Deep Learning-Based Metadata Inference for Photographs Captured with Manual Lenses: Challenges and Opportunities

Luis Alfredo da Silva
Graduate Program in Electrical Engineering
State University of Santa Catarina
Joinville, Santa Catarina, Brazil
luis.ads@edu.udesc.br

Rafael Stubs Parpinelli
Graduate Program in Applied Computing
State University of Santa Catarina
Joinville, Santa Catarina, Brazil
rafael.parpinelli@udesc.br

***Abstract*—This paper presents a deep learning approach to infer missing metadata, specifically focal length, aperture, and subject distance, from photographs taken with manual lenses. The lack of lens metadata on photographs taken with those lenses (due to their inability to electronically communicate metadata) poses challenges for photographers, archivists, and researchers who rely on this information for image organization, forensic analysis, and computational photography applications. Motivated by this, we develop a ResNet-based model adapted for regression tasks and train it on a diverse dataset of 2,200 original images, augmented to approximately 28,000 usable, representative samples. The model demonstrates encouraging results in both aperture and focal length estimation (MAPE $<$ 26%).**

***Index Terms*—Metadata inference, Deep learning, Convolutional Neural Networks, Dataset augmentation, Computational photography**

I. Introduction

In the realm of modern photography, metadata serves as the backbone for efficient image organization, analysis, and understanding. Digital single-lens reflex (DSLR) and mirrorless interchangeable-lens (ILC) cameras rely on electronic communication between the lens and the camera body to capture essential metadata, such as aperture (f-stop) and focal length. However, manual lenses, both vintage and contemporary, lack this electronic interface, leading to missing or incomplete metadata in photographs.

This research addresses the challenge of inferring missing metadata from images captured with manual lenses. Motivated by the needs of photographers and researchers who require accurate metadata for tasks such as image categorization, reproduction, and detailed analysis, we aim to develop a reliable deep learning-based solution. Our primary objective is to estimate the focal length, aperture, and subject distance — key metadata elements — from photographs lacking this information. Achieving this goal will significantly enhance the organizational, educational, and forensic applications of photographic images.

The contemporary photography industry's dependence on metadata is profound, particularly for key parameters like aperture and focal length. These values are typically obtained through electronic communication between the lens and the camera body in DSLRs and ILCs, a feature commonly found in modern autofocus lenses. However, manual lenses, which lack this electronic capability, present a significant metadata gap. These lenses can be vintage models adapted for modern cameras or contemporary manual lenses designed for specific camera mounts, both contributing to the growing popularity of manual photography due to their cost-effectiveness and technological simplicity.

The absence of metadata in images captured with manual lenses poses substantial challenges. For instance, it hinders precise image categorization, analysis, and reproduction, which are essential in forensic investigations, educational contexts, and photo organization. Addressing this issue is crucial for advancing various applications that rely on comprehensive image metadata.

A. Motivation

The motivation behind this research stems from the practical difficulties encountered when working with photographs from manual lenses. Without metadata, understanding the technical settings used to capture an image becomes exceedingly challenging. This information is vital for reproducing similar shots, organizing photo collections, and conducting in-depth analyses. Therefore, there is a compelling need for a robust method to infer missing metadata values, ensuring that photographs taken with manual lenses can be as informative and usable as those from electronic lenses.

B. Objectives

The central objective of this research is to develop a deep learning system capable of accurately inferring missing metadata from photographs taken with manual lenses. Specifically, we aim to estimate focal length, aperture (f-stop), and subject distance—metadata elements crucial for understanding image characteristics. By achieving this, we will facilitate better organization, analysis, and educational use of photographic images. This system has the potential to revolutionize forensic analysis, historical archive digitization, and artistic exploration, making it a valuable tool for both professional and amateur photographers.

C. Contributions

This paper presents several significant contributions to the field of computational photography and metadata inference:

Model Architecture and Adaptation: We introduce a deep learning model based on the ResNet architecture, tailored for inferring missing metadata from photographs captured with manual lenses. ResNet's resilience in image classification tasks inspired our adaptation for metadata estimation, a regression problem. Despite the challenges, our model returned promising results and serves as a foundational framework that can be further enhanced and refined.

Comprehensive Evaluation: Our evaluations include a battery of metrics: Mean Absolute Error (MAE), Mean Absolute Percentage Error, (MAPE), Mean Squared Logarithmic Error (MSLE), and Mean Squared Error (MSE). While aperture estimation showed promising results, focal length and subject distance estimation remained challenging, underscoring the need for further dataset enhancements and model improvements.

Practical Tool and Collaboration: We provide a practical tool for photographers and researchers, offering a starting point for more sophisticated metadata inference systems. By making our code publicly available, we aim to foster collaboration and further research in this domain.

In summary, our contributions include:

- Developing a deep learning model for metadata inference using ResNet.
- Conduct a thorough evaluation of a diverse, augmented dataset.
- Demonstrating the impact of dataset quality and augmentation techniques.
- Providing a practical tool for the photography and research community.

II. RELATED WORK

The domain of metadata inference for photographs, especially those taken with manual lenses, remains relatively unexplored. This section synthesizes relevant literature on optics, convolutional neural networks (CNNs), and the adaptation of CNNs for regression tasks. We also discuss existing metadata inference work and highlight the challenges and opportunities within this field.

A. Optical fundamentals and Related Research

While metadata inference for manual lenses remains largely unexplored, recent work in computational photography provides relevant foundations. Smith [1] and Hecht [2] established optical principles governing aperture-focal length relationships, while Cira et al. [3] demonstrated CNN adaptations for regression tasks in image analysis. Chen et al. [4] further validated residual networks for nonlinear regression. Our work extends these concepts to the novel domain of manual lens metadata inference. Despite the different purpose, a recent study on body measurement inference from 2D images [5] reveals similar challenges related to neural network training for regression tasks on images.

B. Convolutional Neural Networks and Vision Transformers

CNNs have revolutionized computer vision tasks due to their hierarchical feature learning capabilities from images ([6], [7], [8], [9], [10]). When choosing a model architecture for this research, different alternatives were considered. Our choice, ResNet ([6]), introduced residual blocks to mitigate the vanishing gradient problem, enabling the training of very deep networks. It's possible that alternatives like EfficientNet [7], and the MobileNet family [8], [9], [10], could also work for the task at hand, however, ResNet, a powerful and widely recognized architecture in image classification, was chosen as the backbone of our metadata inference model. ResNet's capability to learn intricate features and handle high-dimensional data made it a compelling choice for this regression task.

Unlike CNNs, Transformer-based architectures [11] weren't considered due to computational and dataset size requirements.

C. Adapting CNNs for Regression Tasks

While CNNs are traditionally trained for image classification, they can be effectively adapted for regression tasks, aiming to predict continuous values [4]. This adaptation involves:

Output Layer Modification: Changing the final fully connected layer to output continuous values, such as focal length and aperture.

Loss Functions: Employing loss functions like Mean Absolute Error (MAE), Mean Squared Error (MSE), and Mean Squared Logarithmic Error (MSLE) to guide model training by minimizing the difference between predicted and actual values.

Regularization: Techniques like dropout and L2 regularization prevent overfitting, crucial in regression tasks due to the continuous output nature.

D. Relevant Studies and Challenges

Though limited, existing studies on metadata inference offer valuable insights:

Cira et al. [3] adapted CNNs for regression, predicting arrow orientation on road pavement using rectified orthophotography. Their work demonstrates CNN's potential in regression and the value of adapting architectures like VGGNet and ResNet.

Chen et al. [4]. developed a deep residual learning model for nonlinear regression, showcasing the effectiveness of adapted ResNet architectures in regression tasks.

Despite progress, challenges remain, including, but not limited to, the following:

Dataset Limitations: Obtaining large, diverse datasets with accurate metadata is crucial. Biases can lead to overfitting and poor generalization [5].

Real-World Variability: Capturing real-world variations in lighting, scene composition, and other factors is vital for robust model development.

Future research should address these challenges through dataset expansion, model refinement, cross-validation, and exploration of metadata-aware data augmentation and feature analysis.

III. METHODOLOGY

This study develops a deep learning approach to infer focal length and aperture from image content. We combine: (1) a modified ResNet architecture adapted for regression, (2) a dataset built with metadata-aware augmentation to preserve optical relationships, (3) a training process leveraging transfer learning, and (4) an inference process to test, obtain results and use the model in the real world.

A. Model Architecture

We fine-tuned several ResNet variants, ResNet18, ResNet34, and ResNet50, each offering distinct trade-offs between model complexity and performance. In the results section, comparisons between the different network sizes are made.

To adapt ResNet for metadata inference, we implemented the following strategic modifications:

Output Layer: The final fully connected (FC) layer was changed to produce a vector of two values, representing the estimated focal length and aperture. This customization tailored the model's output to our specific metadata inference objective.

Dropout Regularization: A dropout layer was incorporated at the FC layer to mitigate overfitting. However, as the experimental results demonstrated, the impact of this regularization technique was not as positive at high dropout values (0.5), instead a very small one was used (0.1 to 0.2, depending on the different model sizes trained).

Optimizer: We opted for the Adam optimizer, renowned for its adaptive learning rate and prowess in handling non-convex problems. L2 regularization was enabled to curb overfitting.

Loss Function: Mean Absolute Error (MAE) was chosen as the loss function due to its robustness to outliers and its alignment with the preference for consistently accurate estimates in this application. MAE calculates the average magnitude of errors, providing a reliable measure of model performance.

Layer Unfreezing: Initially, ResNet layers 1, 2, 3, and 4 were frozen to preserve pre-trained features. Layers 4 and 3 were progressively unfrozen to fine-tune the model, allowing for more precise adjustments to the learned features. As the results will indicate, the best checkpoints were consistently achieved before unfreezing layer 3, suggesting that further fine-tuning may not yield performance improvements.

B. Dataset Preparation

The creation of a robust and diverse dataset is foundational to the success of any deep learning endeavor. In this section, we detail the meticulous process of constructing and Pre-processing our dataset, ensuring it adequately represents the various metadata values encountered in manual lens photography.

C. Data Collection

Our dataset was curated in two stages to ensure comprehensive coverage of metadata values and real-world variability:

1) **Systematically Produced Photographs:** We captured approximately 1400 photographs under controlled conditions, intentionally representative of the field of view (often using consequent perspective distortions) and depth of field (often using amount and characteristics of background blur), employing a range of electronic lenses (which provide reliable EXIF lens data for aperture and focal length). Each shot was carefully planned to cover focal lengths from 20mm to 640mm, f/stops from f/2.0 to f/22, and subject distances from 10cm to infinity (all values are 35mm full-frame equivalent). This controlled capture provided us with a precise ground truth for metadata values, essential for training and evaluation.
2) **Real-World Photographs:** To infuse our dataset with authentic diversity, we use an additional 800 photographs from the authors' photographic portfolio. These images were selected to represent a wide array of metadata values, complementing the systematically produced set. The real-world photographs should bring real-life variability in scene content, lighting conditions, and composition, enhancing the model's ability to generalize.

D. Pre-processing and Augmentation

Pre-processing and data augmentation are critical steps in preparing a high-quality dataset for deep learning. We implemented the following procedures:

EXIF Extraction and Normalization: For each image, we extracted the relevant metadata - focal length, aperture, and subject distance - from the EXIF data. These values were then meticulously normalized to a range of [0, 1] to facilitate efficient learning by the model. It is important to note that subject distance was excluded from the training process due to its inconsistent availability in the dataset (not all used lenses provided this metadata).

Metadata-Aware Augmentation: To enhance the dataset's diversity, we applied cropping and rotation techniques. Each augmented image underwent EXIF metadata recalculation to ensure that the augmented version retained accurate metadata values consistent with the original image. This approach preserved the correct metadata relationships, although it was limited in generating shorter focal lengths and larger f/stops. The order of augmentation steps was crucial to prevent blurry interpolation; augmentations were applied to the original-sized images (up to 4000x4000 pixels) to ensure that the final resized output (896x896 pixels) maintained the desired metadata characteristics without introducing artifacts.

Resizing and Normalization: All images, both augmented and original, were resized to a uniform resolution of 896x896 pixels. This resolution struck a balance between preserving image quality and managing computational demands. The order of pre-processing steps is important to avoid undesired interpolation: all augmentations are made on original size

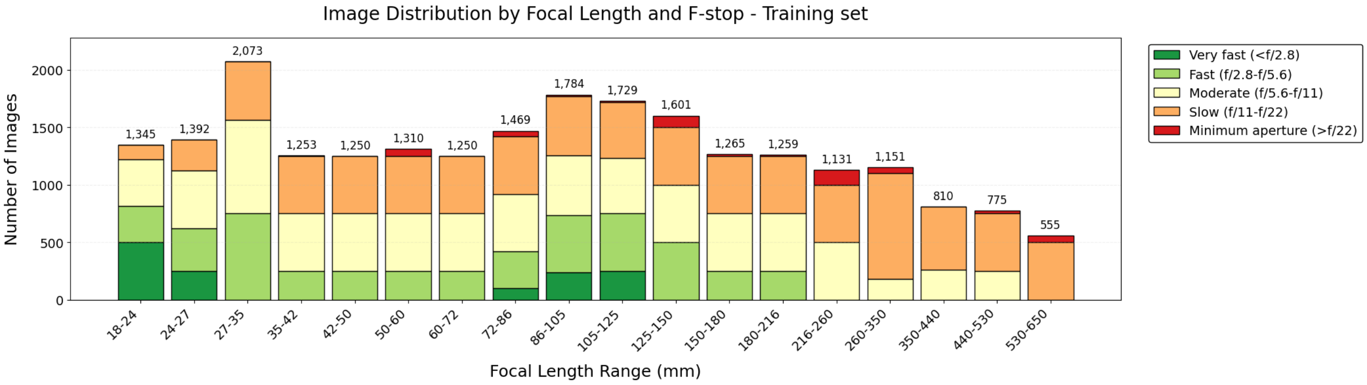

Fig. 1. Dataset distribution across focal length ranges and f-stop categories for training set.

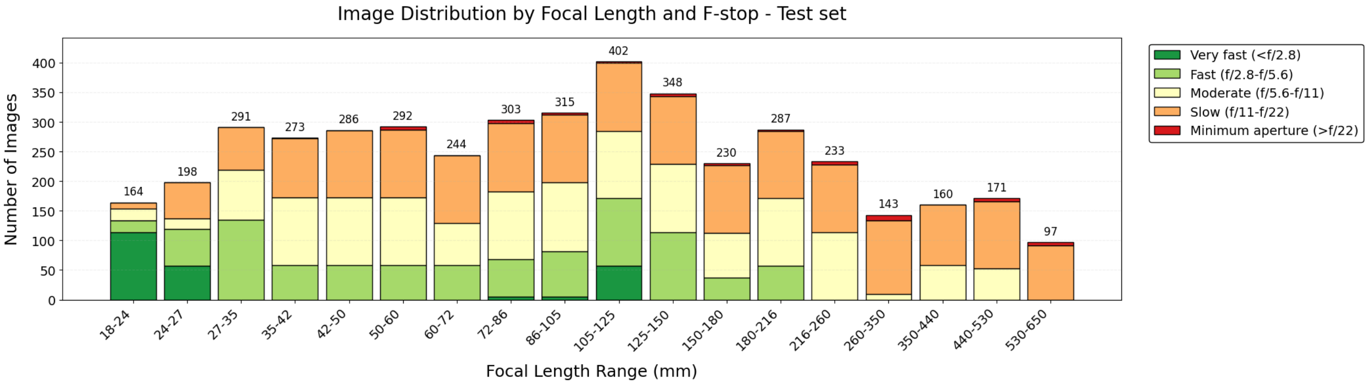

Fig. 2. Dataset distribution across focal length ranges and f-stop categories for test set.

images (up to 4000x4000 pixels) in such a way that augmentations themselves never result in a crop that would be smaller than the 896x896 output (so they can safely be resized to that size without risk of a resize introducing blurry enlargement).

E. Dataset Distribution

The dataset preparation process began with 2,155 original images (real-world photographs, approximately 70% of them captured with the specific intent of being representative of field of view and depth of field effects, the remaining 30% being typical photographs from the authors' portfolio), with diverse metadata values – focal lengths ranging from 20mm to 640mm, f/stops from f/2.0 to f/22 (smaller f/stops up to f/48 were used in initial tests, but discarded later on due to over-representation of apertures not common in real-world scenarios where this research would apply), and subject distances from 10cm to infinity (all values are 35mm full-frame equivalent).

Those originals were augmented to 64,650 samples through metadata-aware transformations (which will be explained in detail below). After careful curation to maintain balanced representation across focal lengths and apertures, the final dataset comprised:

- Training set: 19,803 images
- Validation set: 4,443 images
- Test set: 4,437 images

The remaining augmented pictures were culled and left as a bonus set for further testing (however, unlike the true training/validation/test sets, its distribution is not curated).

The distribution of images across focal length and aperture combinations is shown in Figure 1 for the training set and Figure 2 for the test set. The distribution for the validation set is similar to the test set. The augmentation process successfully expanded the dataset while preserving the physical relationships between focal length, aperture, and subject distance through sensor size equivalence calculations. However, the augmentation was inherently limited to generating longer effective focal lengths and smaller apertures than the original images, as the transformations simulate cropping and consequent changes in the field of view.

Our experiments demonstrated that the quality and diversity of the images, along with the use of metadata-aware augmentation, play a crucial role in the model's ability to learn and generalize effectively. Despite the initial plan to use 20,000 source images (pre-augmentation) for training, we found that many of these images were inadequate due to missing metadata or irrelevant content. For example, most indoor shots were less useful because they rarely provided a significant amount of depth of field. This finding underscores the importance of high-quality, diverse datasets in deep learning applications and the effectiveness of metadata-aware augmentation techniques.

F. Dataset Distribution Technique

To ensure balanced representation across the range of focal lengths and apertures, we implemented a systematic dataset distribution strategy. Our approach categorizes images into bins based on combined focal length and f-stop categories, then distributes samples across training, validation, and test sets while trying to maintain proportional representation.

The focal length bins were designed with emphasis on ultra-wide (18-24mm) and telephoto (260-650mm) ranges, reflecting the greater technical challenge in estimating extreme focal lengths. F-stop categories were simplified into five groups from ”Very fast (<f/2.8)” to ”Minimum aperture (>f/22)” to capture the nonlinear impact of aperture on image characteristics.

Our distribution algorithm follows these key steps:

1) **Bin Identification:** Each image is assigned to a focal length bin and f-stop category based on its metadata.
2) **Original Sample Preservation:** All non-augmented (original) samples are retained in the training set to maintain data authenticity.
3) **Augmented Sample Selection:** Augmented samples are selectively included to balance representation, with a maximum of 200 samples per bin to prevent over-representation.
4) **Validation/Test Set Formation:** Samples not selected for training are distributed to validation and test sets, with care taken to include all bins proportionally.
5) **Bin Balancing:** For bins missing from validation or test sets, up to 20% of the training samples are reallocated to ensure complete coverage.

G. Training Process

The training process leveraged an RTX 3090 GPU with 24 GB of VRAM, on a system with an 18-core Xeon E5-2696v3 CPU and 128 GB of system RAM. System CPU and RAM are important for the metadata extraction and metadata-aware data augmentation techniques used, as those do not leverage GPU resources. As a software platform, PyTorch was selected for its robustness, flexibility in defining custom models, and efficient GPU acceleration capabilities. Training followed by a training > validation > testing methodology.

Early Stopping: The training process was configured to halt if the validation loss did not improve for 10 consecutive epochs, a value determined through empirical testing.

Additional augmentation: Training images had additional randomization of characteristics that wouldn’t interfere with the associated metadata, done in-place at training time. Those augmentations include flipping, contrast, and brightness changes, and explicitly do not include any form of cropping or rotation (as such changes require careful metadata changes in the way we explained in the Metadata-Aware Augmentation section). This exposes the model to a more varied dataset, whose characteristics keep changing slightly during each epoch.

This chapter outlined the meticulous methodology employed in crafting the metadata inference model, emphasizing the critical steps in dataset preparation, model architecture, and training processes. The careful selection of techniques and parameter sets the stage for evaluating the model’s performance and identifying avenues for future enhancements.

IV. RESULTS AND DISCUSSION

Three models were finetuned with optimal settings (batch size, learning rate, and dropout varying for each) and the same dataset: ResNet18, ResNet34 and ResNet50. ResNet18 failed completely to converge to acceptable results on all tried settings (best Test MAPE 39%, when the task needs a value lower than 25% for a usable model), while ResNet34 (best Test MAPE 25%) and ResNet50 (best Test MAPE 20%) provided good results. Figure 4 show test results for each version on MAPE. All other figures show results specific to the ResNet50 version (which gave us the best results), where Figure 3 demonstrates the metrics coming from training, validation, and test results across all checkpoints, and identifies the best checkpoint-based on test results on the given metric, by evaluating through MAE(3a), MAPE(3b), MSLE(3c) and MSE(3d).

During training, all checkpoints are saved, as well as training metadata including, for each epoch, training and validation loss, MAPE, MSLE and MSE metrics. After training is finished, all checkpoints run through the test set to get test metrics on loss, MAPE, MSLE, and MSE as well. The best checkpoints, as presented in I, are found based on test performance.

The Table I also shows the training time, number of epochs, best epoch by each metric and average performance of the model according to our standardized testing on the best epoch by MAPE (which we found out was the metric that indeed gave the best results on our tests, tailored to measure real-world performance of the model more clearly).

After the best checkpoint is found, more thorough testing is done to assess its performance in a real-world scenario:

The model is used for full inference of metadata on every picture on the test set, and the best, worst, and randomly assorted results are shown to the user (figure 5), showing the tested image, the real metadata, and the inferred metadata. This was done to give an easy assessment of model performance in its real application.

The mean percentage errors (figure 6) for FL (6a) and f/stop (6b), looked particularly easy to compensate for using simple math, so simple biasing arithmetic was added to the inference code, eliminating systematic error and further improving performance on the test dataset.

In addition, to give a better understanding of the model performance characteristics and where lie its strengths and flaws, error trends and residuals (figure 7), for focal length (7a, 7c) and f/stop (7b, 7d) were all plotted against the best checkpoint. Published here are the results for the ResNet50 model after biasing.

TABLE I
RESULTS FOR THE THREE MODELS TRAINED

Model	Total training epochs	Total training + testing time*	Best epoch				Average error**	
			Loss	MAPE	MSLE	MSE	f/stop	Focal length
ResNet18	65	3:30h	58	**60**	58	58	26,80%	25,48%
ResNet34	68	9:40h	59	59	59	59	14,31%	20,49%
ResNet50	92	10:58h	46	**68**	61	46	**9,77%**	**16,69%**

* For reference only, some code changes (DataLoader optimizations) and hardware changes (load, power limits) changed between different training runs)

** After applying systematic bias corrections

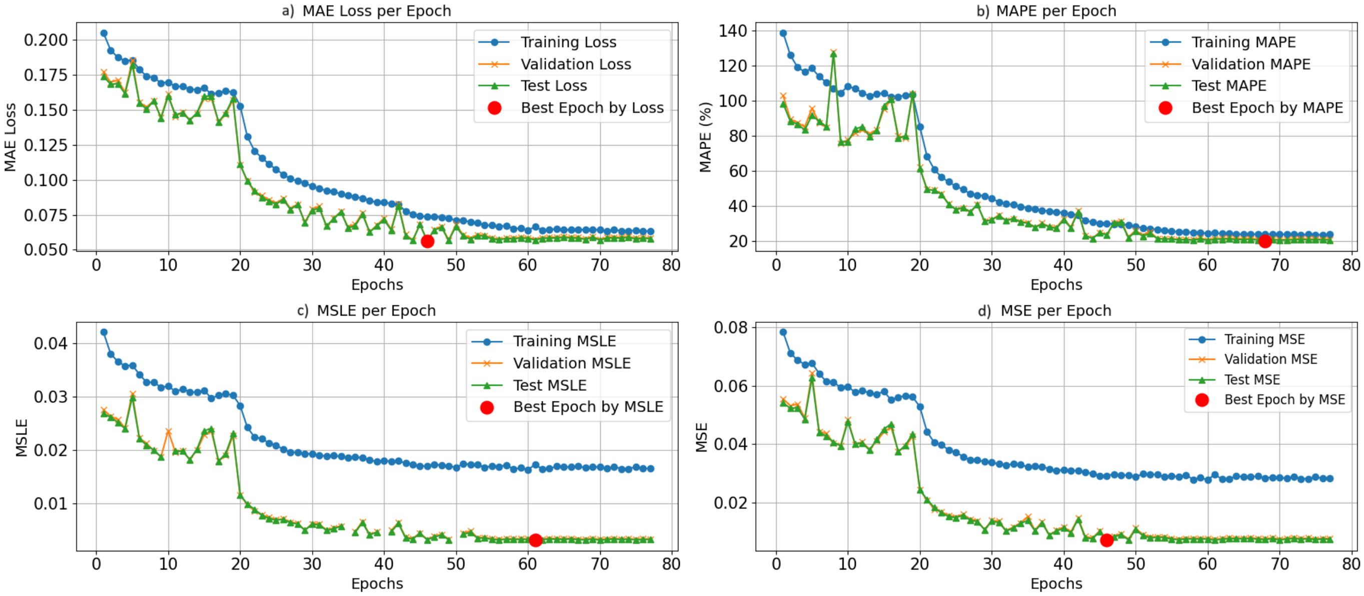

Fig. 3. MAE(a), MAPE(b), MSLE(c) and MSE(d) metrics for training, validation, and test on the ResNet50 model

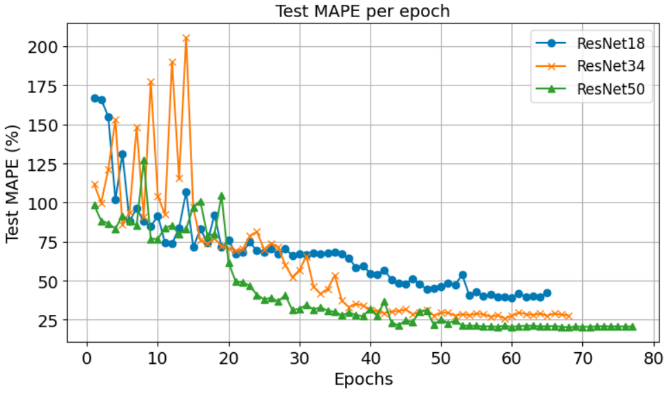

Fig. 4. Test MAPE results per epoch for each model size

Given the detailed results, the model seems to give reasonable performance (we consider errors under 20% acceptable for our usage) for focal length and f/stop on all ranges.

For f/stops, despite the average error being seemingly good, performance is bad for larger f/stops. The average results look good, probably due to the dataset distribution (seen before in this paper) having proportionally more pictures on the 'good performance' range, which suggests performance could be improved with a more representative dataset for the larger f/stops, possibly by using proportionally more metadata-aware data augmentation for the larger f/stops than the smaller ones.

Key Observations:

- The performance gap between ResNet34 (25% MAPE) and ResNet50 (20% MAPE) suggests deeper architectures better capture subtle optical patterns;
- Systematic focal length underestimation (Fig. 6a) indicates potential scaling issues in regression layers;
- Limited large f-stop performance correlates with dataset underrepresentation (Fig. 1).

A. Limitations

The dataset used in this study is limited in size and may contain biases that affect the model's generalization. The dataset consists of approximately 2,200 original images,

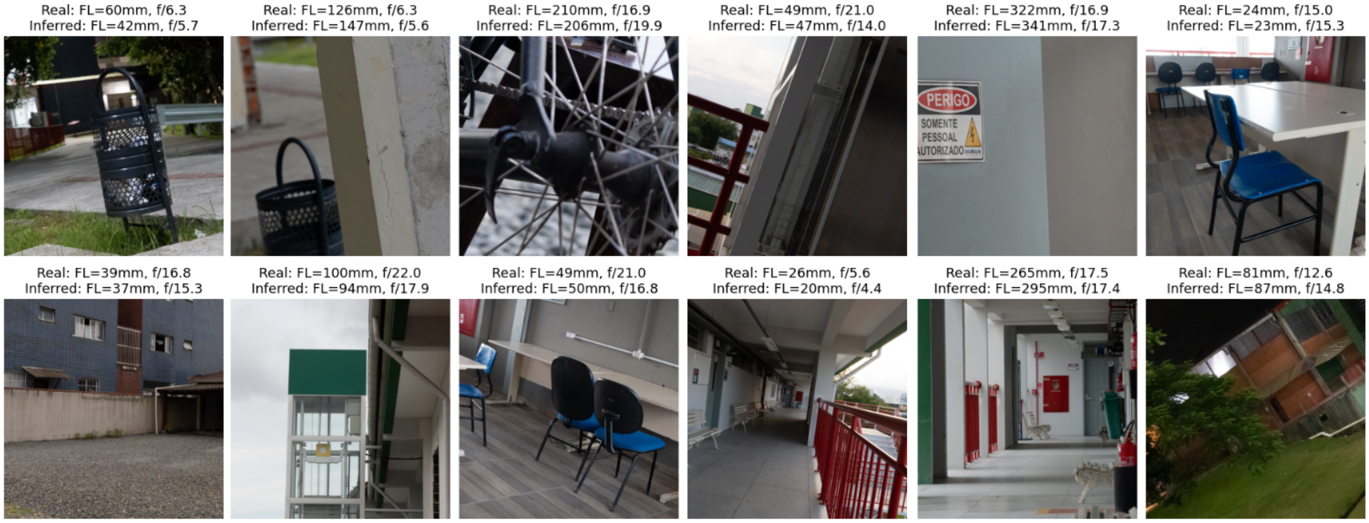

Fig. 5. Example of randomly assorted inference results on the test set.

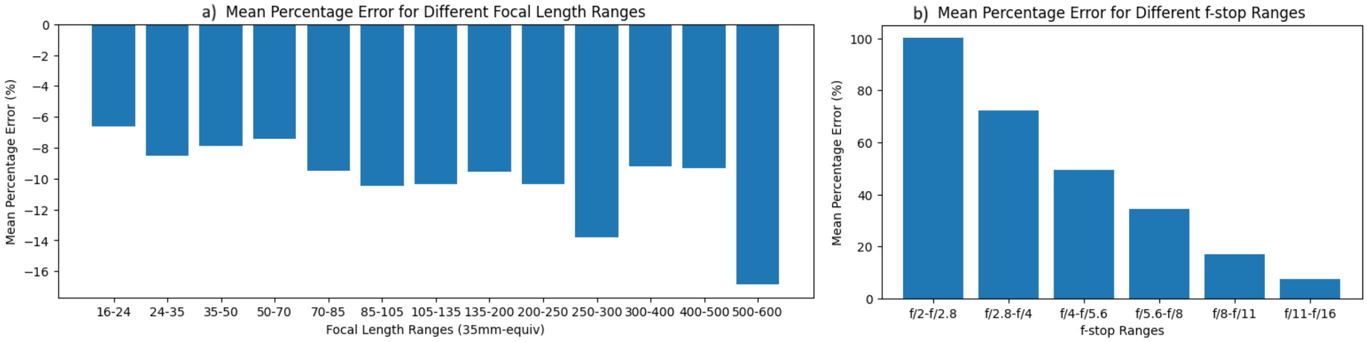

Fig. 6. Mean errors for FL (a) and f/stop (b) before inference-time biasing

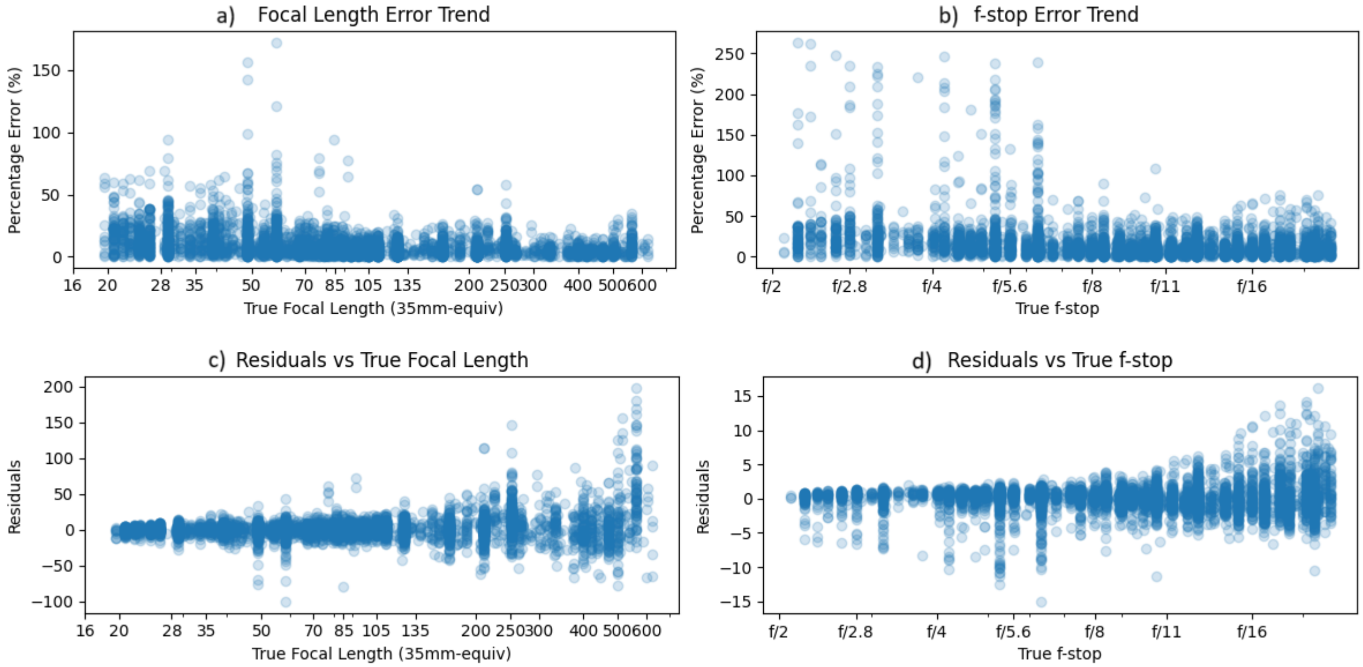

Fig. 7. Error trends (a)(b) and Residuals(c)(d) for FL and f/stop after inference-time biasing.

which were expanded to approximately 64,000 images using metadata-aware augmentation techniques, most of those were then culled, and the remaining were sorted between training, validation, and testing sets. While these techniques help increase the dataset size, they are constrained by physical limitations and cannot produce a wide range of focal lengths and f-stops. This limitation might lead to overfitting to specific characteristics of the training data.

V. Conclusions and Future Work

This paper presents a deep learning model based on the ResNet architecture for inferring missing metadata in photographs captured with manual lenses. The model demonstrates promising results in estimating both focal length and f-stop, which suggests that it might be using an internal representation of pupil size. A correlation analysis between inferred f-stop and actual pupil size showed a moderate correlation, indicating that the model is capturing features related to pupil size.

Despite the current limitations, this study provides a valuable foundation for future research in metadata inference and computational photography. The insights gained from this study highlight the importance of a diverse and high-quality dataset and the need for further exploration of deep learning architectures and training techniques to improve performance.

To address the limitations identified in this study, future work can lead to more accurate and reliable metadata inference systems, benefiting photographers and researchers in various applications, by focusing on:

1) **Dataset Expansion:** Collecting a larger and more diverse dataset with a balanced distribution of focal lengths, f-stops, and subject distances. This will help the model generalize better and capture the full range of relationships between these metadata, as well as allow for more comprehensive testing.
2) **Model Refinement:** Exploring different deep learning architectures and training techniques to improve the model's performance on focal length and subject distance estimation.
3) **Cross-Validation:** Using cross-validation techniques to evaluate the model's performance on different subsets of the dataset and identify any overfitting.
4) **Feature Analysis:** Conducting a detailed analysis of the features used by the model to infer f-stop and focal length to gain deeper insights into the model's behavior.

Acknowledgment

Author Contributions: Luis Alfredo da Silva designed the experimental methodology, created and augmented the dataset, provided computational resources, implemented the ResNet-based architecture, conducted training/validation/testing, and prepared the initial manuscript. Rafael Stubs Parpinelli provided research supervision, contributed to conceptualization and methodology validation, and performed critical manuscript revisions. Both authors analyzed the results and approved the final version.

To promote transparency and reproducibility, the complete source code for the training, testing, and result analysis, as presented in this paper, is available on the author's GitHub [https://github.com/Zidrewndacht/lens-metadata-inference].

This paper, and the associated source code and model, were developed with the assistance of the following open-weight Large Language Models:

Qwen2.5-Coder-32B by Alibaba Cloud and **DeepSeek V3-0324** helped with writing and debugging the source code for training, testing, and using the model.

Qwen2.5-72B by Alibaba Cloud and **Mistral Large 3 2411** by Mistral AI provided guidance in structuring and deciding on the contents of the paper.

Aya-Expanse-32B by Cohere contributed to translation and text quality assurance.

References

[1] Smith, W. J. (2008). Modern Optical Engineering.

[2] Hecht, E. (2017). Optics.

[3] Cira, C. I., Díaz-Álvarez, A., Serradilla, F., & Manso-Callejo, M.-Á. (2023). Convolutional Neural Networks Adapted for Regression Tasks: Predicting the Orientation of Straight Arrows on Marked Road Pavement Using Deep Learning and Rectified Orthophotography. Electronics, 12(18), 3980. https://doi.org/10.3390/electronics12183980

[4] Chen, D., Hu, F., Nian, G., & Yang, T. (2020). Deep Residual Learning for Nonlinear Regression. Entropy, 22(2), 193. https://doi.org/10.3390/e22020193

[5] Mohammedkhan, H., Fleuren, H., Güven, Ç., & Postma, E. (2025). Inferring Body Measurements from 2D Images: A Comprehensive Review. Journal of Imaging, 11(6), 205. https://doi.org/10.3390/jimaging11060205

[6] K. He, X. Zhang, S. Ren and J. Sun, ”Deep Residual Learning for Image Recognition,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, https://doi.org/10.1109/CVPR.2016.90.

[7] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv preprint arXiv:1905.11946. https://doi.org/10.48550/arXiv.1905.11946

[8] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. Chen, ”MobileNetV2: Inverted Residuals and Linear Bottlenecks,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018, pp. 4510-4520, doi: 10.1109/CVPR.2018.00474.

[9] A. Howard et al., ”Searching for MobileNetV3,” 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 1314-1324, doi: 10.1109/ICCV.2019.00140.

[10] Qin, D. et al. (2025). MobileNetV4: Universal Models for the Mobile Ecosystem. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds) Computer Vision – ECCV 2024. ECCV 2024. Lecture Notes in Computer Science, vol 15098. Springer, Cham. https://doi.org/10.1007/978-3-031-73661-2_5

[11] K. Han et al., ”A Survey on Vision Transformer,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 1, pp. 87-110, 1 Jan. 2023, https://doi.org/10.1109/TPAMI.2022.3152247.