Received: 24 July 2024 | Revised: 30 September 2024 | Accepted: 10 October 2024 | Published online: 21 October 2024

RESEARCH ARTICLE

Medinformatics
2025, Vol. 2(4) 279–287
DOI: 10.47852/bonviewMEDIN42023933

BON VIEW PUBLISHING

A Pilot Study of Deep Learning-Based Monocular Depth Estimation from Fundus Photographs

Rony Gelman[1,2,*] and Michael D. Abràmoff[3,4,5,6]

[1]*Courant Institute of Mathematical Sciences, New York University, USA*

[2]*Retina Consultants, USA*

[3]*Department of Ophthalmology and Visual Sciences, University of Iowa, USA*

[4]*Department of Electrical and Computer Engineering, University of Iowa, USA*

[5]*Department of Biomedical Engineering, University of Iowa, USA*

[6]*Digital Diagnostics Inc., USA*

Abstract: The purpose of this study was to evaluate the feasibility of a generalizable deep learning (DL)-based system with no a priori knowledge of fundus photographs to generate monocular depth map information about optic disc structures from this imaging modality. Images of 30 stereo pairs of fundus photographs centered on the optic disc of 30 subjects were analyzed with this DL system to generate monocular depth maps using zero-shot cross-dataset transfer. These maps were registered onto reference standard depth maps derived from optical coherence tomography. Accuracy of the DL system was assessed by the root of mean squared error (RMSE) between the estimate and reference standard. 47% of the total images from the dataset were successfully processed, with mean RMSE of 0.081. Our findings demonstrate that single image, monocular depth estimation with a generalizable DL system using zero-shot cross-dataset transfer applied to retinal color fundus photographs is feasible and has potential.

Keywords: monocular depth estimation, zero-shot cross-dataset transfer, deep learning, stereo fundus photography

1. Introduction

Stereo color fundus photography is an established imaging method used to document the optic disc and estimate damage from optic neuropathies such as glaucoma, using the parallax between the stereo pair taken at slightly different angles [1, 2]. However, the stereo process is subjective as the angle is non-standard in all cameras used clinically. Specialized stereo viewer devices are not always readily available, and personal viewing techniques, such as parallel or cross-eye viewing, can be impractical and time-consuming. Furthermore, automated software tools to quantify depth information, such as enlargement of the cup, from stereo fundus photographs are lacking.

Monocular image acquisition is more patient-friendly and is standardly utilized for evaluation of retinal diseases, such as with telemedicine and artificial intelligence systems for diabetic retinopathy detection [3–6]. However, depth information from monocular images is limited. Monocular estimations, such as cup-disc ratio, or monocular cues, such as vessel overpass or deep localized notching, provide limited qualitative depth information.

***Corresponding author:** Rony Gelman, Courant Institute of Mathematical Sciences, New York University and Retina Consultants, USA. Email: rony.gelman@gmail.com

Optical coherence tomography (OCT) can precisely measure the relative depth information of the optic disc [7, 8]. However, because of their cost, OCT systems are not readily installed in most screening-based settings, such as non-ophthalmic medical offices, rural communities, or underdeveloped nations, whereas fundus cameras are more readily available. Furthermore, software tools to automatically generate depth information from OCT data are not integrated with commercial OCT systems or readily available as standalone.

There have been limited prior studies investigating depth estimation in fundus images. Two studies used pre-deep learning (DL) stereo-based methods: (1) Nakagawa et al. [9] investigated an automatic reconstruction method for the quantitative depth measurement of the optic nerve head from a stereo retinal fundus image pair. (2) Tang et al. [10] developed a depth from stereo algorithm and evaluated it on a set of stereo fundus images that have OCT-based ground truth references.

Chakravarty and Sivaswamy [11] correlated multiple depth estimates from shading, color, and texture gradients in single-color fundus images with an OCT-based depth reference. Ramaswamy et al. [12] performed supervised and unsupervised techniques on monocular fundus images to compute depth maps and found similar accuracy when evaluated on the dataset of Tang et al. [10]. Shankaranarayana et al. [13] developed a DL-based framework for

estimation of depth from a monocular stereo image. They trained and validated their framework using the dataset from Tang et al. [10] and reported they achieved significant improvement in depth estimation over the previously proposed methods. To the best of our knowledge, their framework is the only prior report using DL for the estimation of depth from a monocular fundus image. However, they used a five-fold cross-validation strategy, which may have overestimated the performance of their system given the relatively small dataset.

Ranftl et al. [14] developed a robust DL system that generates monocular depth information from images. Their generalizable DL system was extensively trained and validated on massive datasets, such as 3D films. They tested their system using zero-shot cross-dataset transfer, which tests against images not seen during training of their system. Their rationale for this approach was that "systematically testing models on datasets that were never seen during training is a better proxy for their performance 'in the wild' than testing on a heldout portion of even the most diverse datasets that are currently available".

The goal of our pilot study was to evaluate the feasibility of applying the DL system developed by Ranftl et al. [14] to generate monocular depth maps from fundus images. We utilized zero-shot cross-dataset transfer applied to the dataset from Tang et al. and our primary metrics were the portion of images successfully processed and the root of mean squared error (RMSE) between the estimate and the ground truth generated by OCT. We show the feasibility of using this DL system and believe that this pilot study lays the framework for further work as larger fundus image datasets with ground truth depth references become available. Furthermore, monocular depth estimation may have important potential for applications such as glaucoma screening using conventional monocular fundus images.

2. Research Methodology

2.1. Dataset

Images of 30 stereo pairs of fundus photographs centered on the optic disc of 30 subjects with a depth reference standard for each stereo image based on OCT were obtained from the Iowa Normative Set for Processing Images of the Retina (INSPIRE) stereo dataset [15]. These images were not annotated with clinical diagnoses.

Prior investigators [9–13] utilized the INSPIRE dataset as this was the only publicly available fundus image dataset that has a ground truth depth reference at the time of their respective reports. To the best of our knowledge, this dataset remains the only publicly available fundus image dataset that has a ground truth depth reference, hence we used this dataset for our analysis.

Tang et al. [10] report the image acquisition methodology as follows. Color slide stereo photographs centered on the optic disc of both eyes were acquired using a fixed-base Nidek 3Dx digital stereo retinal camera. The stereo images were down-scaled to 768 × 1,019 pixels by automatically locating the optic disc in the 4,096 × 4,096 images. SD-OCT scans were acquired using a Cirrus OCT scanner in the 200 × 200 × 1,024 mode. Surfaces of the retinal layer were detected in the raw OCT volume using 3D segmentation. Depth information was recorded as intensities and registered manually with the reference stereo photographs to provide ground truth for performance evaluation.

2.2. Depth map generation

Ranftl et al. [14] developed a Python programming language-based DL model for generating monocular depth information from images. Technical implementation details of their model with training, validation, and testing metrics were previously published. Their model was trained with multi-objective optimization on several established datasets containing images along with depth maps from domains such as scenes, landscapes, and movies. For testing purposes, they utilized zero-shot cross-dataset transfer, which tested their model on datasets that were never seen before during model training [16, 17].

We applied zero-shot cross-dataset transfer using the INSPIRE-stereo dataset as test input to the monocular depth map generation system developed by Ranftl et al. [14], termed "Models for computing relative Depth from a Single image" (MiDaS) [18]. We used this technique for two reasons: (1) because there is a lack of large stereo datasets with ground truth depth references, upon which to perform conventional transfer learning [19–22]; and (2) to evaluate MiDaS on a never previously seen during training dataset. MiDaS was run with PyTorch on a Linux-based high-performance computing environment.

Tang et al. [10] reported that the down-scaled stereo images in the INSPIRE-stereo dataset were cropped for their analysis; however, their crop size was not reported. Since the optimal fundus photograph crop size for MiDaS was unknown, we evaluated two sets of crop sizes: (1) 251 × 251 pixel crops and (2) 502 × 502 pixel crops. All crops were manually made by a study author (RG), a practicing ophthalmologist, and were centered around the optic disc. Thus, depth maps were generated for 120 images derived from 60 images across 30 stereo pairs for the 2 crop sizes.

2.3. Registration methods

Tang et al. [10] report that the accuracy of the disparity map generated by their algorithm was measured by the RSME between the estimate and the ground truth generated by the OCT scans. They report that to exclude those nonoverlapping regions and focus only on the main structure—cupping of the optic nerve—both maps were cropped to 251 × 251 pixels centered at the optic disc for comparison. To calculate RMSE, we attempted to register each MiDaS-generated depth map to its corresponding OCT ground truth reference using the automated registration tool bUnwarpJ supplied with ImageJ2 [23].

Since the optimal crop size of the MiDaS-generated depth map for registration to the reference OCT was unknown, we evaluated 3 registration methods: (1) using the 251 × 251 depth map generated from the 251 × 251 pixel crop processed by MiDaS; (2) from a 251 × 251 pixel crop (centered at the optic disc) of the 502 × 502 pixel depth map generated from MiDaS processing of a 502 × 502 pixel color fundus image crop; and (3) from an entire 502 × 502 pixel depth map generated from MiDaS processing of a 502 × 502 pixel color fundus image crop. In the remainder of this paper, methods 1, 2, and 3 are termed: "251 × 251 depth map", "502 × 502 depth map cropped to 251 × 251", and "502 × 502 depth map" respectively. An example illustrating these methods is shown (Figure 1).

2.4. Processing failures

The number of images across stereo pairs and the number of stereo pairs that were successfully processed were tabulated for each method. An image was considered successfully processed if both the MiDaS depth map was generated and image registration to the OCT reference succeeded. Images that could not be successfully processed were excluded from statistical

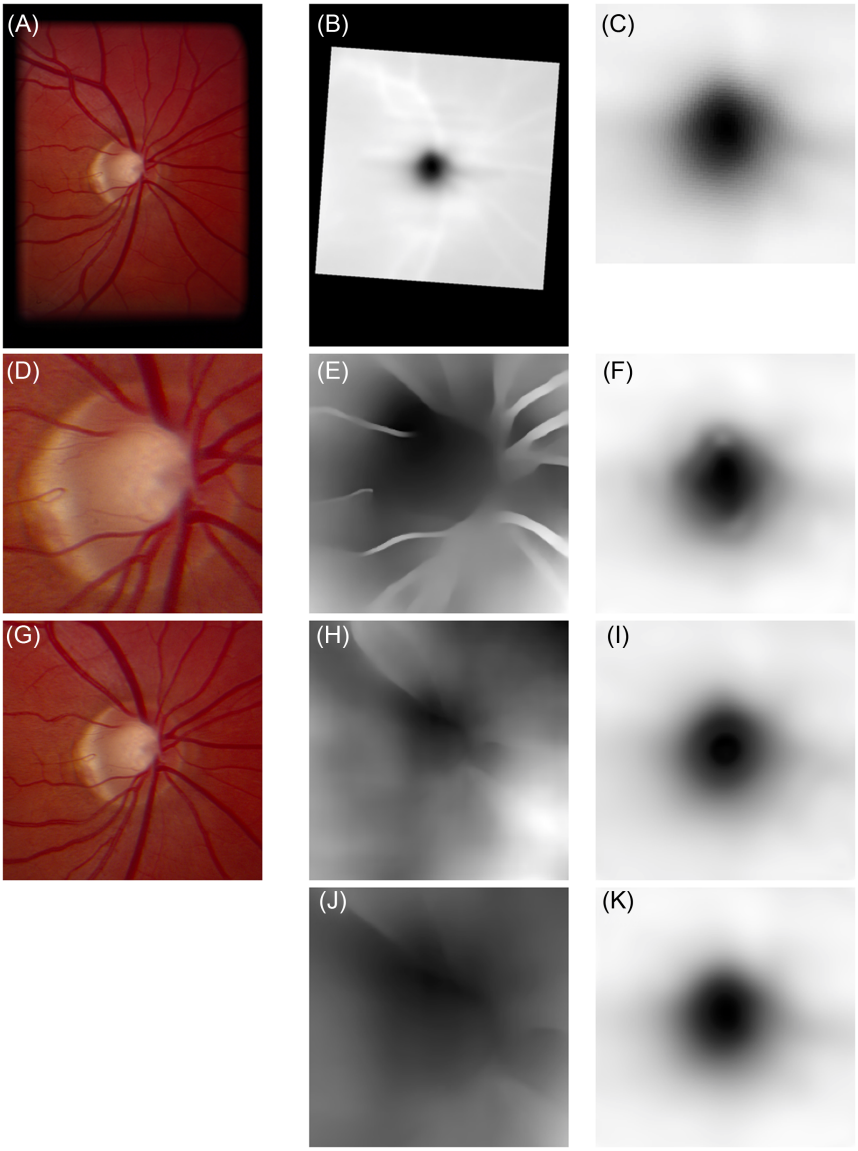

Figure 1. Representative example of an INSPIRE-stereo image successfully processed in this study. (A) A 768 × 1019 pixel image from one of the stereo pairs. (B) Accompanying 768 × 1019 pixel OCT-based depth reference. (C) 251 × 251 pixel crop of (B) centered on the optic disc. (D) 251 × 251 pixel crop of (A) centered on the optic disc. (E) MiDaS-generated depth map of (D). (F) Depth map in (E) registered onto reference (C) with RMSE of 0.032. (G) 502 × 502 pixel crop of (A) centered on the optic disc. (H) MiDaS-generated depth map of image in (G). (I) Depth map in (H) registered onto reference (C) with RMSE of 0.014. (J) 251 × 251 pixel crop centered on optic disc of (H). (K) Cropped depth map (J) registered onto reference (C) with RMSE of 0.020

analysis. RMSE was computed for each image, and summary statistics by method were tabulated. Author RG manually reviewed each case that failed processing and qualitatively assigned them into 1 of 3 categories: (1) the MiDaS-generated depth map was qualitatively sufficient, but the image registration software failed to register the image to the OCT reference; (2) the MiDaS-generated depth map was of inadequate quality; or (3) the MiDaS-generated depth map was inadequate because of over-sensitivity to vessels. Failed cases were tabulated by registration methods.

2.5. Statistical analysis

Statistical analysis was performed in R (version 3.6.3; The R Foundation). The Wilcoxon signed rank test was used to test for differences between pairwise matched data for combinations of methods. The level of significance set at 0.05 was adjusted using the Bonferroni correction accordingly. Boxplots were generated with the Python package matplotlib.pyplot.

3. Results

Summary statistics are shown (Table 1). A greater portion of the total images and stereo image pairs were successfully processed by the 502 × 502 depth map method compared to the other two methods. The median RMSE was lowest for the 502 × 502 depth map cropped to 251 × 251 method (Table 1, Figure 2). To directly compare the 3 methods, an analysis was performed on the intersection of the images successfully processed by 3 pairwise combinations of methods.

Table 1. Summary statistics

	251 × 251 depth map method	502 × 502 depth map cropped to 251 × 251 method	502 × 502 depth map method
No. (%) images successfully processed	18 (30.0)	11 (18.3)	28 (46.7)
Mean (Stdev) RMSE	0.290 (0.502)	0.040 (0.035)	0.081 (0.132)
Median RMSE	0.107	0.022	0.028
Range RMSE	0.019–1.887	0.014–0.127	0.010–0.633

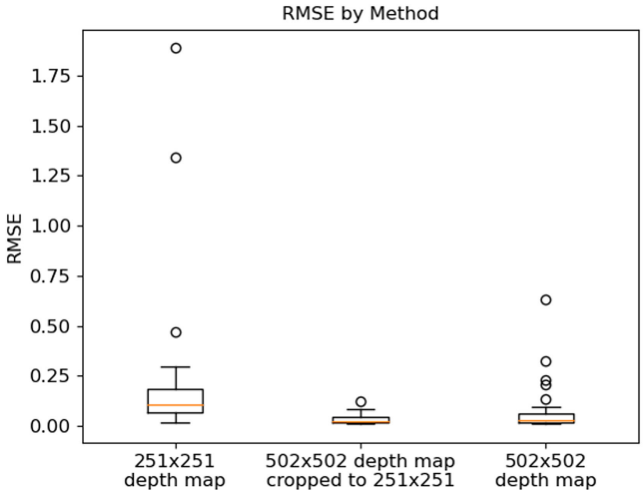

Figure 2. Boxplots of RMSE by registered depth map method. Data displayed per method are derived from all successfully processed images for the corresponding method. Each box indicates the median and first and third quartiles. The whiskers show the 10th and 90th percentiles. Circles: outliers

3.1. Comparison of registration methods

3/60 (5%) images were successfully processed by both the 251 × 251 depth map and 502 × 502 depth map cropped to 251 × 251 methods (combination 1). 6/60 (10%) images were successfully processed by both the 251 × 251 depth map and 502 × 502 depth map methods (combination 2). 8/60 (13.3%) images were successfully processed by both the 502 × 502 depth map cropped to 251 × 251 and 502 × 502 depth map methods (combination 3). Median RMSE for the combinations were as follows. Combination 1: 0.032 and 0.020 for the 251 × 251 depth map and 502 × 502 depth map cropped to 251 × 251 methods ($p = 0.75$), respectively. Combination 2: 0.068 and 0.015 for the 251 × 251 depth map and 502 × 502 depth map methods ($p = 0.03$), respectively. Combination 3: 0.027 and 0.015 for the 502 × 502 depth map cropped to 251 × 251 and 502 × 502 depth map methods ($p = 0.02$), respectively. Using a Bonferroni-adjusted level of significance set to 0.0167, none of these comparisons reached statistical significance. 3/60 (5%) images were successfully processed by the intersection of all 3 methods and were the same images from combination 1.

4/30 (13.3%), 3/30 (10.0%), and 7/30 (23.3%) stereo pairs were successfully processed by the 251 × 251 depth map, 502 × 502 depth map cropped to 251 × 251, and 502 × 502 depth map methods, respectively. Median RMSE values by (left/right) stereo pair were: 0.117 and 0.162 ($p = 0.86$) for the 251 × 251 depth map method, 0.033 and 0.019 ($p = 0.50$) for the 502 × 502 depth map cropped to 251 × 251 method, and 0.029 and 0.026 ($p = 0.94$) for the 502 × 502 depth map method (Figure 3).

3.2. Processing failures

No images were rejected by MiDaS. Causes of failed image processing are listed (Table 2) and examples shown (Figure 4). Across the 3 registration methods, most failed cases occurred because of inadequate quality generated depth maps. The 251 × 251 depth map method had a higher portion of failed cases due to over-sensitivity to vessels compared to the other 2 registration methods. Very few cases were attributable to failed registration to the OCT reference of an adequate quality generated depth map.

Causes of failed processing for each of 6 possible pairwise permutations of registration methods were tabulated (Table 2). Each permutation pair is such that the first image registration method successfully processed an image and the second registration method failed. The following were observed: (1) most cases where the 251 × 251 depth map method succeeded but the other 2 methods failed were because they generated inadequate quality depth maps; (2) most cases where the 251 × 251 depth map method failed but the other 2 methods succeeded were also due to inadequate quality, but 25–32% were due to over-sensitivity to vessels; (3) the vast majority of cases where the 502 × 502 depth map method succeeded but the 502 × 502 depth map cropped to 251 × 251 failed occurred because the cropped depth map was inadequate quality; (4) only 3 cases occurred where a 251 × 251 crop of a failed 502 × 502 depth map changed to successful processing.

4. Discussion

Depth estimation of intraocular structures has important potential clinical application, such as assessment of optic nerve cupping that occurs with glaucoma. OCT can precisely measure relative depth information of structures such as the optic disc, but may be prohibitively expensive in certain clinical settings, whereas fundus cameras are typically more available. Moreover, software tools to generate OCT-based depth maps are not currently integrated or readily available standalone.

Depth estimation techniques that utilized stereo fundus photographs have been reported. Nakagawa et al. [9] found that their depth values obtained from stereo image pairs were in accordance with reference values obtained with the Heidelberg Retina Tomograph. Tang et al. [10] tested the performance of their depth from stereo algorithm utilizing the INSPIRE-stereo dataset. Using a depth reference standard for each stereo image based on OCT, they found a mean RMSE of 0.1592, which

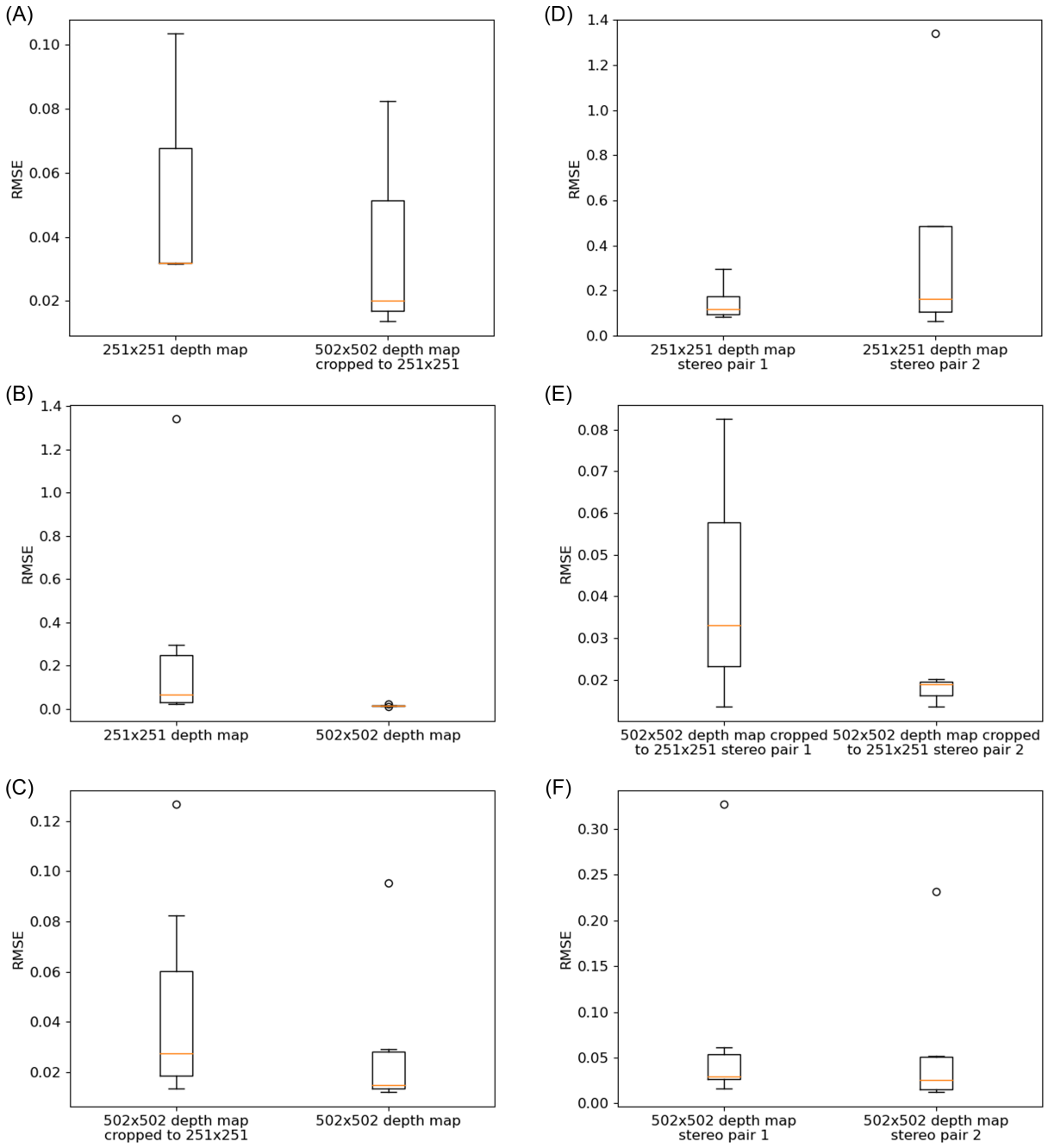

Figure 3. Boxplots of RSME by depth map methods and by stereo pairs. (A)–(C) Boxplots showing 3 pairwise combinations of depth map methods, where images for analysis are common to methods being compared. (D)–(F) Boxplots showing comparisons between first and second images from stereo pairs by depth map method, where images are from cases where both stereo pairs were successfully processed. No comparisons in (A)–(F) reached statistical significance. Each box indicates the median and first and third quartiles. The whiskers show the 10th and 90th percentiles. Circles: outliers

performed well compared to previous algorithms developed by other groups that they tested.

Although stereo fundus photography is an established technique [1, 2], it is not routinely performed in clinical practice, whereas monocular fundus photographs are more commonly acquired. Monocular depth estimation, however, is a challenging task that is less straightforward technically than stereo depth estimation as described in the report by Eigen et al. [24]. Although there have been several reports analyzing monocular image-based depth estimation for generic scenes [24–26], there have been limited reports in the domain of monocular fundus images.

Shankaranarayana et al. [13] developed a DL-based framework for estimation of depth from a monocular stereo image. To the best of our knowledge, their framework is the only prior report using DL for this task, in contrast to the prior report of Chakravarty and Sivaswamy [11] who correlated multiple depth estimates from shading, color, and texture gradients using techniques predating modern DL. Shankaranarayana et al. [13] report they achieved

Table 2. Causes of failed image processing

	Registration	Inadequate quality depth map	Vessel sensitivity
251 × 251 depth map method*	3 (7%)	25 (60%)	14 (33%)
502 × 502 depth map cropped to 251 × 251 method†	2 (4%)	43 (88%)	4 (8%)
502 × 502 depth map method‡	0 (0%)	28 (88%)	4 (12%)
Permutation 1: (A, C)⁑	0 (0%)	10 (83%)	2 (17%)
Permutation 2: (A, B)	0 (0%)	13 (87%)	2 (13%)
Permutation 3: (C, A)	2 (9%)	13 (59%)	7 (32%)
Permutation 4: (B, A)	1 (12%)	5 (63%)	2 (25%)
Permutation 5: (C, B)	2 (10%)	18 (90%)	0 (0%)
Permutation 6: (B, C)	0 (0%)	3 (100%)	0 (0%)

Note: *Denoted as "A"
† Denoted as "B"
‡ Denoted as "C"
⁑ Denoted as (Registration Method 1, Registration Method 2), where first method succeeded, and second method failed.

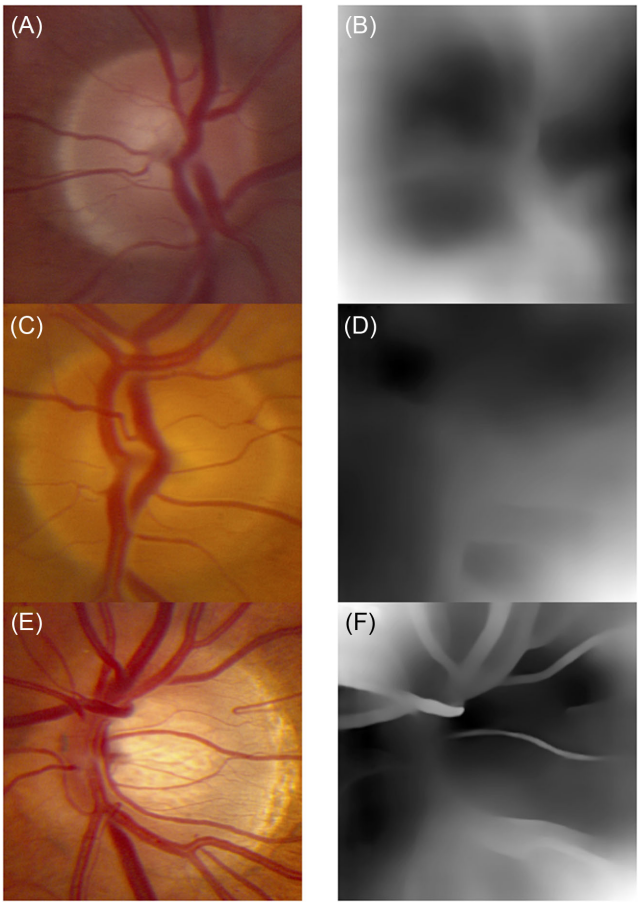

Figure 4. Examples of images that failed processing. (A) Example image with (B) adequate quality generated depth map that failed registration to the OCT reference. (C) Example image with (D) inadequate quality generated depth map. (E) Example image with (F) generated depth map that is sensitive to retinal vasculature and could not be successfully registered to the OCT reference

significant improvement in depth estimation over the previously proposed methods, with their best mean RMSE being 0.0059. Although their study used pretraining and data augmentation techniques, they may have overestimated performance because they utilized five-fold cross-validation on the relatively small INSPIRE-stereo dataset. Moreover, their depth estimation network was architected to target the task of monocular depth estimation from fundus photographs.

The motivation for our study was that stereo fundus photographs are not routinely acquired compared to conventional monocular images. Our goal was to study the feasibility of a DL-based system for generating monocular depth estimation. We believe the strengths and novelties of our study are as follows: (1) we applied a generalizable DL-based system and (2) utilized zero-shot cross-dataset transfer given the small size of the INSPIRE-stereo dataset. The key finding from this study was that monocular depth estimation using digital color fundus photographs of the optic disc with this DL system is feasible. 47% of the total images from the INSPIRE-stereo dataset were successfully processed with a depth map generation method, yielding mean RMSE of 0.081.

Our approach, which has not been applied previously to the considered problem of DL-based monocular depth estimation, was the use of a zero-shot cross-dataset transfer. Furthermore, another novelty of our study is that we analyzed a generalizable DL system with no a priori knowledge of fundus photographs. This is in contrast to the prior report of Shankaranayana et al. [13], who architected their DL system for the target domain of fundus images. Our results indicate the robustness of the generalizable DL system when applied to a never-encountered imaging domain.

A direct performance comparison between the systems developed by Tang et al. [10], Chakravarty and Sivaswamy [11], Ramaswamy et al. [12], Shankaranarayana et al. [13] and Ranftl et al. [14] was not possible because we could not successfully process the complete INSPIRE-stereo dataset. However, a strength of our study was that application of zero-shot cross-dataset transfer resulted in about half of the

INSPIRE-stereo images being successfully processed with a very good RMSE result. This finding has important implications since the MiDaS model was trained on a very different domain such as 3D movie scenes. The INSPIRE-stereo domain in contrast is fine scale in structure relative to these large-scale image domains. Furthermore, the MiDaS model was not architected with a priori knowledge of fundus images.

4.1. Effect of crop size

The optimal crop size for MiDaS depth map generation and registration to the OCT reference was unknown and explored. To the best of our knowledge, prior reports have not investigated the question of optimal crop size. Tang et al. [10] reported that they cropped the images from their dataset but did not report the crop size. Shankaranarayana et al. [13] cropped the optic nerve-head region and did not explore other crop sizes. Our analysis indicated that a larger crop of the original image resulted in a greater percentage of images successfully processed, specifically 46.7% (502 × 502 pixel crop) vs. 30% (251 × 251 pixel crop). Restricting the images to the intersection of images successfully processed by the two crop sizes, we found that there was a trend for the 502 × 502 pixel crop size to have higher mean RMSE (0.068) compared to the 251 × 251 pixel crop size (0.015), though not statistically significant ($p = 0.03$) when corrected for multiple comparisons. While not statistically significant, there was a trend for crops of the larger generated depth maps to have greater mean RMSE compared to the uncropped depth maps. The effect of crop size is illustrated with an example (Figure 1). Further study with larger datasets to understand the behavior of the DL system with variation in crop sizes will be informative.

Overall, across the methodologies we found that most failed processing cases were due to inadequate quality depth map generation. This finding is not surprising given the nature of zero-shot cross-dataset transfer and one would expect this failure rate to decrease as larger datasets become available to employ transfer learning.

4.2. Limitations

Certain study limitations are noted. (1) One of the two steps necessary for successful image processing in this study was image registration to the OCT reference. We did not test image registration tools other than the one described in this study. We note that manual review of failed cases found very few were due to image registration (Table 2). (2) Sample sizes in the statistical analysis of pairwise combinations of depth map methods were small and type I and II errors may have occurred. We used a Bonferroni correction to adjust for multiple comparisons. To the best of our knowledge, the INSPIRE-stereo dataset was the only publicly available dataset at the time our study was conducted. Future reanalysis as larger datasets become available would be insightful. (3) Although not a study goal, we found no statistically significant differences in mean RMSE between stereo image pairs when both were successfully processed. Similar to (2), this sub-analysis was limited by small sample sizes. (4) This study was not designed to assess the performance of MiDaS to discriminate between normal and glaucomatous optic discs based on cupping in-depth maps. This type of analysis requires a very large dataset of normal and glaucomatous images with ground truth references, presently unavailable. (5) Failed processing cases were manually reviewed because automated tools are presently unavailable. Incorporating a validated automated tool in a future study would be beneficial.

5. Conclusions

In summary, we believe that this pilot study demonstrates that monocular depth estimation with a generalizable DL system (MiDaS) using zero-shot cross-dataset transfer applied to color fundus photographs is feasible and has potential. Conceivably, applying conventional transfer learning, such as with fine-tuning, to train and validate MiDaS using a large set of color fundus photographs could result in a higher number of successfully processed images and lower RMSE. As larger datasets with ground truth references become available, exploring this approach will be important future work and could have potential applications for glaucoma screening and decision support systems utilizing monocular fundus photography.

As new monocular depth estimation models using encoder backbones evolve, future work analyzing the performance of models using vision transformer architectures would be an area of interest to explore [27, 28]. Increasing the size of datasets via synthetic data generation with Variational AutoEncoder Generative Adversarial Network [29] would be an interesting avenue of exploration, once such methodologies are further developed and validated.

Future clinical applications can potentially investigate the practical application of monocular depth estimation with this generalizable DL system. For example, it would be insightful to obtain depth information about choroidal lesions, such as choroidal nevi or melanomas, using monocular color fundus photography or multicolor imaging [30]. Future work can also potentially grade the degree of papilledema using monocular depth estimation on images processed with a validated DL-based system that can differentiate among optic disks with papilledema, normal disks, and disks with nonpapilledema abnormalities [31].

Acknowledgement

RG would like to thank Professor Carlos Fernandez-Granda from the Courant Institute of Mathematical Sciences at New York University for his mentorship and support. RG would like to thank the NYU IT High-Performance Computing Center for computing resources, services, and staff expertise.

Ethical Statement

This research study adhered to the tenets of the Declaration of Helsinki. This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in Iowa Carver College of Medicine at https://medicine.uiowa.edu/eye/inspire-datasets, reference number [15]; in Github at https://github.com/isl-org/MiDaS, reference number [18]; in PyTorch at https://pytorch.org/tutorials/beginner/transfer_learning_tutorial.html, reference number [19]; in ImageJ Docs at https://imagej.net/software/imagej2/, reference number [23].

Author Contribution Statement

Rony Gelman: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Michael D. Abràmoff:** Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

References

[1] Tanna, A. P., & Boland, M. V. (2024). Section 10: Glaucoma. In American Academy of Ophthalmology (Ed.), *2024–2025 basic and clinical science course* (pp. 65–72). American Academy of Ophthalmology.

[2] Tyler, M. E. (1996). Stereo fundus photography: Principles and technique. *Journal of Ophthalmic Photography*, *18*(2), 68–81.

[3] Lawrence, M. G. (2004). The accuracy of digital-video retinal imaging to screen for diabetic retinopathy: An analysis of two digital-video retinal imaging systems using standard stereoscopic seven-field photography and dilated clinical examination as reference standards. *Transactions of the American Ophthalmological Society*, *102*, 321–340.

[4] Abràmoff, M. D., & Suttorp-Schulten, M. S. A. (2005). Web-based screening for diabetic retinopathy in a primary care population: The EyeCheck project. *Telemedicine Journal & e-Health*, *11*(6), 668–674. https://doi.org/10.1089/tmj.2005.11.668

[5] Scanlon, P. H. (2008). Article commentary: The English national screening programme for sight-threatening diabetic retinopathy. *Journal of Medical Screening*, *15*(1), 1–4. https://doi.org/10.1258/jms.2008.008015

[6] Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digital Medicine*, *1*(1), 39. https://doi.org/10.1038/s41746-018-0040-6

[7] Geevarghese, A., Wollstein, G., Ishikawa, H., & Schuman, J. S. (2021). Optical coherence tomography and glaucoma. *Annual Review of Vision Science*, *7*(1), 693–726. https://doi.org/10.1146/annurev-vision-100419-111350

[8] Biousse, V., Danesh-Meyer, H. V., Saindane, A. M., Lamirel, C., & Newman, N. J. (2022). Imaging of the optic nerve: Technological advances and future prospects. *The Lancet Neurology*, *21*(12), 1135–1150. https://doi.org/10.1016/S1474-4422(22)00173-9

[9] Nakagawa, T., Suzuki, T., Hayashi, Y., Mizukusa, Y., Hatanaka, Y., Ishida, K., …, & Yamamoto, T. (2008). Quantitative depth analysis of optic nerve head using stereo retinal fundus image pair. *Journal of Biomedical Optics*, *13*(6), 064026. https://doi.org/10.1117/1.3041711

[10] Tang, L., Garvin, M. K., Lee, K., Alward, W. L. W., Kwon, Y. H., & Abràmoff, M. D. (2011). Robust multiscale stereo matching from fundus images with radiometric differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(11), 2245–2258. https://doi.org/10.1109/TPAMI.2011.69

[11] Chakravarty, A., & Sivaswamy, J. (2014). Coupled sparse dictionary for depth-based cup segmentation from single color fundus image. In *Medical Image Computing and Computer-Assisted Intervention: 17th International Conference*, 747–754. https://doi.org/10.1007/978-3-319-10404-1_93

[12] Ramaswamy, A., Ram, K., & Sivaprakasam, M. (2016). A depth based approach to glaucoma detection using retinal fundus images. In *Proceedings of the Ophthalmic Medical Image Analysis International Workshop*, 9–16. https://doi.org/10.17077/omia.1041

[13] Shankaranarayana, S. M., Ram, K., Mitra, K., & Sivaprakasam, M. (2019). Fully convolutional networks for monocular retinal depth estimation and optic disc-cup segmentation. *IEEE Journal of Biomedical and Health Informatics*, *23*(4), 1417–1426. https://doi.org/10.1109/JBHI.2019.2899403

[14] Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., & Koltun, V. (2022). Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(3), 1623–1637. https://doi.org/10.1109/TPAMI.2020.3019967

[15] University of Iowa, Department of Ophthalmology and Visual Sciences. (n.d.). *Inspire datasets* [Data set]. https://medicine.uiowa.edu/eye/inspire-datasets

[16] Larochelle, H., Erhan, D., & Bengio, Y. (2008). Zero-data learning of new tasks. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, 2*, 646–651.

[17] Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In *Conference on Computer Vision and Pattern Recognition*, 1521–1528. https://doi.org/10.1109/CVPR.2011.5995347

[18] Rbirkl. (n. d.). *MiDaS* [data set]. GitHub. https://github.com/isl-org/MiDaS

[19] Chilamkurthy, S. (n. d.). *Transfer learning for computer vision tutorial*. PyTorch. https://pytorch.org/tutorials/beginner/transfer_learning_tutorial.html

[20] Aggarwal, C. C. (2023). *Neural networks and deep learning: A textbook* (2nd ed.). Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-031-29642-0

[21] Kim, H. E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M. E., & Ganslandt, T. (2022). Transfer learning for medical image classification: A literature review. *BMC Medical Imaging*, *22*(1), 69. https://doi.org/10.1186/s12880-022-00793-7

[22] Chan, H. P., Samala, R. K., Hadjiiski, L. M., & Zhou, C. (2020). Deep learning in medical image analysis. In G. Lee & H. Fujita (Eds.), *Deep learning in medical image analysis: Challenges and applications* (pp. 3–21). Springer. https://doi.org/10.1007/978-3-030-33128-3_1

[23] ImageJ Docs. (2023). *ImageJ2*. Retrieved from: https://imagej.net/software/imagej2/

[24] Eigen, D., Puhrsch, C., & Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, *2*, 2366–2374. https://dl.acm.org/doi/abs/10.5555/2969033.2969091

[25] Liu, F., Shen, C., & Lin, G. (2015). Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5162–5170. https://doi.org/10.1109/CVPR.2015.7299152

[26] Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., & Navab, N. (2016). Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth International Conference on 3D Vision*, 239–248. https://doi.org/10.1109/3DV.2016.32

[27] Birkl, R., Wofk, D., & Müller, M. (2023). *MiDaS v3.1 – A model zoo for robust monocular relative depth estimation*. arXiv. https://doi.org/10.48550/arXiv.2307.14460

[28] Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision*, 12179–12188. https://doi.org/10.1109/ICCV48922.2021.01196

[29] Razghandi, M., Zhou, H., Erol-Kantarci, M., & Turgut, D. (2022). Variational autoencoder generative adversarial network for synthetic data generation in smart home. In *IEEE International Conference on Communications*, 4781–4786. https://doi.org/10.1109/ICC45855.2022.9839249

[30] Muftuoglu, I. K., Gaber, R., Bartsch, D. U., Meshi, A., Goldbaum, M., & Freeman, W. R. (2018). Comparison of conventional color fundus photography and multicolor imaging in choroidal or retinal lesions. *Graefe's Archive for Clinical and Experimental Ophthalmology*, *256*(4), 643–649. https://doi.org/10.1007/s00417-017-3884-6

[31] Milea, D., Najjar, R. P., Jiang, Z., Ting, D., Vasseneix, C., Xu, X., . . . , & Biousse, V. (2020). Artificial intelligence to detect papilledema from ocular fundus photographs. *New England Journal of Medicine*, *382*(18), 1687–1695. https://doi.org/10.1056/NEJMoa1917130

How to Cite: Gelman, R., & Abràmoff, M. D. (2025). A Pilot Study of Deep Learning-Based Monocular Depth Estimation from Fundus Photographs. *Medinformatics*, *2*(4), 279–287. https://doi.org/10.47852/bonviewMEDIN42023933