Interpretable Adverse Lens Corruptions

Kai Bäuerle
University of Mannheim, Germany
kaibauerle@gmail.com

Patrick Müller
University of Siegen, Germany
patrick.mueller@uni-siegen.de

Ivo Ihrke
University of Siegen, Siegen
ivo.ihrke@uni-siegen.de

Margret Keuper
University of Mannheim, Germany
keuper@uni-mannheim.de

Abstract

Deep neural networks excel at image classification on benchmarks like ImageNet, yet they remain vulnerable to adverse conditions, including environmental changes and sensor noise, such as lens blur or camera noise. Consequently, the study of these adverse noise corruptions has been extensive. At the same time, each camera lens is expected to have its characteristic blur: while manufacturers optimize the lens to have fewer aberrations, a compromise between lens quality and available budget as well as physical constraints has to be made. However, a study of adverse but realistic blur corruptions is currently still amiss. We introduce Adverse Lens Corruption (ALC), an optical adversarial attack that identifies worst-case lens blurs. ALC works by optimizing Zernike polynomial-based aberrations via gradient descent. This method enables direct optical analysis and complements existing noise and corruption benchmarks, revealing which lens designs pose the greatest challenge to current models.

1 Introduction

Lens design is a multi-step process that relies heavily on optimization of an initial lens configuration, which follows the system specifications such as focal length and field of view. The merit function is then used to optimize the optical parameters, *i.e*. the individual lens surfaces and distances [1]. This process is usually done without incorporating in-depth knowledge of the final computer vision task.

At the same time, adversarial attacks are well-known to find model-specific worst case inputs to the models that cause surprisingly low performance. One drawback of such methods is that standard attacks like PGD [2] are often not interpretable from a physics point of view and the likelihood of encountering an adversarial example naturally is usually low. However, in this work, we introduce Adverse Lens Corruptions (ALC) designed to identify weak points in an optical aberration space that is directly optically interpretable and linked to physics. We consider our method as a first step towards novel robust task-specific lens designs: the ultimate goal is to constrain the lens design parameter space to avoid critical aberration regions, where the computer vision model is most sensitive to. Ideally, this helps also to specify advanced production tolerances, where the tolerance budget can be relaxed or tightened based on an analysis of adverse lens corruptions.

Our contributions are as follows: Complementing existing attacks, we introduce Adverse Lens Corruption (ALC) designed to evaluate the robustness of image classification models against optical blur corruptions, *i.e*. we optimise the optical kernel h that acts on the input signal x via *convolution*. The proposed ALC learns a linear combination of realistic optical aberration effects, such as coma, astigmatism and spherical aberration to represent lens blur. For this, we expand the wave aberration into Zernike polynomials and optimize the coefficients with different constraints, which results in

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Learning to Sense.

dataset and model-specific worst-case lens blur. Due to its optics-based parameterization, ALC comes with *interpretable* optical results for free.

In this first work, we show that ALC yields model and dataset specific lens corruptions that effectively reduce the model accuracy. Compared to the worst-case OpticsBench [3] corruption per model, ALC is stronger at a comparable corruption size. Compared to PGD for single image attacks, ALC provides complementary information, *e.g*. the main class-wise confounders are different.

Model Robustness Model robustness and stability have been addressed from various perspectives, including data augmentation and adversarial training. Our method introduces an optical adversarial attack that can both benchmark state-of-the-art image classifiers and improve them through adversarial training. Unlike previous attacks that focus on point-wise noise, ALC models adverse lens *blur*.

Data Augmentation and Knowledge Distillation Data augmentation improves image classification robustness by simulating real-world variations. AugMix [4] strengthens resistance to common corruptions, while [3] employ optical blur kernels to mitigate aberrations. Other strategies include learned augmentation policies [5], feature perturbations [6, 7], frequency adjustments [8], and style adaptation via generative models [9, 10, 11, 12]. Knowledge Distillation (KD) transfers robustness from a teacher model, enhancing adversarial [13, 14, 15, 16] and out-of-distribution performance [17]. However, KD relies on large pre-trained models and entails substantial computational cost.

Optical Co-design Recent work shows that optics and image-processing can be optimized end-to-end to produce task-specific gains [18, 19]. Although end-to-end co-optimization including the downstream application has been demonstrated [18], it is not yet widely adopted in industry due to high computational cost and practical challenges for manufacturability and certification [18, 20].

Adversarial Attacks Neural networks can be deceived by applying imperceptible perturbations to specific image regions [21]. Attack strategies are usually split into *white-box* and *black-box* methods [2, 21, 22, 23, 24]. Recent work combines both modalities to increase robustness [25]. Physical-world attacks exploit optical effects rather than digital pixel changes. Superpixel-guided attention targets salient regions [26], while structured illumination and Fourier-domain modulation introduce adversarial light patterns that remain invisible to humans [27, 28]. Rolling-shutter manipulation yields adversarial shadows or reflections [29, 30], and camera stickers can mislead object detectors [31]. Similar techniques threaten X-ray security systems via metal modifications [32]. Unlike these works, our Adverse Lens Corruption (ALC) framework probes model vulnerability to realistic optical degradations that are non-malicious and useful for design feedback. ALC perturbs the entire image by optimizing within a physically meaningful aberration space rather than per-pixel noise or local patches. The attack is unconstrained in ℓ_2 norm but bounded by the feasible range of optical parameters, mirroring realistic lens distortions. Adversarial training can mitigate such attacks and even shift texture bias toward shape bias [33]. Our experiments confirm that ALC-based training preserves this effect (see Appendix I).

2 Optimizing Adverse Lens Corruptions

Photographic lenses are aberration-limited, while lens design aims to reduce the amount of aberrations by optimizing the distances and radii of lens elements [34]. However, because of physical constraints on materials, physical dimensions and economical pressure, optical aberrations are to be expected and subject to a trade-off between costs and lens quality.

We study which aberrations constitute worst-case scenarios for image classification and segmentation. To this end, we model the point spread function (PSF), which describes lens blur as the spatial impulse response of an optical system [35]. In general, this PSF varies with wavelength, location, and distance [35]. However, we make several practical simplifications to this general setting, as detailed in the Appendix B. The PSF depends then only on wavelength.

Our Adverse Lens Corruption generates a PSF kernel h to corrupt a clean image x:

$$y = h * x. \tag{1}$$

Unlike classical adversarial attacks [21], which optimize additive noise n, we optimize an adversarial kernel h, interpretable as PSF, based on a wave optics model [35]. We parameterize the wave aberration model in the exit pupil using Zernike polynomials [36], which describe optical imperfections such as defocus, astigmatism or coma. The coefficients A_n^m of this polynomial expansion are optimized

per color channel to construct worst-case aberration combinations for computer vision tasks. Using the resulting kernel h, the corrupted image at iteration t is defined as

$$adv_x^t = adv_x^{t-1} + \nu \cdot (x * h_\lambda^{t-1} - adv_x^{t-1}), \quad \nu \in [0, 1], \tag{2}$$

with $adv_x^0 = x$. Typically, $\nu = 1$, yielding purely convolutional corruption, but $\nu < 1$ allows interpolation with additive perturbations. To optimize aberrations adversarially, we iteratively update the Zernike coefficients via gradient ascent on the model loss L:

$$A_n^{m,t} = A_n^{m,t-1} + \alpha \nabla_{A_n^{m,t-1}} L(\theta, adv_x^{t-1}, label_x), \tag{3}$$

where the coefficients $A_n^{m,t}$ determine the optical kernel h_λ^t for all three color channels. We also provide a detailed description of the kernel generation h in the Appendix in Section B.

3 Experimental Evaluation

In the following, we evaluate different computer vision models against our proposed, optical adversarial attack. The DNNs we evaluate are trained on different publicly available subsets of ImageNet [37] to allow for extensive experiments. To complement image classification, we also conduct experiments on image segmentation on the COCO segmentation dataset [38] using SAM [39] in Section F in the Appendix. All experiments are conducted with restriction parameters τ and τ_e values of 4 to provide compatibility to OpticsBench [3]. We also report an ablation of how τ affects the accuracy in the Appendix in Section C.

3.1 Model comparison

To evaluate the ALC attack, we train models on multiple datasets and optimize a single optical kernel for each model - dataset combination (see Appendix A.1 for details). We evaluate CNN-based models, such as ResNet50 [40], AlexNet [41], DenseNet161 [42], EfficientNet b0 and b4 [43], VGG16 [44], MobileNetV3 large [45], ConvNeXt [46] and ConvNeXt v2 [47], as well as transformer-based models like Vision Transformer [48] and Swin Transformer v2 [49]. Furthermore, state-of-the-art foundation models are attacked by the ALC attack, such as CLIP [50], DINO v2 [51] and SAM [39]. Note that for this experiment, we optimize a single adverse kernel per model and dataset, enabling efficient robustness testing at inference.

	Clean ↑	**OpticsBench ↑**	**ALC ↑**	Δ	ℓ_2
ResNet50	0.809	0.130	0.093	0.716	44.618
VGG16	0.716	0.044	0.052	0.664	40.623
DenseNet161	0.772	0.141	0.074	0.698	46.822
EfficientNet b0	0.777	0.116	0.049	0.728	45.738
EfficientNet b4	0.834	0.127	0.149	0.685	47.264
MobileNet v3	0.753	0.072	0.029	0.724	43.500
ViT (base)	**0.853**	**0.246**	**0.177**	0.676	43.478
SwinV2-tiny	0.821	0.132	0.085	**0.736**	46.002
SwinV2-small*	0.837	0.183	0.119	0.718	45.014
CLIP	0.761	0.104	0.078	0.683	48.274
Σ	-	-	-	-	45.13

Table 1: Accuracy for different classification models, including clean accuracy, accuracy for our ALC attack and the difference (Δ), evaluated on ImageNet1k. *=Multiple Seeds. For comparison, we also report the accuracy for the hardest corruption on OpticsBench [3] per model for severity 5, where the ℓ_2-distance is 46.0 for OpticsBench. The average ℓ_2-distance across OpticsBench corruptions is 44.32 is comparable to ALC with an average 45.13 for $\tau = 4.0$. The single optimized ALC corruption per dataset is stronger than the OpticsBench corruption at a comparable corruption level.

Table 1 reports ImageNet1k accuracies under clean and ALC attacked conditions. While all evaluated models perform poorly, ViT (base) remains most robust (0.177), followed by EfficientNet-b4 (0.149). Both of these models are also able to classify best on the clean dataset. Yet, all models see a substantial decrease in accuracy upon being attacked by the optical attack, with the SwinV2-tiny suffers the largest drop (0.736) and MobileNet v3 sees the lowest absolute accuracy while being attacked.

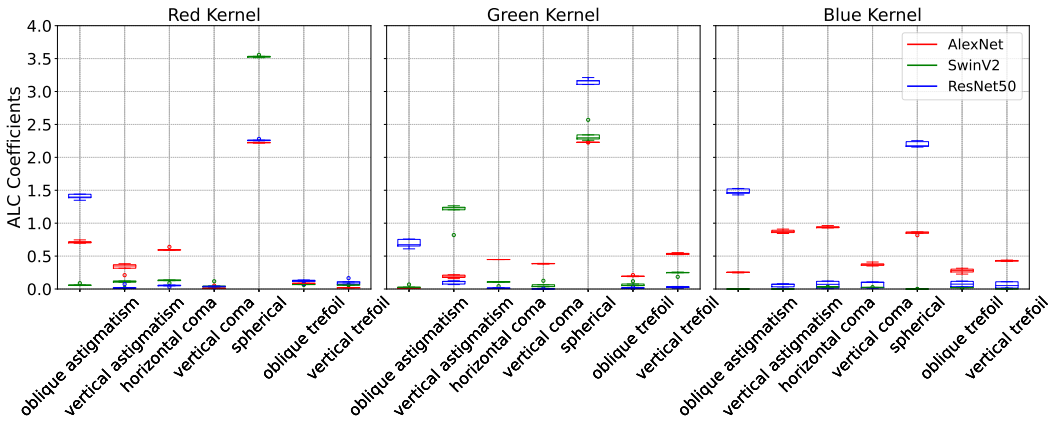


Figure 1: Comparison of the absolute coefficient values A_n^m without defocus on three different ImageNet subsets (ImageNette, ImageNet100 and ImageNet1k) with ResNet50. The models are trained with the same seed, however ALC was evaluated on 5 different seeds per dataset.

3.2 Corruption Ablation Study

In the previous subsection (Section 3.1), we showed, that Adverse Lens Corruptions decrease the accuracy signficantly via introducing defocus into the adversarial kernel. In this subsection, we elaborate, which alternative corruptions are getting increasingly important, while defocus, which has been shown in Table 2 to be the dominating adverse aberration, is not available. We justify this choice by noting that defocus can be corrected by changing the position of the image sensor.

Therefore, we dismiss the defocus corruption as an additional restriction and run the same adversarial attack as in Section 3.1. The result of this adjusted optical attack can be examined in Table 2 with three different models (Resnet50, EfficientNet & CLIP) on ImageNet1k.

Table 2: Ablation of ALC parameter space on ResNet50 [40], EfficientNet b4 [43] and CLIP [50].

Model	Clean	full ALC	No Defocus	No Spherical
ResNet50	0.809	0.102	0.267	0.564
EfficientNet b4	0.834	0.149	0.374	0.619
CLIP	0.761	0.078	0.148	0.410

After defocus, the most prominent aberration is spherical aberration, which is caused by spherical lenses. Spherical lenses are used in most systems, because of the simple manufacturing process. Spherical aberration can also be corrected using aspherical lenses, which are *e.g.* common in mobile plastic lenses, but are often costly to manufacture. However, the last column shows that correcting for spherical aberration would lead to significantly higher accuracies.

Figure 1 shows that the learned kernels are equally dataset characteristic when defocus is removed as an aberration. See also Figures 20 and 21 in the Appendix.

Discussion While there is a large research body that discusses standard adversarial attacks that generate artificial noise [2, 21, 22, 52], our Adverse Lens Corruptions follows a different goal. ALC relates to realistic lens blur by constraining the disturbance to operate in an aberration space spanned by Zernike polynomials. Therefore, ALC is not ℓ_2-bounded (see Appendix, Section J).

Our approach helps to gain actionable insights of which aberration combinations to avoid during lens design. Specifically, ALC could serve as means to guide the optimization process in lens design. Despite these advantages, there are a few practical limitations, which are discussed below and in the Appendix Section K in more detail. The method is mainly useful for aberration-limited lenses such as photographic and machine vision lenses. Modeling other optical effects such as space-variance, vignetting, wavelength-dependent coupling and the full camera pipeline including sensor noise and the image signal processor is left for future work [1].

Several promising extensions are discussed below. Future work includes sampling polytopes of aberration configurations using our method, exploring an epsilon-bounded distortion budget around real lens representations in Zernike space, and applying ALC for tolerancing by emphasizing model-specific worst-case points. Designers could also use our approach to identify aberration configurations within acceptable accuracy budgets, helping balance lens quality and manufacturing cost.

4 Conclusion

Adverse Lens Corruptions addresses the robustness of models to lens blur: ALC finds the worst-case lens blur. In contrast, existing benchmarks like OpticsBench require testing a huge variety of blur kernels and challenges models less severely. With optimizing the linear combination of Zernike Polynomials, ALC can identify the aberrations that contribute most, providing an optical interpretation for lens designers.

Acknowledgments

This work was supported by the DFG research unit DFG-FOR 5336 "Learning to Sense" with additional support through the ZESS-Scholarship from the NRW-Zentrum für Sensorsysteme, Siegen.

References

[1] Gross, Herbert, editor. *Handbook of Optical Systems*, volume 3. John Wiley & Sons, Ltd, 1 edition, 2006.

[2] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[3] Patrick Müller, Alexander Braun, and Margret Keuper. Classification robustness to common optical aberrations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2023.

[4] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[5] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

[6] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[7] Benjamin Erichson, Soon Hoe Lim, Winnie Xu, Francisco Utrera, Ziang Cao, and Michael Mahoney. NoisyMix: Boosting model robustness to common corruptions. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*. PMLR, 2024. ISSN: 2640-3498.

[8] Mehmet Kerim Yucel, Ramazan Gokberk Cinbis, and Pinar Duygulu. Hybridaugment++: Unified frequency spectra perturbations for model robustness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023.

[9] Minui Hong, Jinwoo Choi, and Gunhee Kim. StyleMix: Separating content and style for enhanced data augmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[10] Han Xue, Zhiwu Huang, Qianru Sun, Li Song, and Wenjun Zhang. Freestyle layout-to-image synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[11] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 2020.

[13] Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04), April 2020.

[14] Bojia Zi, Shihao Zhao, Xingjun Ma, and Yu-Gang Jiang. Revisiting adversarial robustness distillation: Robust soft labels make student better, 2021.

[15] Bo Huang, Mingyang Chen, Yi Wang, Junda Lu, Minhao Cheng, and Wei Wang. Boosting accuracy and robustness of student models via adaptive adversarial distillation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023.

[16] Shiji Zhao, Jie Yu, Zhenlong Sun, Bo Zhang, and Xingxing Wei. Enhanced accuracy and robustness via multi-teacher adversarial distillation. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*. Springer Nature Switzerland, 2022.

[17] Andy Zhou, Jindong Wang, Yu-Xiong Wang, and Haohan Wang. Distilling out-of-distribution robustness from vision-language foundation models. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[18] Ethan Tseng, Ali Mosleh, Fahim Mannan, Karl St-Arnaud, Avinash Sharma, Yifan Peng, Alexander Braun, Derek Nowrouzezahrai, Jean-François Lalonde, and Felix Heide. Differentiable compound optics and processing pipeline optimization for end-to-end camera design. *ACM Trans. Graph.*, 40(2), June 2021.

[19] Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Transactions on Graphics (TOG)*, 37(4):1–13, 2018.

[20] Alice Fontbonne, Hervé Sauer, and François Goudail. Comparison of methods for end-to-end co-optimization of optical systems and image processing with commercial lens design software. *Opt. Express*, 30(8):13556–13571, Apr 2022.

[21] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[22] Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017.

[23] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations (ICLR) Workshop Invite*, 2017.

[24] Francesco Croce and Matthias Hein. Minimally distorted Adversarial Examples with a Fast Adaptive Boundary Attack. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*. PMLR, 2020.

[25] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*. PMLR, 2020.

[26] Xiaoyi Dong, Jiangfan Han, Dongdong Chen, Jiayang Liu, Huanyu Bian, Zehua Ma, Hongsheng Li, Xiaogang Wang, Weiming Zhang, and Nenghai Yu. Robust superpixel-guided attentional adversarial attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[27] Abhiram Gnanasambandam, Alex M. Sherman, and Stanley H. Chan. Optical Adversarial Attack. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021.

[28] Athena Sayles, Ashish Hooda, Mohit Gupta, Rahul Chatterjee, and Earlence Fernandes. Invisible Perturbations: Physical Adversarial Examples Exploiting the Rolling Shutter Effect. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021.

[29] Yiqi Zhong, Xianming Liu, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Shadows can be Dangerous: Stealthy and Effective Physical-world Adversarial Attack by Natural Phenomenon. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022.

[30] Donghua Wang, Wen Yao, Tingsong Jiang, Chao Li, and Xiaoqian Chen. RFLA: A Stealthy Reflected Light Adversarial Attack in the Physical World. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023.

[31] Juncheng Li, Frank Schmidt, and Zico Kolter. Adversarial camera stickers: A physical camera-based attack on deep learning systems. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*. PMLR, 2019.

[32] Aishan Liu, Jun Guo, Jiakai Wang, Siyuan Liang, Renshuai Tao, Wenbo Zhou, Cong Liu, Xianglong Liu, and Dacheng Tao. X-Adv: Physical Adversarial Object Attacks against X-ray Prohibited Item Detection, 2023.

[33] Paul Gavrikov, Janis Keuper, and Margret Keuper. An Extended Study of Human-Like Behavior Under Adversarial Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[34] Rudolf Kingslake and R. Barry Johnson. *Lens Design Fundamentals*. Elsevier/Academic Press, 2nd ed edition, 2010.

[35] Joseph W Goodman. Introduction to fourier optics, 4w. h. *Freeman, New York*, 2017.

[36] M. Born, E. Wolf, and A.B. Bhatia. *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*. Cambridge University Press, 1999.

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3), 2015.

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

[39] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[40] 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 (CVPR)*, 2016.

[41] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

[42] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[43] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[45] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3, 2019.

[46] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[47] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders, 2023.

[48] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Represenations (ICLR)*, 2021.

[49] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[51] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research (TMLR)*, 2023.

[52] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. RobustBench: A standardized adversarial robustness benchmark, 2021.

[53] Jeremy Howard. Imagenette. `https://github.com/fastai/imagenette/`, 2023.

[54] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision – European Conference on Computer Vision (ECCV)*. Springer-Verlag, 2020.

[55] Norberto López-Gil, Frances J. Rucker, Lawrence R. Stark, Mustanser Badar, Theodore Borgovan, Sean Burke, and Philip B. Kruger. Effect of third-order aberrations on dynamic accommodation. 47(6):755–765, 2007.

[56] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proceedings of the International Conference on Learning Represenations (ICLR)*, 2019.

[57] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2018.

[58] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[59] Guang ming Dai. Zernike aberration coefficients transformed to and from fourier series coefficients for wavefront representation. *Opt. Lett.*, 31(4), Feb 2006.

This supplementary material provides detailed information on models, datasets and implementation, and additional experimental results. It is structured as follows:

- Dataset and Implementation Details in Section A
- Generation of the ALC Attack in Section B
- Ablation on threshold parameters of ALC and its restrictions in Section C
- Kernel visualizations in Section D
- Implementation details for adversarial training with ALC and the experiments in Section E
- ALC Results of a Segmentation task in Section F
- Image wise ALC Attack in Section G
- Ablation on the optical corruptions in Section H
- Discussing the Shape-Texture bias in the Context of ALC in Section I
- Discussing the non-ℓ_2-boundedness of ALC in Section J
- Discussing additional practical aspects for lens and system designers in Section K

A Datasets and Implementation Details

To evaluate the robustness of multiple image classification models, subsets of the ImageNet [37] dataset are used. The ALC attack, in Section 2, uses the data and implementation process with the details from Subsection A.1 & A.2.

A.1 Implementation Details

The generic ALC attack, which is trained over the whole dataset, generates one data and model specific kernel. This kernel combines the worst case corruptions by iterating through the whole dataset and updating the coefficients A_n^m to determine the optical kernel h_λ^t as described in Equation 2. A visualization of kernels for single coefficients A_n^m without mixing and the effect of mixing with defocus is provided in Fig. 2. To achieve a fair distributed updating process for each image, we iterate through the dataset in whole epochs. In each epoch, the coefficients A_n^m of the ALC attack are only updated with the training subset and subsequently conducted to the validation subset. A_n^m is updated via stochastic gradient descent with an initial learning rate of 0.001, a momentum of 0.9, and a weight decay of 0.0001 The ALC attack has no major adjustments on the coefficients in less than 1M images. Thus, we attack the models for one epoch on the ImageNet21k dataset, one epoch on the ImageNet1k, 10 epochs on the ImagenNet100 dataset, and 100 epochs on the ImageNette dataset, if not stated differently. All experiments can run on CPU or GPU with less than 24 GB of memory.

The image specific ALC attack is not conducted over the whole dataset at once, rather it iterates over each image multiple time to generate an image specific adversarial kernel (h_λ^t). In our experiment, we iterate over each image 40 times and update the coefficients (A_n^m) to determine the optical kernel (h_λ^t) accordingly. After each image, the coefficients are initialized, again randomly, in the range of $[-0.1, 0.1]$.

A.2 Datasets

In Section 3, we evaluate different computer vision models against our proposed, optical adversarial attack. The DNNs we evaluate are trained on different publicly available subsets of ImageNet [37] to allow for extensive experiments and investigate the attack behavior on datasets with different complexity, *i.e*. ImageNette [53] is a dataset consisting of 10 ImageNet classes. It has 9,469 training and 3,925 validation images [53]. ImageNet-100 [54] uses 100 ImageNet classes with a total of 128k training and 5,000 validation images [54]. The ImageNet-1k dataset [37] has 1000 ImageNet classes and contains over 1,281k training images and 50k validation images [37]. Additionally, we also incorporate the full ImageNet-21k dataset [37] to ensure a comprehensive assessment of the effectiveness of our optical attack across varying scales and complexities [37]. To complement image classification, we also conduct experiments on image segmentation on the COCO dataset [38], a comprehensive dataset designed for object detection and segmentation tasks, using SAM [39].

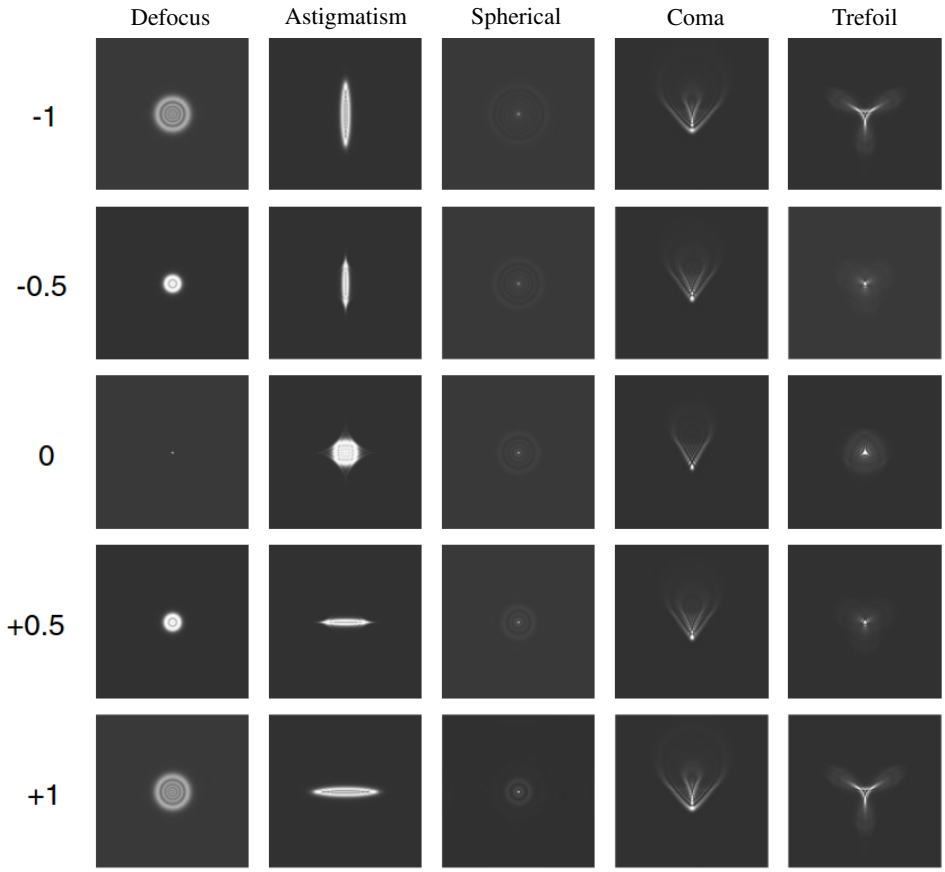


Figure 2: Different aberrations produced by single Zernike Fringe polynomials (central row) and their effect, when mixed with defocus. Figure modified from [55]. All kernels are normalized to the maximum intensity for display purposes. ALC learns a color-dependent linear combination of the corresponding Zernike polynomial to find the worst-case lens blur for a model and dataset.

B Adverse Lens Corruptions Computation

To investigate optical aberrations scenarios, we model the point spread function (PSF), which mimics the characteristic lens blur describing the spatial impulse response of a linear optical system on an image.

More specifically, the PSF describes the effect to a point source that emanates a spherical wave that subsequently passes through a lens. A perfect lens converts this diverging spherical wave into a wave converging near the geometric image point. However, the limiting aperture and imperfections of a real lens cause the point response to spread according to the theory of diffraction and aberrations [35].

In general, this PSF varies with wavelength, location, and distance [35]. However, we make several practical simplifications to this general setting: 1) the local PSF can be assumed to be approximately constant within a small portion of the image. In this work, we therefore assume that the small ImageNet images are regions from a larger image (several megapixels) and that the PSF can be treated as constant. 2) the depth-dependence of the PSF is negligible beyond the hyperfocal distance, we therefore assume imaging at optical infinity to enable a convolutional treatment of the lens aberration effects. With these simplifications, the PSF only depends on the wavelength.

ALC generates a PSF for an aberration-limited system for each color channel, which we then use as a convolution kernel h to generate a corrupted image y from the clean image x via convolution, denoted by '$*$', as described in Eq. (1).

In order to acquire the kernel h tailored to optical aberrations, we use the linear system model for diffraction and aberration as in [35]:

$$h_\lambda(u,v) = \left| \mathcal{F}\left\{ \text{Circ}(\omega_u, \omega_v) \cdot e^{-j\frac{2\pi}{\lambda z} W_\lambda(\omega_u,\omega_v)} \right\} \right|^2 , \tag{4}$$

with j being the imaginary number and W_λ the wave aberration for wavelength λ. For a circular exit pupil, the wavefront is propagated from the exit pupil with coordinates (ω_u, ω_v) to the image space at point (u, v) at distance z via the squared Fourier transform $\mathcal{F}$, to yield an incoherent PSF h for wavelength λ [35].

In Eq. (4), we expand the wave aberration $W_\lambda(\omega_u, \omega_v)$ into the orthogonal Zernike polynomials [36] denoted by Z, which allows for a parameterization of individual aberrations defined in multiples of wavelength λ:

$$W_\lambda(\omega_u, \omega_v) = \lambda \cdot \sum_{n,m} A_n^m(\lambda) \cdot Z_n^m(\omega_u, \omega_v) \tag{5}$$

Each optical basis function Z_n^m represents a meaningful physical representation of specific types of lens aberrations, such as astigmatism and coma. Our optical attack learns the coefficients $A_n^m(\lambda)$ depending on λ, *i.e.* for each color channel separately, to form a worst-case combination of optical aberrations for, *e.g.*, an image classification model.

Using the kernel from Eq. (4), we can corrupt the clean image x with a convolution with the optical kernel h_λ. We can further scale the blur contribution via a corruption size ν, which results in a corrupted image adv_x^t at time step t, as in Eq. (2). In which h_λ^{t-1} denotes the optimized lens corruption at time step $t-1$ of a gradient ascent based optimization and $adv_x^0 = x$, *i.e.* h^0 is the Dirac impulse. In every iteration, adv_x will be clamped to the valid image range. Note that in our default setting, ν will be set to one such that the corruption is purely convolutional as in Eq. (1). In this case, the corruption can be ℓ_2 bounded as in [2].

To use this process in an adversarial technique, we use our model parameters θ and the target $label_x$ to calculate the model loss L (typically cross entropy loss) and update subsequently A_n^m to increase L in an iterative manner via Eq. (3), where the coefficients $A_n^{m,t}$ determine the optical kernel h_λ^t from Eq. (2) for all three colors λ as in Eqs. (4) and (5).

C ALC Attack Restrictions

To restrict the ALC to 8 out of the first twelve Zernike Fringe modes, which can be mixed into one 3D kernel (u, v, λ). Table 3 highlights the selected modes in bold.

#	Name	#	Name
1	piston	7	**horizontal coma**
2	tilt u	8	**vertical coma**
3	tilt v	9	**spherical**
4	**defocus**	10	**oblique trefoil**
5	**oblique astigmatism**	11	**vertical trefoil**
6	**vertical astigmatism**	12	sec. vert. astigmatism

Table 3: First twelve Zernike Fringe modes. We select numbers 4-11 for our optical attack (bold). The exclusion of modes 1-3 are due to optical considerations: the piston term does not cause a kernel intensity change and modes 2 and 3 only cause an image shift without blur. We therefore exclude these modes from our analysis. The remaining selected modes are all primary and two secondary aberrations (mode 10 and 11) – these modes are of direct relevance to optics designers that use them to guide meeting specifications for their designs. A visualization is given in the appendix in Fig. 2.

Next to the limitation of the selected modes, the adversarial attack is restricted by the magnitude of the corruption coefficients A_n^m, for each color channel $k_\lambda \in \{r, g, b\}$ if the overall corruption size is larger than a threshold τ:

$$A_{n,k_\lambda}^m = \begin{cases} \frac{\tau \cdot A_{n,k_\lambda}^m}{\sum_{n,m} |A_{n,k_\lambda}^m|}, & \text{if } \sum_{n,m} |A_{n,k_\lambda}^m| > \tau \\ A_{n,k_\lambda}^m, & \text{otherwise.} \end{cases} \tag{6}$$

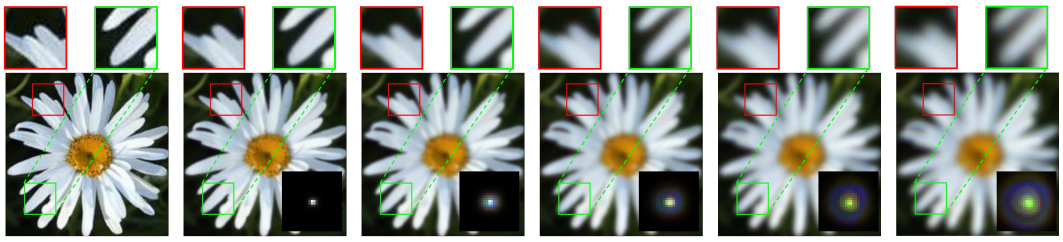


Figure 3: Adversarial optical attack examples with increasing thresholds τ. The image was created with an optical kernel from the PSF attack with the ResNet50. From left to right, τ increases from 0 to 5, where each inlet shows the corresponding kernel.

Similar to the restriction of the combined magnitude of the corruption, we also restrict each coefficient to a given element-wise threshold τ_e via:

$$A^m_{n,k_\lambda} = \begin{cases} \text{sign}_{A^m_{n,k_\lambda}} \cdot \tau_e, & \text{if } |A^m_{n,k_\lambda}| > \tau_e \\ A^m_{n,k_\lambda} & \text{otherwise,} \end{cases} \tag{7}$$

where $\text{sign}_{A^m_{n,k_\lambda}}$ denotes the sign of the coefficient.

This restriction implements a fair distribution over the corresponding corruptions. The resulting kernel is then ℓ_1-normalized per color-channel to avoid any brightness changes, when applied to an image. An example of an optically corrupted image with mixed corruption can be examined in Figure 3 for increasing corruption budget τ.

In Figure 4, we demonstrate the accuracy dependency on the restriction values τ and τ_e with the ResNet50 model on the ImageNette dataset. All other experiments are conducted with $\tau = 4$ and $\tau_e = 4$ to have a similar ℓ_2-distance as in OpticsBench [3]. Furthermore, to restrict the ℓ_2-distance of the permuted image, we integrate a δ bound and use $\nu < 1$.

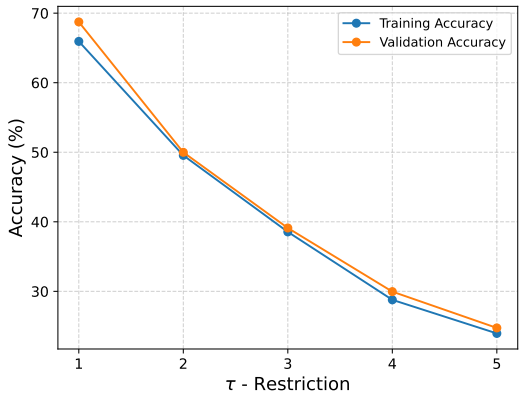


Figure 4: Ablation study of τ and τ_e threshold from 1 to 5. A minor attack, with a low τ & τ_e threshold, only marginally lowers the accuracy of the classification model. This study is conducted by using the ResNet50 model on the ImageNette dataset.

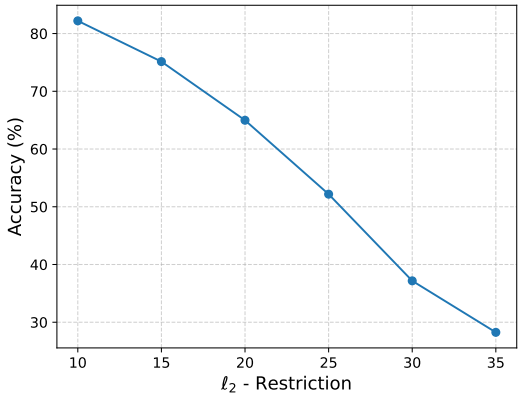


Figure 5: Ablation study of the ℓ_2-based restriction, where we bound the ℓ_2-distance to a maximum δ from $\delta \in [10, 35]$. A minor attack, with a low δ threshold, only marginally lowers the accuracy of the classification model. This study is conducted using the ResNet50 model on the ImageNette dataset.

D Model and Dataset Comparison

The initialization of ALC coefficients are the same for all experiment runs. However, over the ALC attack procedure, they start deviating from each other. The progression of the coefficients through the ALC attack is displayed in Figure 6.

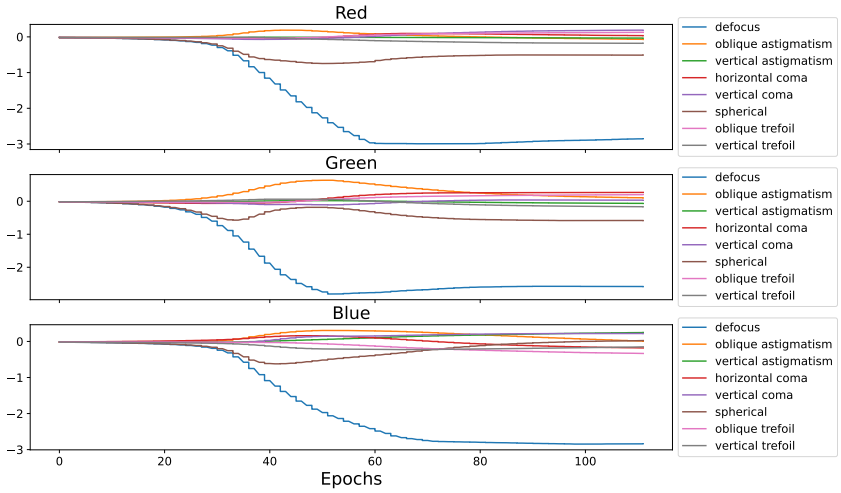


Figure 6: Updating the ALC kernels over the number of 100 epochs. The ResNet50 model [40] is trained on ImageNette [53] and subsequently used for the ALC Attack.

Figure 7 to 9 show the variance of the ALC coefficent by mulitple ALC runs. While the model architecture and training settings are constant over the five runs per model and dataset, the ALC seed is different for each run.

In Figure 10, the ALC kernels corresponding to the results in Table 1 are visualized. The first three columns represent the individual color channels, while the last column illustrates the combined ALC kernel.

D.1 ImageNet Subsets

To study the effect of ALC on dataset size and complexity, we also evaluate on different image datasets. To differentiate these datasets in size and complexity, we used subsets of the ImageNet dataset, such as the ImageNette dataset [53], the ImageNet100 dataset [54], the ImageNet1k dataset [37] and the whole ImageNet21k dataset [37].

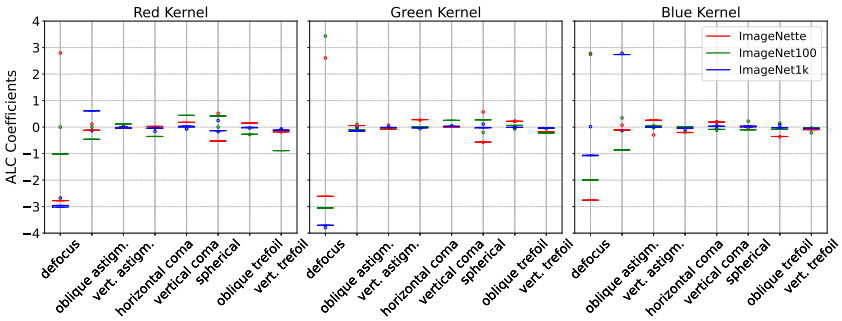


Figure 7: Comparison of the resulting ALC coefficients ResNet50 on multiple ImageNet subsets (ImageNette, ImageNet100, and ImageNet1k). The coefficient variances are low, while their respective differences are significant. Red = ImageNette, green = ImageNet100, and blue = ImageNet1k.

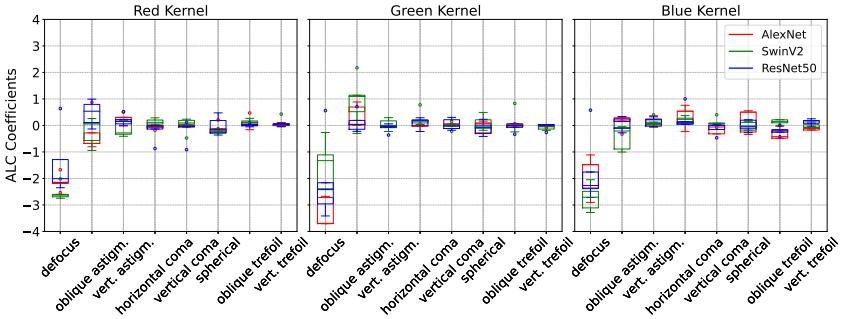


Figure 8: Comparison of the resulting ALC coefficients on multiple models (AlexNet, SwinV2, ResNet50) using ImageNette dataset. Red = AlexNet, green = SwinV2, and blue = ResNet50.

Figure 11 compares the generated optical kernels, while using the ALC attack with the same model architecture for the four different ImageNet subsets. The number of images and classes increases from top to bottom by row. Where in the top row only 10 classes are present, while adversarially attacking the ResNet50 model, in the bottom row, there are over 21k classes. However, not only the complexity, also the number of images in the dataset increases. In the figure, the first three columns show the generated optical kernel for each color channel. The fourth column displays the combined kernel, and the last column presents an example image perturbed by the attack. Figure 11 and the evaluation results in Subsection 3.1 indicate that the adverse lens corruptions are model and dataset specific. Figure 12 shows the mean magnitudes and variances for the learned aberration coefficients for ResNet50. It confirms that the variance across different ALC seeds is low while differences across datasets are significant. Figure 8 in the appendix shows that the variance increases for models trained with different seeds, while the adversarial blur corruption remains characteristic.

D.2 Transfer Attack

To evaluate to which extent the resulting corruption combinations from Subsections 3.1 and D.1 are transferable to other model architectures or datasets, we conducted a transfer attack experiment set. Therefore, we use the generated kernels from ResNet50, which were generated by the ALC attack on the four ImageNet subsets and evaluate ResNet50 models on ImageNette and ImageNet100. Results in Table 4 indicate the dataset specificity of ALC. Corruption combinations from the same dataset are more efficient than corruptions from different datasets.

To further investigate the generalizability of the ALC attack, we evaluate the performance of computation expensive state-of-the-art foundation models against a corruption combination from ResNet50 on the same dataset (ImageNet1k). Specifically, we apply these optical perturbations to three diverse architectures: CLIP, ConvNextV2, and DINOv2. The results are summarized in Table 5. ALC, originally crafted using ResNet50, remains effective across model design. However, the model

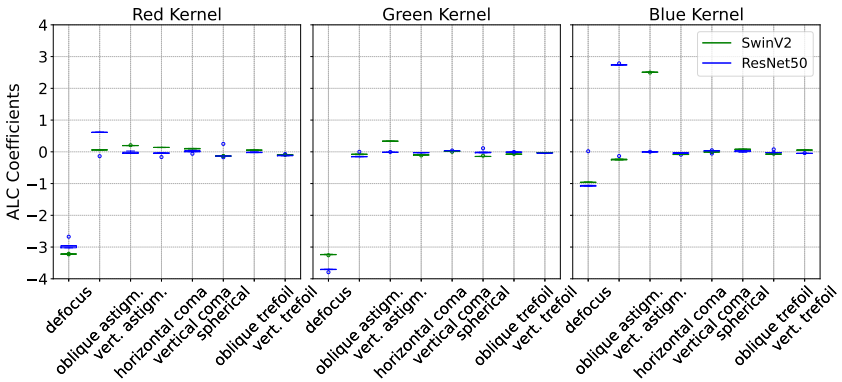


Figure 9: Comparison of the resulting ALC coefficients on multiple models (SwinV2, ResNet50) using ImageNet1k dataset. Green = SwinV2 and blue = ResNet50.

Model	ImageNette	ImageNet100	ImageNet1k	ImageNet21k
ImageNette	0.299	0.460	0.337	0.485
ImageNet100	0.237	0.158	0.224	0.318

Table 4: Accuracy of different ResNet50 models, while being attacked by the ALC attack generated kernel optimized on a *different* dataset. Columns: Datasets used by our ALC attack to create the adversarial kernel. Rows: The dataset used for *evaluation*.

specific ALC attack is reducing the accuracy of certain models, such as CLIP, significantly more, *i.e.* to 0.078 vs. 0.303 (refer to Table 1).

Attack	CLIP	ConvNeXtV2	DINOV2
-	0.761	0.923	0.912
ALC Attack	0.303	0.540	0.670

Table 5: Accuracy of state-of-the-art models (CLIP, ConvNextV2, DINOV2) with and without the corruption combination from an ResNet50 ALC attack on the same dataset.

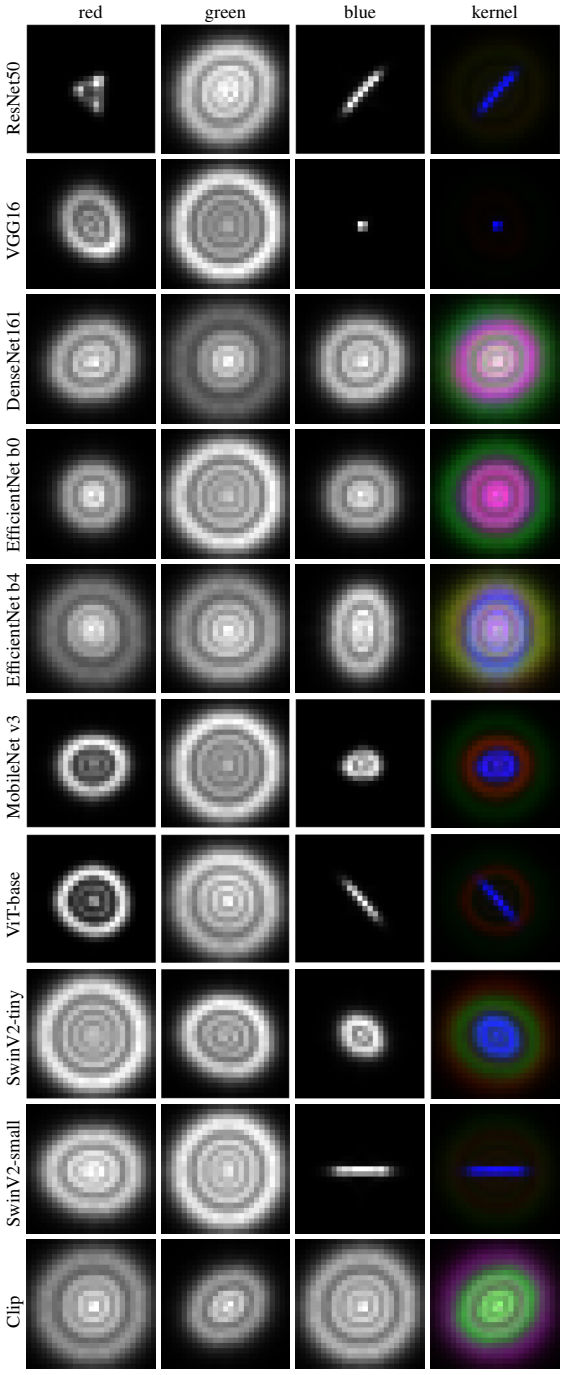


Figure 10: A comparison of the generated kernel with the ImageNet1k dataset [37] and all models from Table 1. The plot shows the generated kernel for the three color channels (red, green, and blue) and the combined kernel for all channels.

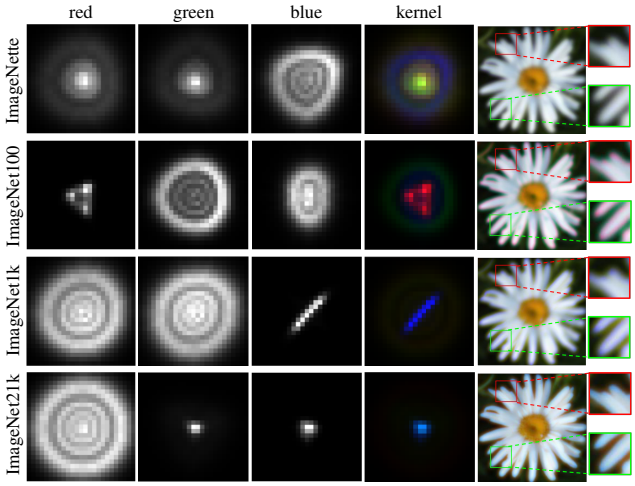


Figure 11: Difference in optical kernel generation for ImageNette, ImageNet100, ImageNet1k and ImageNet21k. The kernels for individual color channels are normalized to one for better visibility.

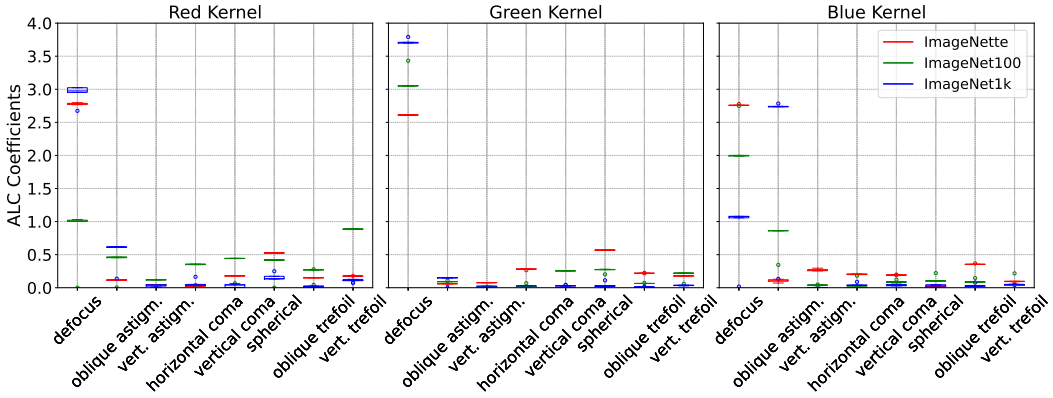


Figure 12: Comparison of the resulting ALC coefficients ResNet50 on multiple ImageNet subsets (ImageNette, ImageNet100, and ImageNet1k). See Figure 7 in the Appendix for a corresponding plot of signed ALC coefficients. The coefficient variances are low, while their respective differences are significant.

E Adversarial Training

After finding the most harmful optical kernel for a dataset-model combination in Subsections 3.1 & D.1 and the image-model combination in Section G, in this subsection, we adversarially train the models to increase the robustness against optical corruptions. For a stable training process, we conducted multiple training processes, varying the restriction values τ, as well as delaying the adversarial part of the training by varying the number of epochs, and fine-tuning with pretrained models via the optical attack from Section 2. More details and the result of these different approaches can be examined in the Appendix in Section E.

Figure 13 shows AlexNet, ResNet50 and SwinV2 adversarially trained models against the models trained only on clean data. These models are evaluated on the Common Corruptions [56] and OpticsBench [3] dataset, divided into five sections (Noise, Blur, Compression, Weather, Color) and five severities (left to right). All three models, adversarially trained, outperform the respective baseline over all sections and severities marginally. In particular, it is interesting to observe that adversarial training on blur corruptions can improve the model behavior on noise corruptions. While training image classification models adversarially with our proposed ALC attack, multiple hyperparameters for the adversarial process are adjustable. ALC specific hyperparameters are:

- τ & τ_e to control the intensity of the corruptions
- ν to bound the ℓ_2-distance
- *Attack frequency*, which determines to not attack every training batch
- *Start epoch*, which sets the first epoch, in which the ALC attack will start to attack the model on the training batches

To have a stable training process, we used multiple hyperparameter combination for our experiments. The results of these experiments can be examined in 6.

Model	τ	*AF*	*SE*	CD	ALC
Resnet50	1	1	0	0.781	0.272
	2	1	0	0.770	0.301
	3	1	0	0.772	0.324
	4	1	0	0.761	0.372
	5	1	0	0.713	0.331
	4	3	0	0.761	0.354
	4	1	20	0.755	0.294
SwinV2	4	3	0	0.851	0.299
AlexNet	4	3	0	**0.864**	**0.393**

Table 6: Ablation of adversarial training with the ALC attack on ResNet50 [40], SwinV2-tiny [49], and AlexNet [41]. First column section: The trained model. Second column section: The ALC hyperparameter combination with τ, attack frequency (*AF*) and starting epoch (*SE*). Third column section: The result on clean data (CD) and ALC attack result.

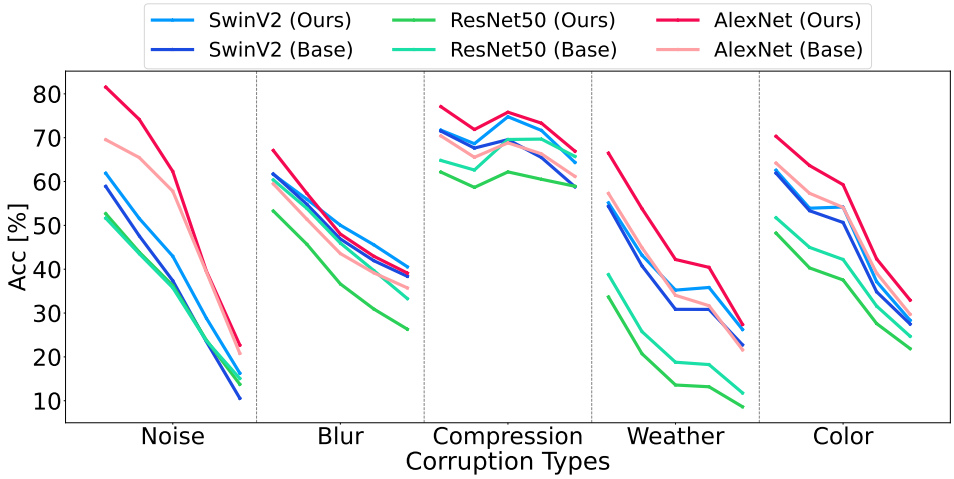


Figure 13: Adversarially trained models with ALC (ours) compared to baselines (Base). The models are evaluated on ImageNette for 2D Common Corruptions [56] and OpticsBench [3] corruptions.

F Segmentation Experiments

To evaluate the effectiveness of adverse lens corruptions in a segmentation task, we used two SAM (Segment Anything Model) models [39]. Specifically, we evaluate on the COCO Segmentation Dataset [38] to assess whether the ALC could compromise the performance of the SAM model. In Table 7, the two SAM models (base & huge) are evaluated in terms of AP50 and AP75 prior and post adverse lens corruptions. Especially the SAM base variant, with a significant drop of 50% in comparison to the clean data, is vulnerable to blur coruptions. The SAM huge variant is quite robust against ALC when comparing AP50. However, on a more fine granular level, the performance on AP75 drops drastically. We can conclude that the rough shapes can still be well segmented by SAM huge while optical aberrations blur the image data such that the segmentation becomes incorrect on fine details.

G ALC Attack per Image

In the previous sections (Sections 3.1 to F), we considered the ALC attack as an analysis tool for a given optical system, *i.e.* to generate one mixed corruption kernel for the entire dataset. Here, we alter this process to an image-specific optical attack. Therefore, instead of iterating over the whole dataset and updating the kernel after each batch, the attack iterates over one image multiple times (max. 40) and updates. This altered approach provides us insights in image-specific corruptions.

In comparison to the previous approach, the kernel and corruption diversity is increased significantly, as some images are more vulnerable against different corruptions. As a result, the accuracy between the classes also differs significantly. Table 8 in the appendix shows the drop in accuracy for each ImageNette class for the image-specific ALC on ResNet50, SwinV2, and AlexNet. Some classes, such as the *golf ball* class, have significantly lower accuracy drops while beeing attacked by ALC than other classes (e.g. *church* class), which holds for all tested models.

Model	Attack	AP50	AP75
SAM base	-	0.405	0.241
	ALC	0.198	0.102
SAM huge	-	0.818	0.460
	ALC	0.621	0.267

Table 7: Segmentation results on the COCO validation dataset. Evaluated on pre-trained SAM base [39] and SAM huge [39]. ALC significantly reduces the Average Precision (AP).

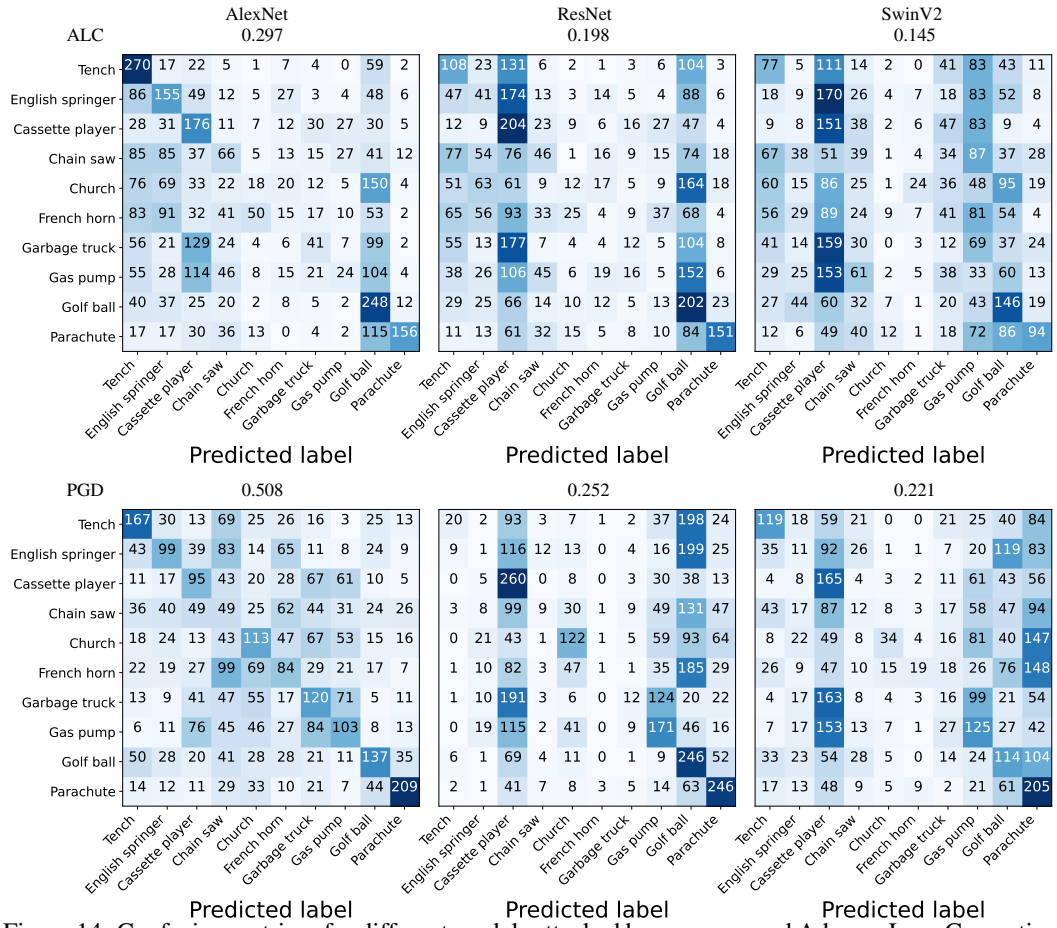


Figure 14: Confusion matrices for different models attacked by our proposed Adverse Lens Corruption (ALC) (top) and PGD (bottom) evaluated on the ImageNette dataset. The PGD attack was conducted with an ϵ bound of $\frac{2}{255}$. Each confusion matrix originates from a different model - from left to right: AlexNet, ResNet and SwinV2. The accuracies of the attacked models are displayed on top of each confusion matrix. While the absolute accuracies between the differently bounded attacks can not be directly compared, it is interesting to see that different classes are most confused by the different nature of attack (blur versus noise).

Other adversarial attack methods, such as PGD [2] and FGSM [21], attack the input images on a pixel-level and can not be reproduced by creating attack-based camera lenses. To compare the pixel-level attacks against our optical attack, we attacked the same models as in Figure 14 with an ℓ_{inf} bounded 40 step (same as for ours) PGD attack with $\epsilon = \frac{2}{255}$. These different adversarial attack approaches have significantly different performance on the different classes.

Figure 17 shows examples of image specific kernels, which are generated by the ALC attack over each image separately. The examples are randomly chosen examples from five different classes from the validation set. Furthermore, Figure 16 provides an overview of the corruption coefficient values from all ImageNette validation set images.

Table 8 shows the accuracy per class for ResNet50, SwinV2, and AlexNet. All three models have similar accuracy drops over the same classes. Furthermore, Table 9 has the ten highest (left side) and the ten lowest accuracy drops on the ImageNet1k dataset for the ResNet50 model.

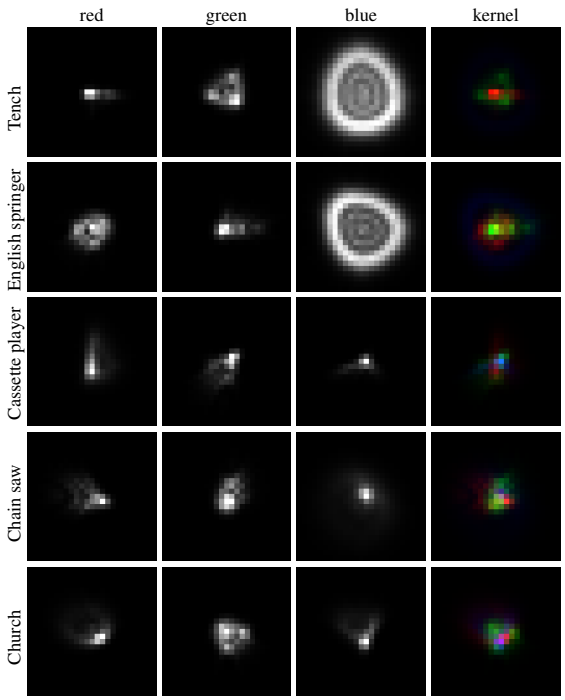


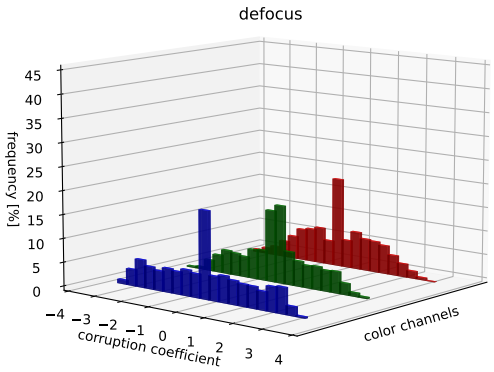
Figure 15: A comparison of the generated kernels with the ImageNette dataset [37] and the ResNet50 model. The plot shows generated image specific kernels for the three color channels (red, green, and blue) and the combined kernel for all channels.

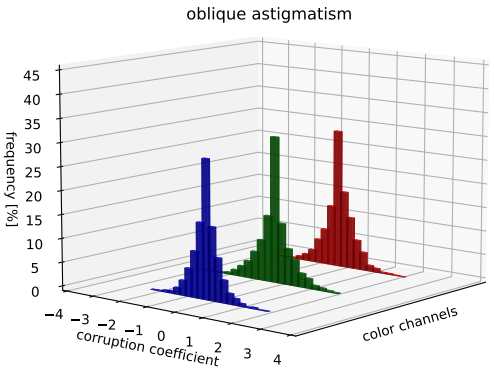
	ResNet50	**SwinV2**	**AlexNet**
Tench	0.646	0.729	0.630
English springer	0.798	0.883	0.753
Cassette player	0.241	0.355	0.361
Chain saw	0.595	0.627	0.678
Church	0.819	0.885	0.800
French horn	0.815	0.834	0.706
Garbage truck	0.815	0.876	0.777
Gas pump	0.716	0.763	0.652
Golf ball	0.306	0.468	0.197
Parachute	0.462	0.643	0.431

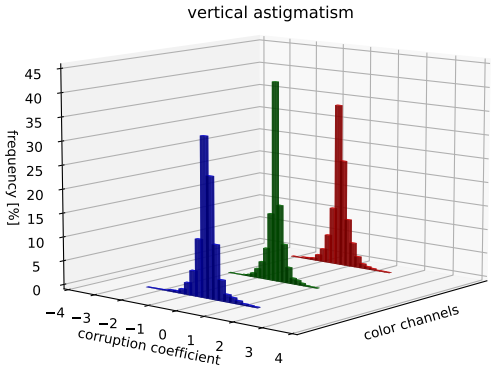
Table 8: Accuracy drop due to the per-image ALC attack, for three models (ReNet50, SwinV2, AlexNet). The attack is conducted on the ImageNette validation set.

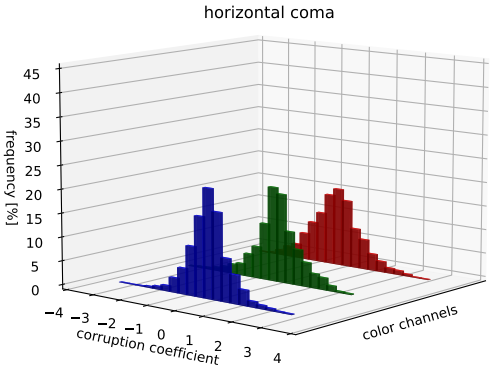
Class	**Acc. Delta**	**Class**	**Acc. Delta**
Projector	1.000	**Espresso**	0.1867
Admiral	0.980	**CD Player**	0.2222
Crash Helmet	0.980	**Jackfruit**	0.2444
Quill	0.980	**Shopping Basket**	0.2511
Bighorn	0.978	**Trailer Truck**	0.2600
Standard Poodle	0.978	**Cairn**	0.2667
Armadillo	0.978	**Water Snake**	0.2800
Table Lamp	0.978	**Sock**	0.2911
Chime	0.978	**Tile Roof**	0.2956
Prairie Chicken	0.978	**Vestment**	0.2956

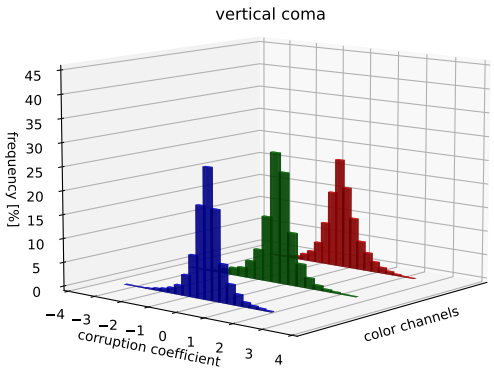
Table 9: Accuracy delta due to the per-image ALC attack on the validation ImageNet1k dataset with ResNet50. The ten largest accuracy delta values (left side) and the ten smallest accuracy delta (right side).

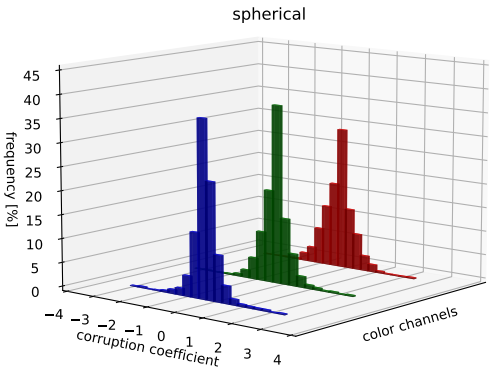

defocus
frequency [%]
corruption coefficient
color channels

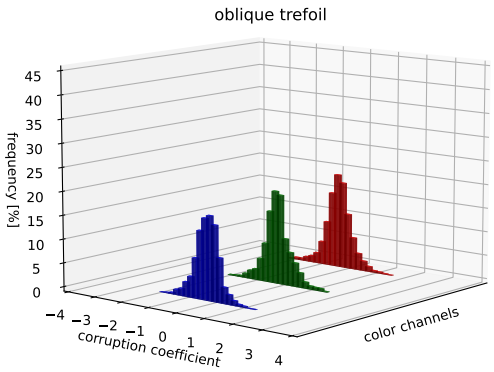

oblique astigmatism
frequency [%]
corruption coefficient
color channels

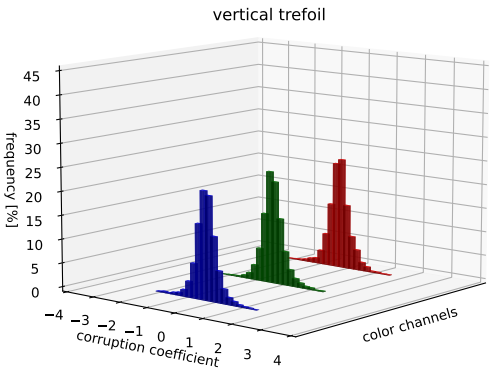

vertical astigmatism
frequency [%]
corruption coefficient
color channels


horizontal coma
frequency [%]
corruption coefficient
color channels


vertical coma
frequency [%]
corruption coefficient
color channels


spherical
frequency [%]
corruption coefficient
color channels


oblique trefoil
frequency [%]
corruption coefficient
color channels


vertical trefoil
frequency [%]
corruption coefficient
color channels

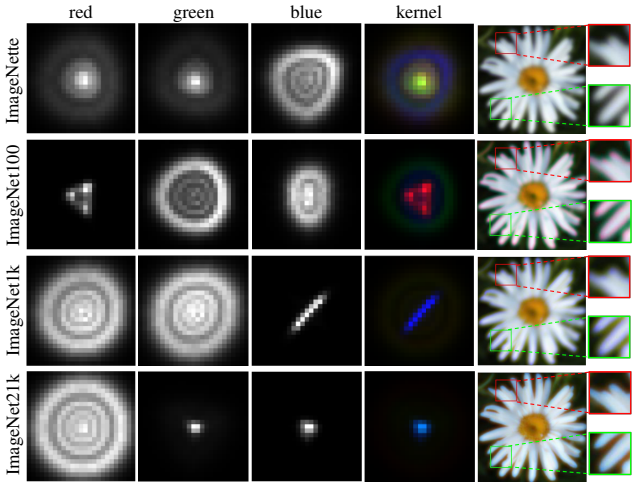


Figure 17: Difference in optical kernel generation for different Datasets. a) ImageNette, b) ImageNet100, c) ImageNet1k and d) ImageNet21k.

H Ablation on Optical Corruptions

In Figure 18 and 19 the corruption with the highest coefficients is removed and the history of the coefficients indicate a shift to spherical corruption in absents of defocus. Subsequently, vertical and horizontal coma are the highest corruption coefficients at the ALC attack.
Figure 20 and 21 show the variance of ALC coefficients on multiple dataset with five seeds per attack.

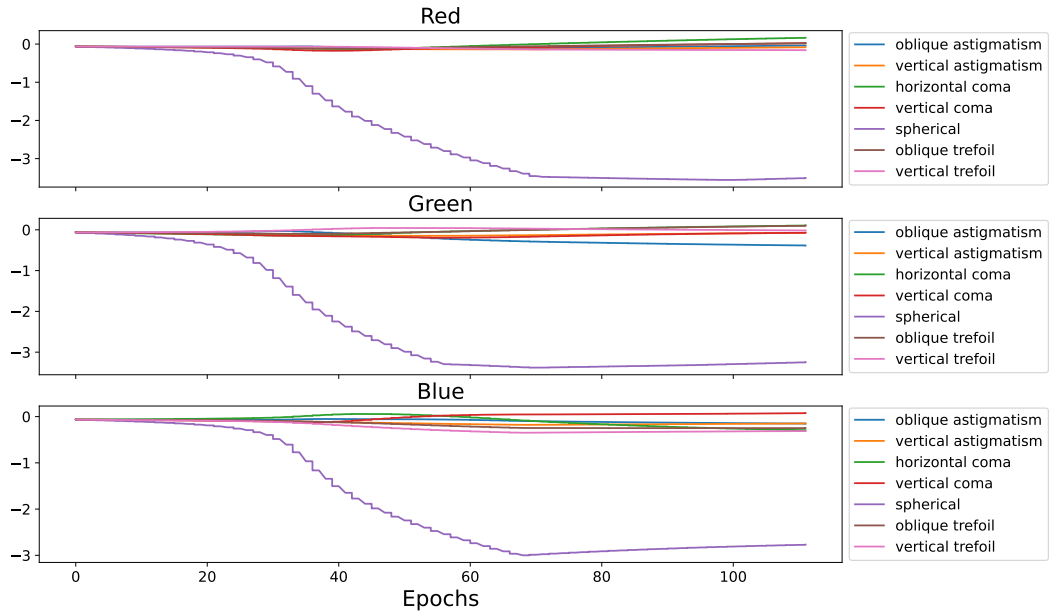


Figure 18: Running an ALC attack without the defocus optical aberration. By excluding defocus the performance decrease is lower by 4.2% in comparison to an optical attack with a defocus aberration.

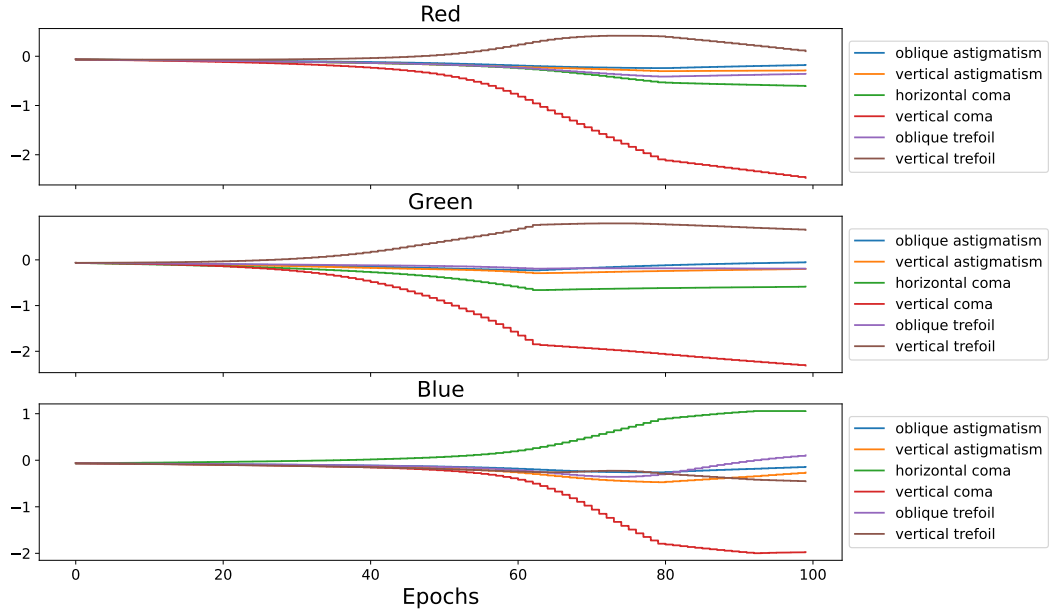
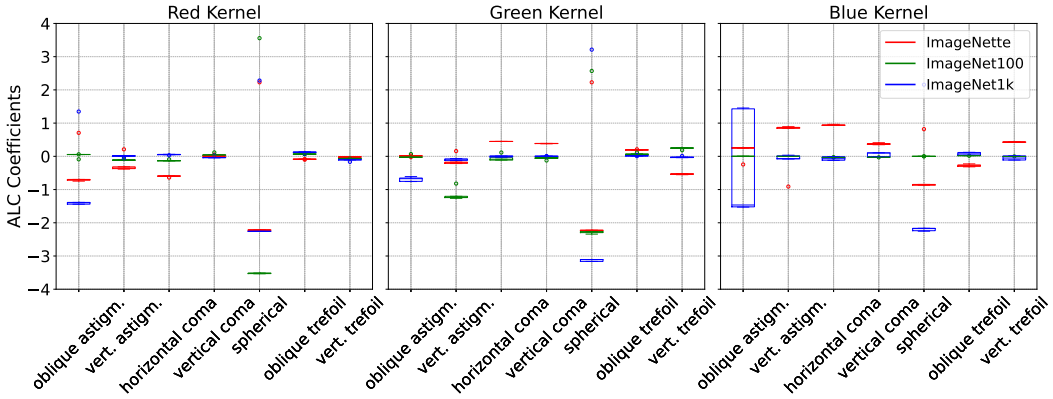
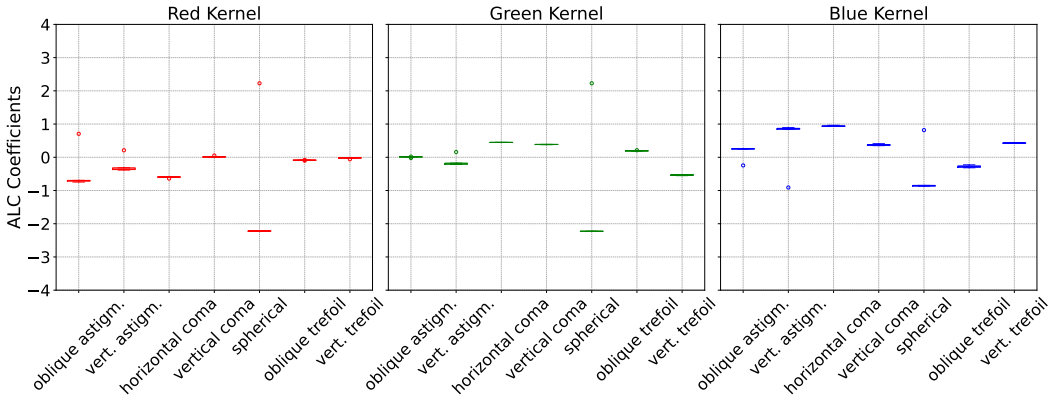


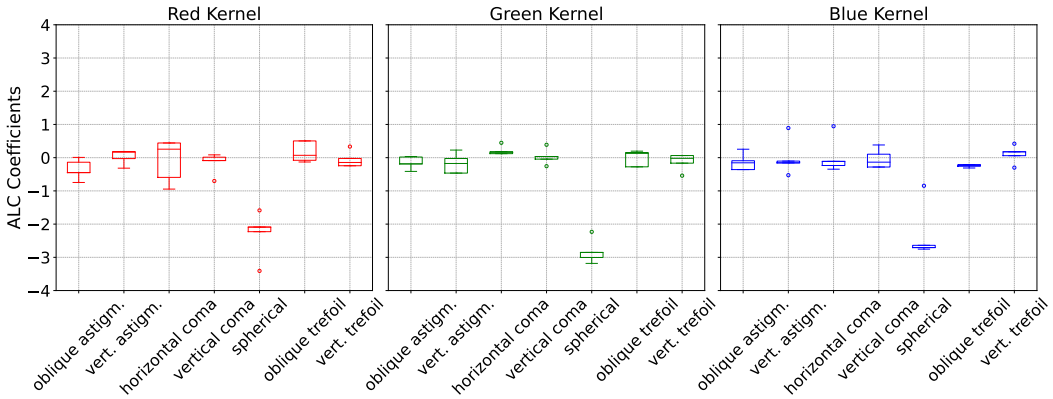
Figure 19: Running an ALC attack without the defocus and spherical optical aberration. By excluding defocus and spherical aberration, the performance decrease is 24% lower than with an optical defocus aberration.



(a) Comparison of the resulting corruptions without defocus on the 3 different ImageNet subsets (ImageNette, ImageNet100 and ImageNet1k). The adversarial attack was conducted with the same model architecture (ResNet50). The models are trained with the same seed, however the adversarial attack was conducted on 5 different seeds per dataset.

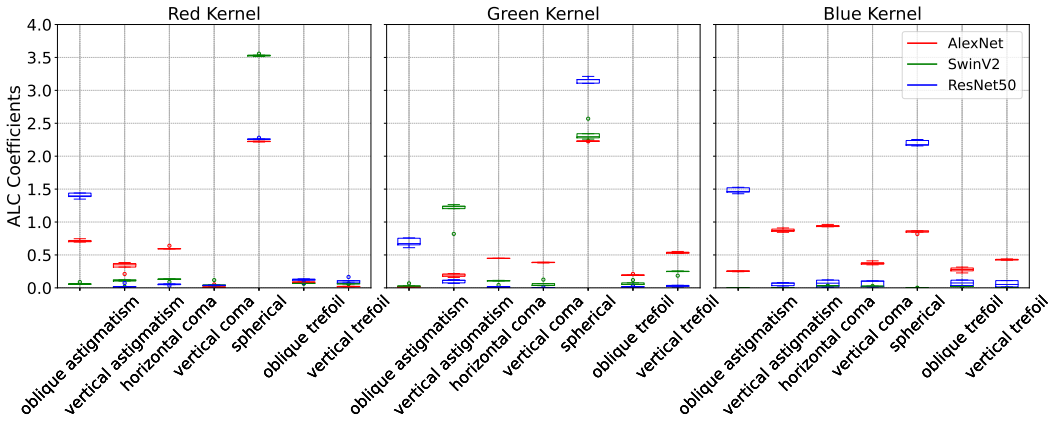


(b) Comparison of the resulting corruptions without defocus on 5 different seeds. The adversarial attack was conducted with the ResNet50 model on the ImageNette dataset. The models are trained with the same seed, however the adversarial attack was conducted on 5 different seeds.

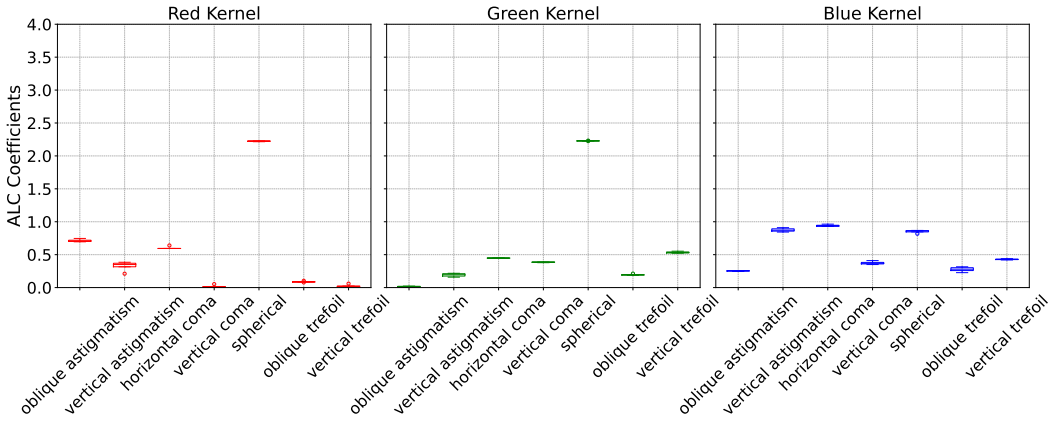


(c) Comparison of the resulting corruptions without defocus on 5 different seeds. The adversarial attack was conducted with the ResNet50 model on the ImageNette dataset. The models are trained with different seeds, and the adversarial attack was also conducted on 5 different seeds. The performance of the trained models for the multiple seeds only differ marginally (0.825, 0.816, 0.827, 0.817, and 0.806).

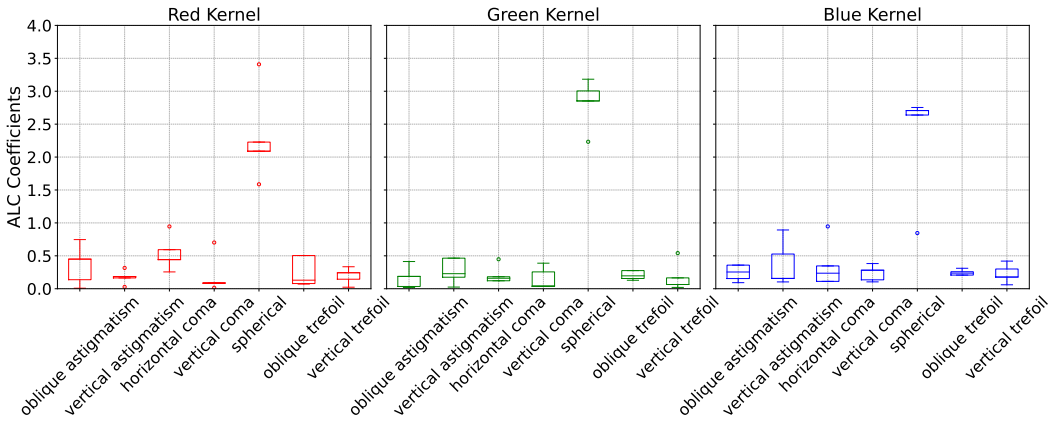
Figure 20: Variance of ALC coefficients on multiple dataset with five seeds per attack.



(a) Comparison of the resulting absolute corruptions values without defocus on the 3 different ImageNet subsets (ImageNette, ImageNet100 and ImageNet1k). The adversarial attack was conducted with the same model architecture (ResNet50). The models are trained with the same seed, however the adversarial attack was conducted on 5 different seeds per dataset.



(b) Comparison of the resulting absolute corruptions values without defocus on 5 different seeds. The adversarial attack was conducted with the ResNet50 model on the ImageNette dataset. The models are trained with the same seed, however the adversarial attack was conducted on 5 different seeds.



(c) Comparison of the resulting absolute corruptions values without defocus on 5 different seeds. The adversarial attack was conducted with the ResNet50 model on the ImageNette dataset. The models are trained with different seeds, and the adversarial attack was also conducted on 5 different seeds. The performance of the trained models for the multiple seeds only differ marginally (0.825, 0.816, 0.827, 0.817, and 0.806).

Figure 21: Variance of the magnitudes of ALC coefficients on multiple datasets with five seeds per attack.

I Texture bias

To measure the texture-shape bias of image classification models, we use the cue-conflict classification problem [57]. This dataset consists of 1,280 images where shape and texture cues conflict. The conflicting images are synthetically generated using a style transfer model [58] and ImageNet examples [37]. The conflicting cues are 16 super-classes of ImageNet. From an information standpoint, it is technically correct to predict either label (or both). Nevertheless, humans prioritize shape cues for categorization, unlike most models [57]. Employing either the shape or texture cue label as the accurate classification allows for the measurement of shape accuracy and texture accuracy, respectively. According to these assessment metrics, cue accuracy is the ratio of predictions that align with either the shape or texture label to the instance of misclassification:

$$\textit{Cue Accuracy} = \textit{Shape Accuracy} + \textit{Texture Accuracy} \tag{8}$$

We also formalize the shape bias [57] as the proportion of decisions made owing to shape over the number of correct decisions:

$$\textit{Shape Bias} = \frac{\textit{Shape Accuracy}}{\textit{Cue Accuracy}} \tag{9}$$

The evaluation of models against this metric shows that an optical attack forces the models to shift from their dependence on texture-based decisions to preferring shape-based recognition. This effect is strongly demonstrated through the shape bias of a ResNet50 model, which increases from 0.222 to 0.684. The state-of-the-art foundation model CLIP base, is known to have stronger shape bias in comparison to models, which were only trained on ImageNet1k. Comparing the CLIP shape bias from pre and post optical attack, we see a stable shape accuracy, however a drop in texture accuracy and thus also an increase in shape bias from 0.671 to 0.825. In the same way, EfficientNet-b4 and Vision Transformer models also increase shape bias from 0.411 to 0.868 and 0.398 to 0.884, respectively, but also a considerable increment in the number of correctly classified categories by shape.

Adversarial training of the AlexNet model as described in Subsection E increases the shape bias without the optical attack by 4%, while also increasing the cue accuracy by 3%. Thus, the model not only shifts its attention from texture to shape, but also is able to classify more cues correctly.

J Non-ℓ_2-boundedness

In the following, we elaborate on the non-ℓ_2-boundedness of the attack and showcase the underlying reasons. We proceed in two steps: 1) we demonstrate that the attack PSF in our approach *is* ℓ_2-bounded in the complex amplitude, but that the square in Eq. 4 removes this property, in step 2) an argument is made that the convolution itself also results in unbounded pixel differences, regardless of the PSF properties.

Regarding step 1), we note that the Zernike basis is orthonormal in the Fourier space of the lens, as is its Fourier transform [59] in the spatial domain. We can therefore consider Parseval's theorem to hold in the complex amplitude (the wave-optical PSF). Computing the intensity PSF from the wave-optical

Model	Attack	Cue Acc	Shape Acc	Texture Acc	Shape Bias
ResNet50	-	0.671	0,149	0.522	0.222
ResNet50	ALC	0.161	0.112	0.050	0.684
EfficientNet-b4	-	0.701	0,291	0,417	0.411
EfficientNet-b4	ALC	0.374	0.311	0.045	0.868
CLIP	-	0.540	0.363	0.177	0.671
CLIP	ALC	0.452	0.365	0.088	0.868
ViT-base	-	0.682	0.271	0.411	0.398
ViT-base	ALC	0.338	0.298	0.039	0.884
AlexNet	-	0.612	0.161	0.451	0.264
AlexNet	Adv. Train	0.642	0.200	0.443	0.312

Table 10: Cue accuracy, shape accuracy, texture accuracy and shape bias for multiple models with and without the ALC attack. In the last row we trained an AlexNet adversarially and this against the model without the adversarial training.

complex amplitude requires taking a square which removes the linearity. As for step 2), consider a PSF that is a shifted Dirac pulse: convolution with this kernel shifts an image by a discrete amount. The pixel-wise differences are therefore unbounded.

To conclude: it may be possible to ℓ_2-bound ALC by working in the complex amplitude domain and removing the possibility of image shifts. We have not done so because it relates to the unpractical assumption of coherent illumination, which is usually only available in laboratory settings.

K Additional practical aspects

We continue here the discussion of the practicality and possible limitations of our approach, which may be relevant to lens and system designers[1].

While we can identify worst-case aberration configurations, there is no guarantee that the obtained lens blur is related to a real lens. This can be addressed by more carefully constraining the parameter space and search directions.

Although most of the aberrations are covered by the lens design and tolerances, the manufacturing process itself can introduce aberrations of higher order[2], which should also taken into account during tolerancing. To illustrate this, we give a brief example. A synchro-speed polishing machine leads to a linear combination of Fringe modes $A_n = \{9, 16, 25\}$, which introduces locally high gradients in the pupil phase. The lens blur of such a lens is more irregular and stronger compared to the expected lower-order blur from the lens design. This holds true when both lens blurs would have the same RMS wave aberration. Therefore, the interpretation of our results with ALC should be guided with further testing of real lenses of typical effects. Including the manufacturing process in our simulation is a promising future research direction, which could *e.g.* serve as a recommendation system of which manufacturing process to use for a given image classification model.

However, in the current definition, ALC is not integrated into an end-to-end or co-optimization process of an actual lens design. Without this, it may still be hard to directly make use of the results obtained by ALC for a specific lens design. ALC therefore offers only a first step into a promising method of deep optics, and we are open for improvements and collaboration, such as testing the integration into an actual lens design process.

[1] One of the authors has worked at a leading optics company for several years and discussed with expert lens designers. The expert lens designers have confirmed the practicability and value of knowing the combination of worst-case primary aberrations for optical lens design with ALC.

[2] This detail was, among others, confirmed in a personal communication with an expert optic designer.