Facial Expression Recognition Using Light Field Cameras: A Comparative Study of Deep Learning Architectures

Sabrine Djedjiga Oucherif, Mohamad Motasem Nawaf, Jean-Marc Boï, Lionel Nicod, Djamal Merad, Séverine Dubuisson

FACIAL EXPRESSION RECOGNITION USING LIGHT FIELD CAMERAS: A COMPARATIVE STUDY OF DEEP LEARNING ARCHITECTURES

Sabrine Djedjiga Oucherif [1] *Mohamad Motasem Nawaf* [2] *Jean-Marc Boï* [2]
Lionel Nicod [3] *Djamal Merad* [2] *Séverine Dubuisson* [2]

Aix-Marseille University, CNRS, [1]IMM, [2]LIS, [3]CERGAM, Marseille, France

ABSTRACT

This paper presents our contribution to facial expression recognition using images obtained from the Light Field Face Dataset(*LF*). We compare several variants of neural network architectures to demonstrate the potential benefits of using this relatively new optical system in the field of facial expression recognition. We propose the use of the EfficientNetV2-S convolutional neural network as the base architecture, combined with various recurrent neural networks (LSTM, GRU, BiLSTM, and BiGRU) in our experiments. Furthermore, we investigate different sets of sub-aperture images, each varying in terms of the number of images and virtual position. The results demonstrate a significant improvement in accuracy for two specific configurations, depending on the sets of sub-aperture images used. The first configuration involves using the EfficientNetV2-S model in a two-branch configuration combined with an LSTM. The second configuration uses a single branch model with a BiLSTM.

Index Terms— Facial Expression Recognition, Light Field Camera, Convolutional Neural Networks, Recurrent Neural Networks.

1. INTRODUCTION

The Light Field (*LF*) camera, also called plenoptic camera, is an optical system that allows to capture the intensity and direction of light rays[1], thanks to a micro-lenses array placed in front of the image sensor. Therefore, with a single camera shot we can capture several images, called sub-aperture images, representing the same scene with different points of view. Using LF imaging system, we can also obtain a depth map [2] and a super-resolution image [3] as illustrated in Fig. 1. Hence, it provides 3D information that standard digital cameras lack, and is used in different fields of computer vision applications ($3D$ reconstruction [4], $3D$ robotics [5], detection [6], classification and recognition [7]).

In this paper, we exploit sub-aperture images captured by the LF camera for human facial expression recognition. These images provide us three-dimensional information about facial structure and depth, allowing us to obtain insights into facial geometry and variations in expressions. Furthermore, we can leverage post-capture reconstruction techniques to emphasize specific facial regions. flexibility afforded by the LF camera enables the detection of subtle nuances in facial expressions that may go undetected by standard cameras.

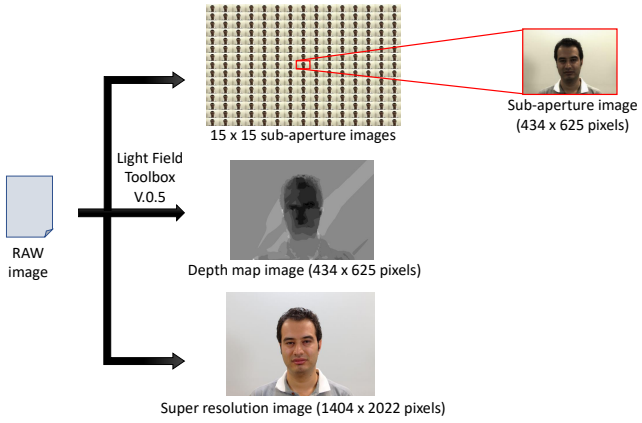

Fig. 1. Information returned by the light field camera from the LFFD Dataset [12].

To this end, we sequentially use (1) a Convolutional Neural Network (CNN) to extract relevant spatial features and (2) a Recurrent Neural Network (RNN) to extract the angular features. In this context, we compare combinations of a CNN (EfficientNetV2-S [8]) and an RNN (LSTM [9], GRU [10], Bidirectional LSTM and Bidirectional GRU [11]), with different sets of sub-aperture images, in terms of facial expression recognition accuracy. This paper is organised as follows: we first present in Section 2 some works on facial expression detection and recognition using the LF images. Then, in Section 3, we introduce our methodology, and in Section 4 we analyse the results obtained with different approaches, and highlight the two most relevant approaches. Finally, we give concluding remarks and perspectives in Section 5.

2. RELATED WORK

For the past four decades, the problem of facial expression recognition has been widely studied in the scientific community [13]. However, only a few studies have addressed this

problem using LF cameras data.

Shen *et al.* [14] used their own database to get depth maps, from which they extracted features using a Histogram Oriented Gradient (HOG). Using an SVM, they classify the facial expressions. As the LF cameras provide several images of different viewpoints from the same scene, Sepas-Moghaddam *et al.* proposed to extract two types of features' information from these views: spatial features and angular features. The former, also called intra-view, can be obtained by using a CNN to extract features from an individual sub-aperture image, and then comparing each of these characteristics with the neighboring attributes. [15]. Angular features, also called inter-view, are obtained by using an RNN to extract information from various sub-aperture images with the same position. This approach captures the relationship information among the images, which in turn represents implicit depth information. [16]. In [17], the authors proposed a combined CNN with a capsul network, and in [18, 19], a pretrained VGG16 on VGG Face Dataset [20] is combined with a bidirectional LSTM. The authors have demonstrated that exploiting multiple views, rather than single views, improve accuracy in emotion recognition. They also demonstrated that adding an attention mechanism learning layer to an RNN increases the accuracy.

New architectures of DL help to improve the precision for facial expression recognition. In the literature, a popular CNN is VGG16 [21]. However, in the past few years, more accurate models have emerged, such as EfficienNet, CoCa [22] and Model Soups [23]. With the development of Deep Learning (DL), new architectures of CNN have emerged, such as EfficientNetV2, which involve fewer parameters and can be trained faster than VGG16. Concerning RNN, we can cite the GRU [10] and LSTM [9] models: both have similar architecture except GRU that has only two gates and no output gate, and then necessitates fewer parameters than LSTM. In the next section, we present our combinations of architecture: EfficientNetV2-S as CNN with different kinds of RNN for facial expression recognition. We also analyse the impact of the used sub-aperture images set, and the various selection strategies, on the classification accuracy.

3. PROPOSED APPROACH

In this section, we propose various combinations using EfficientNetV2-S for facial expression recognition on LF system data with different kinds of RNNs, tested on different sets of sub-aperture images. We chose to use EfficientNetV2-S because it requires fewer parameters, is faster, and performs well on low-resolution images.

3.1. Dataset

The IST-EURECOM Light Field Database (LFFD) [12] is the only dataset publicly available for facial and emotion recognition. It contains raw images obtained with a Lytro Illum camera, $2D$ rendered images, and depth maps.

It is composed of two subsets: `session1` and `session2`, each one contains facial images taken from 100 subjects with three kinds of expression (angry, happy and surprised), neutral images, but also actions, poses, occlusions and illumination images. Images of `session2` show the same subjects of `session1`, but with a temporal delay of 1 to 6 months.

In our study, we only rely on sub-aperture images extracted from the raw data representing three facial expressions and the neutral one. The depth maps from the LFFD are not used due to the absence of face-centered calibration and their inability to provide comprehensive information about the facial region.

3.2. Pre-Processing

To extract sub-aperture images from the raw data, we use the Light Field Toolbox V. 0.5 [24]. For each raw image, we obtain a set of 15x15 images (also called mosaic or matrix) representing the same scene with different viewpoints (see Fig. 1). Each sub-aperture image has a resolution of 434×625 pixels. The images are then cropped and scaled down to obtain $60 \times 60 \times 3$ pixel images. This resolution was chosen according to a compromise between the available GPU memory (24 GB), and the capability of EfficientNetV2-S without a loss of performance.

3.3. Sub-aperture Image Sets

One of our objectives is to compare the accuracy of facial expression recognition depending on the selected sub-aperture images. As we mentioned, for each subject, on each session, we have a mosaic of 15×15 sub-aperture images. Using all sub-aperture images for facial expression recognition is not only computationally expensive, but it is also not necessary because of the small variation of the angular information between two neighbouring images. We rather propose to select sub-aperture images with a large variation between them. Hence, we will only consider subsets of the 15×15 image mosaic, as detailed in the following:

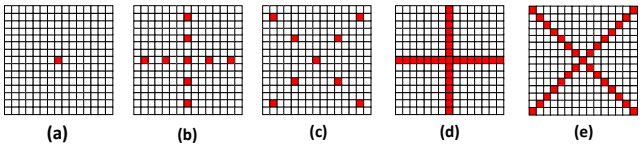

Fig. 2. Subsets of sub-aperture images. (a) Single image (b) 5 vertical and 5 horizontal images (c) 5 upward and 5 downward diagonal images (e) 15 vertical and 15 horizontal images (f) 15 upward and 15 downward diagonal images.

- *Single:* only the image located in the center (Fig. 3.(a)).

- *5 vertical and 5 horizontal:* five images from the middle column and the middle row, with a step of 3 images (Fig. 3.(b)).
- *5 upward and 5 downward diagonal:* five images from the upward and downward diagonals, with a step of 3 images (Fig. 3.(c)).
- *15 vertical and 15 horizontal:* all images from middle column and row (Fig. 3.(d)).
- *15 upward and 15 downward diagonal:* all images from the upward and downward diagonal (Fig. 3.(e)).

In our experiments, we will evaluate each sub-aperture image set for facial expression recognition. In particular, we will compare vertical/horizontal and the upward/downward diagonal subsets.

3.4. Deep Architecture

In this study, we compare 5 different deep learning architectures for facial expression recognition. Some are solely composed of a CNN (EfficientNetV2-S pre-trained with ImageNet), while others are a combination of the aforementioned CNN and an RNN. Note that we have chosen EfficientNetV2-S as CNN since it is more compatible with the resolution of images in our tests, compared to EfficientNetV2-M or EfficientNetV2-L. Each architecture will proceed with the sub-aperture images selection presented in Section 2. Fig. 3 shows the different architectures used in our experiments, and whose features are as follows:

- *Simple EfficientNetV2-S:* only one EfficientNetV2-S model is used, followed by two sequences of dense, batch normalization and dropout layers to prevent the overfitting. A softmax is used at the end for classification (Fig. 3.(a)).
- *Double branch of EfficientNetV2-S:* two EfficientNetV2-S models are used, one for the horizontal line or upward diagonal, and another for the vertical line or downward diagonal. Each CNN will be composed of a dense layer followed by batch normalization and dropout layers. We add an attention learning layer, to improve the results, then dense, batch normalization and dropout layers. Each branch will end with a softmax. It will be regrouped with a fusion layer to get the average results (Fig. 3.(b)).
- *Double branch of EfficientNetV2-S with RNN:* an RNN is added after EfficientNetV2-S and the first dense layer. In the scope of our work, we will compare LSTM, GRU, BiLSTM and BiGRU (Fig. 3.(c)). Using a bidirectional RNN exploit backward and forward information of the input sequences.
- *Single branch of EfficientNetV2-S:* same as the double EfficientNetV2-S model, except we only use one branch instead of two. The input regroups horizontal line/upward diagonal, and vertical line/downward diagonal images (Fig. 3.(b)).
- *Single branch of EfficientNetV2-S with RNN:* same as the double branch of EfficientNetV2-S with RNN model, except that we only have one branch regrouping all images (Fig. 3.(c)).

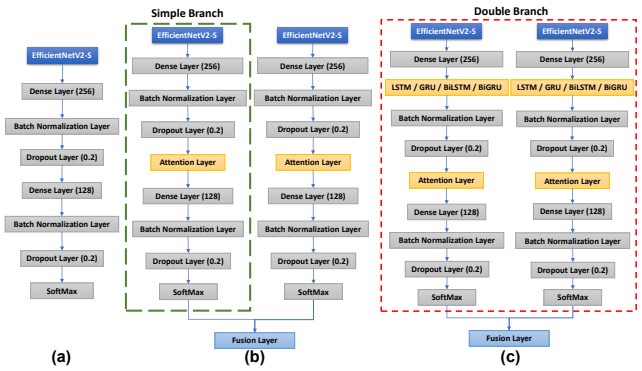

Fig. 3. Tested architectures. (a) Simple EfficientNetV2-S model (b) Simple and double branch of EfficientNetV2-S model (c) Simple and double branch of EfficientNetV2-S with RNN model.

4. EXPERIMENTS AND RESULTS

In this section, we present our experimental protocol, the chosen hyper-parameters, and the results obtained for the different combinations of sub-aperture images subsets and the architectures presented in the previous sections.

4.1. Experimental Protocol

To compare the performance of our models with the different sub-aperture images, we define a protocol which uses `session1` for training and `session2` for testing (see Section 3.1). We repeat the process by interchanging between the two sessions and average the results.

4.2. Hyper-Parameters

For all the sub-aperture image subsets, the resolution is $60 \times 60 \times 3$ pixels. The batch size and the epochs are respectively fixed to 45 images and 100. An early stopping is added, to avoid overfitting, by setting the patience at 10 and saving the best weights for validation accuracy.

4.3. Performance Analysis

Table 1 gives the results obtained with different kinds of architectures and sub-aperture images subsets described in the previous sections. First, we observed that our EfficientNetV2-S model for a single sub-aperture image gives a better average accuracy (81.75%) than the VGG16-EmotiW model (75.5%), the VGG19-PAM model (78.75%) or the AlexNet-PAM model (78.37%) of Sepas-Moghaddam *et al.* [18]. Furthermore, it is important to mention that a better accuracy is achieved by our double branch of EfficientNeV2-S with LSTM (82.88%). Next, we analyse and compared the different features of our approaches.

Table 1. Performance of different architectures using sub-aperture images from LFFD dataset for facial expression recognition.

Architecture	Sub-Aperture image subsets	Proposed Methods	Angry(%)	Happy(%)	Neutral(%)	Surprised(%)	Average(%)	std (%)
Simple Architecture	**Single**	EfficientNetV2-S	76.5	93.5	83.5	73.5	**81.75**	8.88
Double Branch Architectures	**5 vertical and 5 horizontal**	EfficientNetV2-S	69	89.5	76.5	72.5	76.88	8.96
		EfficientNetV2-S + LSTM	72.5	92.5	81	**77**	**80.75**	8.57
		EfficientNetV2-S + BiLSTM	73.5	89	**81.5**	68	78	9.19
		EfficientNetV2-S + GRU	**75**	85.5	76.5	74.5	77.88	5.15
		EfficientNetV2-S + BiGRU	72.5	**93**	75	64	76.13	12.2
	5 upward and 5 downward diagonal	EfficientNetV2-S	70.5	85	78	80	78.38	6.02
		EfficientNetV2-S + LSTM	72	**90**	75	76.5	78.38	7.97
		EfficientNetV2-S + BiLSTM	74.5	87.5	**82.5**	77.5	**80.5**	5.72
		EfficientNetV2-S + GRU	**78**	87	65	79	77.25	9.11
		EfficientNetV2-S + BiGRU	75	89.5	72.5	**80.5**	79.38	7.53
	15 vertical and 15 horizontal	EfficientNetV2-S	70.5	86	77	80	78.38	6.45
		EfficientNetV2-S + LSTM	**83.5**	88	71	72	78.63	8.44
		EfficientNetV2-S + BiLSTM	82.5	91.5	75.5	76.5	**81.5**	7.35
		EfficientNetV2-S + GRU	76.5	88	**81**	78	77.25	9.11
		EfficientNetV2-S + BiGRU	75	**94.5**	66.5	**81.5**	79.38	11.81
	15 upward and 15 downward diagonal	EfficientNetV2-S	77.5	90.5	77	77	80.5	6.67
		EfficientNetV2-S + LSTM	**80**	**92.5**	78.5	**80.5**	**82.88**	6.4
		EfficientNetV2-S + BiLSTM	78.5	88	81	73	80,13	6.22
		EfficientNetV2-S + GRU	75.5	88	**82.5**	80	81.5	5.21
		EfficientNetV2-S + BiGRU	75.5	89	75.5	80	80	6.36
Simple Branch Architectures	**5 vertical and 5 horizontal**	EfficientNetV2-S	74.5	85.5	**80.5**	72.5	78.25	5.91
		EfficientNetV2-S + LSTM	78	91	74.5	78	80.38	7.27
		EfficientNetV2-S + BiLSTM	76	**90.5**	79	**82**	**81.88**	6.25
		EfficientNetV2-S + GRU	**84**	88.5	73	78	80.88	6.79
		EfficientNetV2-S + BiGRU	72	88	74.5	76.5	77.75	7.08
	5 vertical and 5 horizontal	EfficientNetV2-S	73.5	89	72	83	79.38	8.06
		EfficientNetV2-S + LSTM	79	89.5	67	**84**	79.88	9.59
		EfficientNetV2-S + BiLSTM	**82**	**95**	73	71.5	80.38	10.8
		EfficientNetV2-S + GRU	70.5	90	**85**	78	**80.88**	8.49
		EfficientNetV2-S + BiGRU	70.5	87.5	80	76	78.5	7.15
	15 vertical and 15 horizontal	EfficientNetV2-S	77	90.5	76.5	76.5	80.13	6.92
		EfficientNetV2-S + LSTM	74.5	90.5	**82.5**	69.5	79.25	9.22
		EfficientNetV2-S + BiLSTM	78.5	90.5	82	79	**82.5**	5.55
		EfficientNetV2-S + GRU	77	**96.5**	72.5	77.5	80.88	8.49
		EfficientNetV2-S + BiGRU	**80**	91.5	76.50	**78**	81.5	6.82
	15 upward and 15 downward diagonal	EfficientNetV2-S	**81**	86	74.5	76.5	79.5	5.12
		EfficientNetV2-S + LSTM	70.5	**90.5**	80	79.5	80.13	8.18
		EfficientNetV2-S + BiLSTM	69	90	80	**80.5**	79.88	8.59
		EfficientNetV2-S + GRU	**81**	90	**84.5**	71.5	**81.75**	7.77
		EfficientNetV2-S + BiGRU	79.5	89	75.5	80	81	5.7

- Single branch versus double branch of EfficientNetV2-S with RNN: although the best results are obtained with a double branch, results also show that using double branch of neural networks does not provide more information. Using vertical/upward diagonal and horizontal/downward diagonal sub-aperture images as two separate inputs is not necessary.

- LSTM versus GRU: Using the double branch configuration of EfficientNetV2-S combined with LSTM yields higher recognition accuracy compared to GRU, as LSTM excels at capturing long-term dependencies in sequential data. However, in a single branch configuration of CNN-RNN, GRU outperforms LSTM due to its ability to maintain a more focused memory of information flow.

- RNN versus Bidirectional RNN: the comparison between RNN and bidirectional RNN is difficult. For 30 images as input, a double EfficientNetV2-S with LSTM performed better than all the other methods. But, the single branch of EfficientNetV2-S with the BiLSTM model also achieved a better accuracy ($82.5\% \pm 5.55\%$).

- Vertical/horizontal versus upward/downward diagonal images: the results obtained with diagonals are mostly better than those obtained with vertical/horizontal images. This comes from the large disparity between the images obtained with the light field camera.

- 10 versus 30 images: using 10 images as input is better than using a single image considering the standard deviation. For example, the double branch of EfficientNetV2-S with BiLSTM has 80.5% and 81.5% of accuracy with a respective standard deviation of 5.72% and 6.4%. It means that this model has less disparity, therefore more stable than the single architecture. However, using 30 images instead gives higher results. Indeed, The double branch model of EfficientNetV2-S with LSTM ($82.88\% \pm 6.47\%$) gives the best score in this paper.

Some images representing the emotions Angry, Neutral and Surprised do not have significant facial expression variations. That is why their recognition is poor compared to Happy.

5. CONCLUSION AND FUTURE WORK

In this paper, we introduced several deep learning architectures using LF images, and we compared their performances for facial expression recognition. The simple EfficientNetV2-S model with a sample 2D image achieves a better recognition than VGG16, VGG19 and AlexNet with $81.75\% \pm 8.88\%$ of accuracy. Using this model in two branches with diagonal images and an LSTM achieves the best result with $82.88\% \pm 6.47\%$ accuracy. Regrouping all images as input provides good results similar to EfficientNetV2-S with BiLSTM model which achieves $82.5\% \pm 5.55\%$ accuracy. In this context, we can affirm that using the LF system improves the performances of facial expression recognition.

6. REFERENCES

[1] G. Wu *et al.*, "Light field image processing: An overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 926–954, 2017.

[2] H. G. Jeon, J. Park, G. Choe, and J. Park, "Accurate depth map estimation from a lenslet light field camera," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1547–1555.

[3] Y. Wang, F. Liu, K. Zhang, G. Hou, Z. Sun, and T. Tan, "Lfnet: A novel bidirectional recurrent convolutional neural network for light-field image super-resolution," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4274–4286, 2018.

[4] J. Peng, Z. Xiong, Y. Zhang, D. Liu, and F. Wu, "Lf-fusion: Dense and accurate 3d reconstruction from light field images," in *IEEE Visual Communications and Image Processing (VCIP)*, 2017, pp. 1–4.

[5] D. Tsai, D. G. Dansereau, T. Peynot, and P. Corke, "Image-based visual servoing with light field cameras," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 912–919, 2017.

[6] Z. Ji, H. Zhu, and Q. Wang, "Lfhog: A discriminative descriptor for live face detection from light field image," in *IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 1474–1478.

[7] S. Wanner, C. Straehle, and B. Goldluecke, "Globally consistent multi-label assignment on the ray space of 4d light fields," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. pp. 1011–1018, 2013.

[8] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[9] F. Karim et al., "Lstm fully convolutional networks for time series classification," *IEEE access*, vol. 6, pp. 1662–1669, 2017.

[10] K. Cho *et al.*, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[11] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[12] A. Sepas-Moghaddam *et al.*, "The ist-eurecom light field face database," in *5th International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 2017, pp. 1–6.

[13] H. Ge *et al.*, "Facial expression recognition based on deep learning," *Computer Methods and Programs in Biomedicine*, vol. 215, pp. 106621, 2022.

[14] T. W. Shen *et al.*, "Facial expression recognition using depth map estimation of light field camera," in *IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, 2016, pp. 1–4.

[15] A. Sepas-Moghaddam *et al.*, "A double-deep spatio-angular learning framework for light field-based face recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4496–4512, 2019.

[16] B. Girod et al., "Light field compression using disparity-compensated lifting," vol. 4, pp. IV–760, 2003.

[17] A. Sepas-Moghaddam*et al.*, "Capsfield: Light field-based face and expression recognition in the wild using capsule routing," *IEEE Transactions on Image Processing*, vol. 30, pp. 2627–2642, 2021.

[18] A. Sepas-Moghaddam *et al.*, "Facial emotion recognition using light field images with deep attention-based bidirectional lstm," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3367–3371.

[19] A. Sepas-Moghaddam *et al.*, "A deep framework for facial emotion recognition using light field images," in *8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 1–7.

[20] Q. Cao et al., "Vggface2: A dataset for recognising faces across pose and age," in *13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[21] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun, Eds., 2015.

[22] Jiahui Yu and Zirui *et al.* Wang, "Coca: Contrastive captioners are image-text foundation models," *arXiv preprint arXiv:2205.01917*, 2022.

[23] M. Wortsman et al., "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," in *International Conference on Machine Learning*. PMLR, 2022, pp. 23965–23998.

[24] "D, dansereau, light field toolbox v. 05," https://dgd.vision/Tools/LFToolbox/, Accessed: 2023-04-02.