniform scheme. The advantage is even more remarkable considering it is achieved with a single arbitrary fixation, while the cues for answering the questions can be located anywhere in the scenes.

Interestingly, we note that the Top-5 marginal difference question types on VQAv2 have a very different distribution for both the uniform and the variable schemes. The Variable resolution scheme yields marginal gains of up to 8.2% (see Fig. 3a). On the contrary, the uniform sampling scheme prevails with *at most* 1.0% and in most cases orders of magnitude less (see Fig. 3b). The same pattern is evident in both the GQA and the SEED-Bench datasets (Fig. 3c,d).

Variable resolution induces bidirectional information flow between center and periphery. Inspection of the models' self-attention maps reveals insight into the improved performance: while the uniform scheme spreads attention only locally, the variable scheme integrates contextual information by attention spreading to the periphery, similar to human perception (Supp. Section 3).

4 On central image bias

The datasets under review, GQA, VQAv2, and SEED-Bench, may occasionally concentrate on objects positioned at the center of the image, directly aligning with our high-resolution fixation. Consequently, it is not surprising that the variable-resolution model gains an edge in performance due to high-resolution input focused on critical image areas. At this juncture, the reader might be tempted to accredit the benefits of variable resolution to merely coincide with the pre-existing photographer bias in the information-dense region of these datasets.

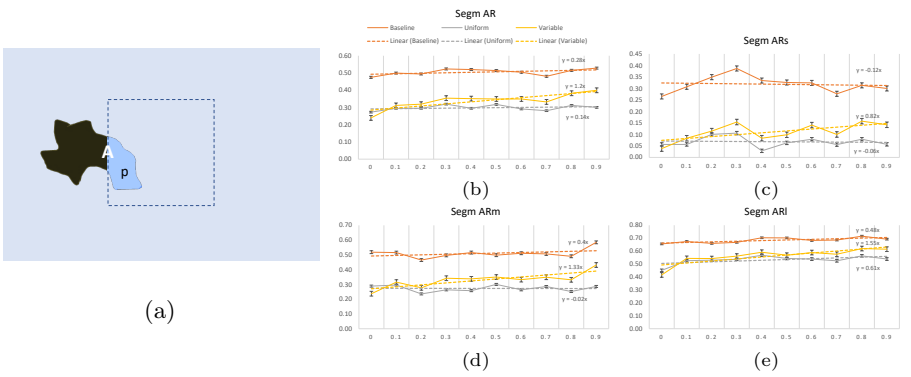

Fig. 4: Bin experiment. (**a**) The $\frac{P}{A}$ HRR allows for measuring the degree to which a ground-truth object is contained in the high-resolution middle area. (**b-e**) Evaluation variable vs uniform sampling performance w.r.t. degree of inclusion (HRA = 200×200).

As a gateway to explore the results achieved on the complex VQA task, and to investigate such concerns in depth, we evaluated the behavior of several models on the underlying task of object detection (16), which is essential for VQA. For this task, we utilized the DETR (also used for VQA with modulation, as MDETR) and Mask-RCNN architectures. The semantic richness of COCO, such as object location, size, type etc. allows us to precisely measure the degree to which a photographer bias artificially benefits the variable model. The following are two experiments **(1), (2)** that demonstrate a variable model outperforms a uniform one irrespective of fixation point location.

Creating annotation bins (1). Consider the COCO validation set $V = I_1, I_2, \ldots, I_{5,000}$, where $I_i \in \mathbb{Z}^{W_i \times H_i \times 3}$. Define a square of size $D \times D$ ($D < W_i$, $D < H_i$ for all i). For instance, D could be 200, since even the smallest images in COCO are larger. Center this square on each validation image, calling the area inside it the high-resolution area (*HRA*).

For each ground truth annotation, compute $\frac{P}{A}$, the fraction of its pixel mask area inside the *HRA* (Figure 4a). This metric, the high-resolution inclusion degree (or inclusion degree), measures how much of the annotation is within the *HRA*. The inclusion degree varies with D. For example, an object might have an inclusion degree of 0.5 for a 200×200 *HRA*, but 0.3 for a 150×150 *HRA*. We tested *HRA* sizes from 100×100 to 250×250 in 10-pixel increments, with consistent results across sizes.

Annotations are then binned by inclusion degree. The set $G_{0.0-0.1}$ contains annotations with inclusion degrees between 0.0 and 0.1, $G_{0.1-0.2}$ includes those between 0.1 and 0.2, and so on, forming ten bins: $G_{0.0-0.1}$, $G_{0.1-0.2}$, ..., $G_{0.9-1.0}$. Their union covers the entire validation set. Figure 4(b-e) shows model performance across these bins, revealing that variable sampling significantly benefits when objects fall within the high-information density area, while uniform sampling offers limited advantages even in peripheral bins.

Counting samples (2). As an additional control, we constructed an annotation set containing only objects encompassed by an identical sample count, varying only in its distribution pattern: variable or uniform (see Fig. 2). The results show a similar performance gap of $\sim$2.0% in favor of the variable model (Table S3 in the Supplementary).

5 Related Work

Several prior computational work explored aspects of foveal schemes, where samples are distributed densely around a fixation point and more sparsely in the periphery of the FOV (1; 17; 23). Early studies developed models to evaluate the capabilities and limitations of human peripheral vision (5). Others (4), studied the impact of foveated texture-based input representations in artificial vision systems on the task of scene classification. They showed that peripheral texture encoding leads to representations with greater generalization, sensitivity to high-spatial frequency and robustness to occlusion. Another study (30) explored a neurocomputational modeling of central and peripheral vision for scene recog-

nition. The study suggested that the advantage of peripheral over central vision is due to intrinsic usefulness of features carried by peripheral vision, generating a greater spreading transform in the internal representational space. They predicted that the two pathways correlate with their neural substrates, LOC and PPA in the brain. However, scene classification may provide only limited insight, as it can be often performed well at extremely poor resolutions (28). A recent study (25) suggested that blurry peripheral vision may have evolved to optimize object recognition. Applying DNNs to foveated images around objects of interest, the study showed that the performance is peaked at the human blur decay setting, also benefiting from reduced false detections in the blurry periphery. Cortical magnification is a brain mechanism that allocates more processing units to the densely sampled area of the foveal image. This approach was applied to videos to fit models into embedded systems (11). This study achieved a $4\times$ speed-up in frame rate, but showed only a small decrease in recall within the restricted foveal region.

6 Conclusions

We address the problem of current artificial vision systems in covering a large FOV, while enabling the acquisition of fine details in high resolution to perform complex visual tasks. Inspired by the human visual system, we employ and evaluate a variable resolution sampling scheme, under a limited budget of samples, with a high resolution area at the center of the FOV and linearly decreasing resolutions in the periphery.

When applied to the complex VQA task, the variable sampling scheme consistently outperforms the uniform sampling scheme across most question types, as demonstrated in Section 3 for the ViLT, MDETR and BLIP2 VLMs on VQAv2, GQA and SEED-Bench datasets, respectively. This is an outstanding finding, mainly from two perspectives. First by considering the fact that we arbitrarily choose the highest resolution area location in the center of the image, while the cues required to answer the questions can be anywhere in the scene. Second, the improvement is achieved with a single fixation, without any scanning across the FOV. In early studies, (28) showed that humans and machines can perform well on the task of scene classification only with a uniformly sub-sampled *gist* of an image. Our results indicate that a variable resolution scheme, is a better alternative than the uniform sampling scheme (e.g., the question "what room is?" in VQAv2 yields an improvement of 4.3%, the "object relation" question type in GQA and the "instance counting" question type in SEED-Bench, both gains 2.9%; see Fig. 3). This gain in performance at a single arbitrary fixation, suggests the dissemination of high resolution information from the center of the FOV to the periphery and of low resolution contextual information from the periphery to the center.

Overall, the results show the potential of the biologically-inspired image representation to improve the design of visual acquisition and processing models in future AI-based systems.

Acknowledgements

We are grateful to Botond Szabó and Carlo Baldassi for their insightful advice and helpful discussion. This work was supported by MBZUAI-WIS Program for Collaborative Research in AI and by a research grant from the Carolito Stiftung. D.H. is supported by the Robin Neustein AI research fellowship.

Bibliography

[1] Akbas, E., Eckstein, M.P.: Object detection through search with a foveated visual system. PLoS Computational Biology **13**, 1–28 (2017)

[2] Bar, M.: A cortical mechanism for triggering top-down facilitation in visual object recognition. Journal of cognitive neuroscience pp. 600–609 (2003)

[3] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229 (2020)

[4] Deza, A., Konkle, T.: Emergent properties of foveated perceptual systems (2021)

[5] Freeman, J., Simoncelli, E.P.: Metamers of the ventral stream. Nature neuroscience **14**(9), 1195–1201 (2011)

[6] Goyal, Y., Khot, T., Agrawal, A., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. International Journal of Computer Vision **127**(4), 398–414 (2019)

[7] Han, Y., Roig, G., Geiger, G., Poggio, T.: Is the Human Visual System Invariant to Translation and Scale ? pp. 564–568 (2017)

[8] He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask R-CNN. Proceedings of the IEEE International Conference on Computer Vision pp. 2980–2988 (2017)

[9] He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 770–778 (2016)

[10] Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)

[11] Jaramillo-Avila, U., Anderson, S.R.: Foveated Image Processing for Faster Object Detection and Recognition in Embedded Systems Using Deep Convolutional Neural Networks. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) pp. 193–204 (2019)

[12] Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr - modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1780–1790 (October 2021)

[13] Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning. pp. 5583–5594 (2021)

[14] Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2024)

[15] Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023)
[16] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
[17] Lukanov, H., König, P., Pipa, G.: Biologically Inspired Deep Learning Model for Efficient Foveal-Peripheral Vision. Frontiers in Computational Neuroscience **15** (2021)
[18] Marr, D., Poggio, T., Hildreth, E.: Smallest channel in early human vision. Journal of the Optical Society of America **70**(7), 868–870 (1980)
[19] Martin, G.R.: Visual fields and their functions in birds. Journal of Ornithology **148**(Suppl 2), 547–562 (2007)
[20] Nguyen, P., Quach, K.G., Duong, C.N., Le, N., Nguyen, X.B., Luu, K.: Multi-camera multiple 3d object tracking on the move for autonomous vehicles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2569–2578 (2022)
[21] Oliva, A., Torralba, A.: The role of context in object recognition. Trends in Cognitive Sciences **11**(12), 520–527 (2007)
[22] Pang, Y., Cao, J., Li, Y., Xie, J., Sun, H., Gong, J.: Tju-dhd: A diverse high-resolution dataset for object detection. IEEE Transactions on Image Processing **30**, 207–219 (2021)
[23] Paula, B., Moreno, P.: Learning to Search for and Detect Objects in Foveal Images Using Deep Learning. In: Conference on Pattern Recognition and Image Analysis. pp. 223–237. Springer Nature Switzerland (2023)
[24] Poggio, T., Mutch, J., Isik, L.: Computational role of eccentricity dependent cortical magnification. Tech. Rep. CBMM Memo 017 (2014)
[25] Pramod, R.T., Katti, H., Arun, S.P.: Human peripheral blur is optimal for object recognition. Vision Research **200** (2022)
[26] Read, J.C.: Binocular Vision and Stereopsis across the Animal Kingdom. Annual Review of Vision Science **7**, 389–415 (2021)
[27] Said, Y., Barr, M.: Human emotion recognition based on facial expressions via deep learning on high-resolution images. Multimedia Tools and Applications **80**(16), 25241–25253 (2021)
[28] Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: A large data set for nonparametric object and scene recognition. IEEE Trans Pattern Anal Mach Intell **30**, 1958–70 (2008)
[29] Trevarthen, C.B.: Two mechanisms of vision in primates. Psychologische Forschung **31**, 299–337 (1968)
[30] Wang, P., Cottrell, G.W.: Central and peripheral vision for scene recognition: A neurocomputational modeling exploration. Journal of Vision **17**(4), 9–9 (06 2017)
[31] Wilson, H.R., Bergen, J.R.: A four mechanism model for spatial vision. Vision Research **19**(1), 19–32 (1979)