CAVLI - Using image associations to produce local concept-based explanations

Pushkar Shukla
Toyota Technological Institute at Chicago
pushkarshukla@ttic.edu

Sushil Bharati
Teladoc Health
SBharati@teladoc.com

Matthew Turk
Toyota Technological Institute at Chicago
mturk@ttic.edu

Abstract

*While explainability is becoming increasingly crucial in computer vision and machine learning, producing explanations that can link decisions made by deep neural networks to concepts that are easily understood by humans still remains a challenge. To address this challenge, we propose a framework that produces local concept-based explanations for the classification decisions made by a deep neural network. Our framework is based on the intuition that if there is a high overlap between the regions of the image that are associated with a human-defined concept and regions of the image that are useful for decision-making, then the decision is highly dependent on the concept. Our proposed CAVLI framework combines a global approach (**TCAV**) with a local approach (**LIME**). To test the effectiveness of the approach, we conducted experiments on both the ImageNet and CelebA datasets. These experiments validate the ability of our framework to quantify the dependence of individual decisions on predefined concepts. By providing local concept-based explanations, our framework has the potential to improve the transparency and interpretability of deep neural networks in a variety of applications.*

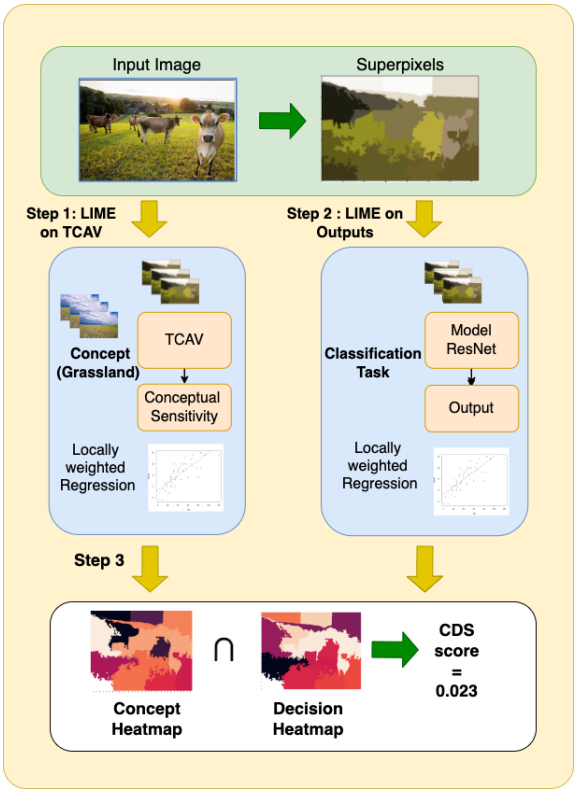

Figure 1. Overview of our proposed approach to estimate the dependence of a concept (e.g., "grassland") on a decision (e.g., "buffalo detection"). After decomposing the input image into superpixels, in Step 1 we find the regions of the image that have the highest association with the concept, defined by a set of images. In Step 2 we identify image regions with the highest involvement in the classification decision. Finally, we measure the overlap between the two in order to quantify the dependence.

1. Introduction

As the capacity and use cases of deep neural networks continue to expand, the need for interpreting and explaining decisions made by these systems also increases. Understanding and reasoning about these networks can be critical in settings with high risk and high impact. Furthermore, improved interpretations and explanations of decisions made by deep neural networks enhance their trustworthiness and enable greater opportunities for human intervention in case a mistake has been made.

Most explainability techniques can be classified into local and global methods. Local methods focus on explaining decisions for individual inputs and aim to identify the factors involved in the decision. These methods employ heatmaps [6, 25, 26], counterfactuals [1, 3, 5, 7, 10, 16, 21], or feature perturbation [20, 22] methods to explain model decisions. However, each of these methods has its own limitations. In contrast, global [14, 15, 18] methods explain the model, independent of decisions on inputs, and include methods that attempt to understand the model's decision-making dependence on human-defined concepts or interpret different layers of the model and their overall role in decision-making. Concept-based methods [8, 12, 13, 18, 24]

are popular global methods that try to quantify the dependence of decisions of a neural network into a set of human-defined concepts or label neurons or neuron combinations [4, 11] with semantic attributes. A concept-based method tries to establish a relationship between decisions made by a machine learning model and human defined notions (e.g., gender, race, patterns) that cannot be measured directly as inputs or outputs of a machine learning system.

In this paper we propose a novel framework for generating local human concept-based explanations that combine aspects of both global and local approaches. As shown in Figure 1, our proposed approach aims to quantify the impact of human concepts on the decision-making process of a neural network for a particular input. For example, given a decision made by a smile classifier, our aim is to estimate the dependence of the decision on human-defined concepts like gender, race, eyeglasses, etc. The high-level architecture of our approach consists of three main steps. In Step 1, we identify the regions of an image that have the highest association with a given human concept. We use a heatmap to highlight the regions that are most relevant to the concept of interest. In Step 2, we identify the image regions that have the highest involvement in the decision-making process of the neural network for a given input image. To achieve this, we use attribution methods to compute the contribution of each image region to the final decision. Finally, in Step 3, we measure the overlap between the image regions identified in Steps 1 and 2. This overlap gives us a measure of the degree to which the decision of the neural network is dependent on the human concept of interest. Our framework is based on the insight that if a model's decision depends on a concept, then there should be a high overlap in regions that are used by the model for decision-making and regions of the image that is used for concept modelling. For example, if parts of the image that are used for classifying whether an animal is a zebra or not are also associated with the concept of stripes, then the decision `zebra` depends on the concept `stripes`.

We propose CAVLI, a method that aims to explain local decisions made by a deep neural network regarding human concepts by combining two well-known methods, **TCAV** [18] and **LI**ME [22]. Even though both TCAV and LIME are approaches used for model explainability, they differ widely in their usage. TCAV is a global technique that explains how well the model understands human concepts. It requires an understanding of model layers and weights and uses extensive images for training concepts. LIME is model agnostic, treats the model as a black box, and works on individual inputs. Merging these two approaches leads to a unique method capable of local concept-based explainability. Overall our major contributions can be summarized as follows.

- We propose a novel approach for building local concept-based explanation models that focuses on understanding the overlap between image pixels involved in decision-making and image pixels that are related with a concept.
- Through our framework we provide both a quantitative explanation in terms of a *Concept Dependency Score (CDS)* and a visual explanation using concept heatmaps that indicate the dependence of the model on a given concept.
- We perform qualitative and quantitative experiments on the ImageNet dataset [9, 23] and the CelebA [19] to validate the utility of our methods.

2. Methodology

2.1. Notation

Consider a trained neural network $F : X \rightarrow \{1, ...K\}$, on a dataset $X = \{\mathbf{x_1}, \mathbf{x_2}, ...\mathbf{x_t}\}$ and associated labels $Y = \{y_1, y_2, ...y_t\}$, where $y_i \in \{0, 1\}^K$ with K classes. $F_k(\mathbf{x_i}) := h_l(f_l(\mathbf{x_i}))$, where $f_l(\mathbf{x_i})$ are the output logits of the l^{th} layer and h_l is the activation function of the l^{th} layer.

2.2. TCAV and LIME

TCAV [18] uses human-defined *concepts* (e.g., "gender" or "stripes") instead of input features to provide explanations for a machine learning model. To express a concept it finds a Concept Activation Vector (CAV) $a \in \mathbb{R}^d$ (a layer with dimension d) in the network's activation space [24] that points in the direction of the concepts. This is achieved by training a classifier that distinguishes concept activations ("striped" or "dotted") from activations of negative samples and taking a unit norm vector $\mathbf{v_c}$ orthogonal to its decision boundary. The inner product in Equation 1. denotes the similarity of the activstion to the required concept and $\mathbf{v_c}$ denotes the direction of the concept vector. This is defined as the Conceptual Sensitivity CS of a given layer l for the network's output class k and the concept C:

$$CS^k_{C,l}(F, \mathbf{x_i}) = \nabla h_l(f_l(\mathbf{x_i}))^T \mathbf{v_c} \tag{1}$$

The TCAV score is given as the ratio of the number of inputs with positive conceptual sensitivity to the number of inputs for a class.

LIME [21, 22] is a black box method for understanding local explanations of a machine learning model. In order to explain the prediction of a model F on an image $\mathbf{x_i}$ it:

1. Decomposes $\mathbf{x_i}$ in r homogeneous image patches or superpixels.
2. Creates a set of new images $\mathbf{x_{i_1}}, ...\mathbf{x_{i_n}}$ by selecting n subsets of the superpixels
3. Queries the model for each of these images $y_{i_j} = F(x_{i_j}) \; \forall j \in \{1, 2...n\}$

4. Builds a local weighted surrogate model $\hat{\beta}_i$ fitting the y_{i_j}'s to the presence or absence of superpixels.

Each coefficient of $\hat{\beta}_i$ is associated with a superpixel of the original image $\mathbf{x_i}$. Intuitively, the more positive the value of the coefficient, the more important the superpixel is for the prediction of the model. Generally, the user visualizes $\hat{\beta}_i$ by highlighting the superpixels associated with the top positive coefficients.

2.3. Proposed Approach

We propose a hybrid TCAV-LIME-based approach as a solution to the problem. Algorithm 1 describes our proposed framework. First, we try to understand how well the model captures a concept for an individual decision. We frame the question by trying to understand *what parts of the image the model associates the most with a specific attribute*. Techniques like TCAV cannot be used directly because of their global nature. We start by dividing the image into r homogeneous superpixels using the SLIC [2] superpixel algorithm. The second part of our pipeline investigates the question, *for a given decision, what parts of the image were the most influential in making the decision?* We make use of LIME to generate these regions in the image (Steps 1, 2, 6, and 7). Finally, we measure the overlap between the two parts. The intuition behind the approach is that if there is a clear overlap between image patches that have a high dependency on a concept and image patches that have the highest weight in decision-making, then the decision made by the network is heavily dependent on the concept (Steps 8 and 9).

The coefficients of $\hat{\alpha}_i$ indicate the level of association between different superpixels and a particular concept. A higher coefficient value suggests that the model considers the concept to be more closely related to that region, and vice versa. Similarly, coefficients of $\hat{\beta}_i$ corresponds to a superpixel in the original image $\mathbf{x_i}$. The higher the weight of the superpixel, the more significant its contribution to the model's decision-making process. We are interested in measuring whether the superpixels associated with the the given concepts are also associated with decisions made by the algorithm. We calculate the Pearson correlation γ_i correlation of $\hat{\beta}_i$ and $\hat{\alpha}_i$ to measure the overlap between the two decisions. A larger value of γ_i indicates that there is a high overlap between the regions of the image that the model associates with the concept and those it uses for the decision. The Concept Dependency Score CDS_i, is calculated as the product of γ_i and $CS^k_{C,l}(F, x_i)$, ensuring that relevant concepts are given higher values. For a qualitative understanding, the coefficients of $\hat{\alpha}_i$ associated with the superpixels can be represented as a concept heatmap. This heatmap gives us a visualization of what parts of the image are more likely to be associated with a concept.

Algorithm 1 CAVLI

1: Train a TCAV model for a given concept C, a model F, and a layer l, resulting in the CAV vector $\mathbf{v_c}$.
2: Decompose the input image $x_i \in X$ in a set of r homogeneous superpixels $\{S\}$.
3: Create a new set of images $\{\mathbf{x_{i_1}}, ..., \mathbf{x_{i_n}}\}$ from $\mathbf{x_i}$ by randomly masking parts of the image and selecting n uniformly sampled subsets of $\{S\}$
4: Calculate the Conceptual Sensitivities z_{i_j}=$CS^k_{C,l}(F, \mathbf{x_{i_j}})$ $\forall \mathrm{x}_{i_j} \in \{\mathbf{x_{i_1}}, ...\mathbf{x_{i_n}}\}$.
5: Build a local weighted surrogate model $\hat{\alpha}_i$ fitting the z_{i_j}'s to the presence or absence of superpixels.
6: Query the model for each of these image patches $y_{i_j} = F(x_{i_j})$ $\forall \mathrm{x}_{i_j} \in \{\mathbf{x_{i_1}}, ..., \mathbf{x_{i_n}}\}$.
7: Build a local weighted surrogate model $\hat{\beta}_i$ fitting the y_{i_j}'s to the presence or absence of superpixels.
8: Calculate the Pearson correlation γ_i between the coefficients of $\hat{\alpha}_i$ and $\hat{\beta}_i$.
9: Calculate the Concept Dependency Score, CDS_i= $\gamma_i \cdot CS^k_{C,l}(F, x_i)$

3. Evaluation

3.1. ImageNet Dataset

In our initial experiments, we assess the effectiveness of CDS in explaining model decisions in terms of concepts. To establish a baseline, we compare the performance of TCAV with the average CDS scores across different samples. We propose a hypothesis that if there exists a correlation between the mean CDS scores and global concept methods like TCAV, it indicates that our metric is capable of accurately capturing the dependence between the model decisions and underlying concepts.We conducted experiments on the ImageNet dataset, using similar settings to Kim et al. [17] and Schrouff et al. [24] to validate our model. Specifically, we focused on the Zebra and basketball classes, using three different models (GoogleNet, ResNet-50, and InceptionNet) for each class. Our goal was to measure the average statistics for each class using 100 images per set and calculating the mean correlation across all CDS scores. The experiments were conducted on the penultimate layer of all models.

Zebra. We ran experiments similar to Schrouff et al. [24] that focus on four different concepts: "stripes," "zigzagged," "dotted," "horse," and "grasslands." The results of the experiments are presented in Table 1, which shows that ResNet and GoogleNet both exhibited the highest mean CDS score for the concept stripes and the lowest mean CDS score for the concept indoor within the Zebra class. For InceptionNet, the concept grassland was more strongly associated with the Zebra class. The TCAV scores, which serve as global indicators of concept dependency, followed a similar trend. This pattern suggests that, on average, the CDS scores resemble the TCAV score. It is worth

	Stripes		Grassland		Indoor		Horse	
Model	CDS	TCAV	CDS	TCAV	CDS	TCAV	CDS	TCAV
GoogleNet	0.17	0.78	0.26	0.62	0.13	0.12	0.02	0.41
ResNet	0.23	0.87	0.26	0.81	0.11	0.48	0.11	0.51
InceptionNet	0.45	0.84	0.21	0.71	-0.13	0.43	0.16	0.35

Table 1. A comparison of mean CDS values and TCAV values of different concepts for the class Zebra in the ImageNet dataset.

	Ball		Jersey		Female		Race	
Model	CDS	TCAV	CDS	TCAV	CDS	TCAV	CDS	TCAV
GoogleNet	0.29	0.56	0.38	0.93	-0.03	0.26	0.24	0.46
ResNet	0.27	0.68	0.21	0.46	-0.20	0.45	0.22	0.73
InceptionNet	0.41	0.87	0.05	0.31	0.09	0.31	0.18	0.57

Table 2. A comparison of mean CDS values and TCAV values of different concepts for the class basketball in the ImageNet dataset.

	Male	Female
Smile	0.004	0.013
Non-Smile	0.005	0.007

Table 3. Average CDS scores for different subgroups in the CelebA dataset.

noting that higher CDS values indicate greater dependency on a concept, while lower values indicate lower dependency on the concept.

Basketball. We examined four human concepts ("ball," "jersey," "gender," and "race") in a manner similar to the Zebra class. The results in Table 2 show higher mean CDS scores for "jersey" and "ball," and a lower score for "female." We trained a race concept classifier with positive class images of African American faces. Our results further confirm the previous findings of a correlation between decision made on the basketball class and concept race. [24].

3.2. CelebA dataset

We are interested in exploring whether our approach can detect biases in model decisions caused by unbalanced data. Through our experiments we are interested in measuring whether these confounds can be detected by our metric. The CelebA dataset [19] is known to have naturally occurring confounds. We train a smile classifier in a biased setting, where the training set is subsampled to create a higher positive correlation between the female-smiling and male-non-smiling attributes. We analyze the average CDS scores for different subgroups on the test data, as shown in Table 3. We observe that the highest average CDS scores were for the "female smiling" group, while the lowest were for the "male smiling" group. These experimental findings align with the existing biases present in the dataset.

3.3. Qualitative Analysis

Our method generates concept heatmaps that illustrate the image regions and their dependence on human-defined concepts. These heatmaps aid in visually interpreting a model's image dependency, as shown in Figure 2 for the Buffalo-grasslands class-concept pair. The qualitative

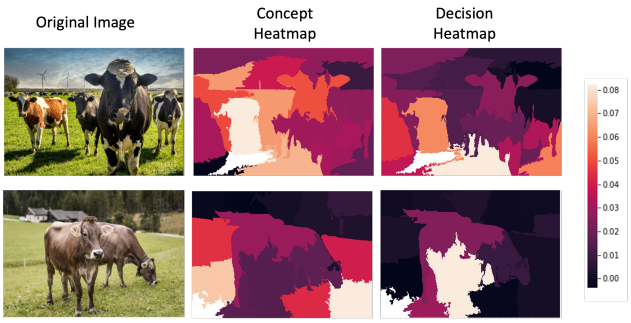

Figure 2. We use concept and decision heatmaps to analyze a classifier's decisions and their dependence on a specific concept, such as identifying whether an image contains a buffalo and what parts are the most useful in decision making. Here we focus on grasslands, and the concept heatmap displays the areas of the image that the model associates most strongly with this concept.

analysis reveals not only the parts of the image used for decision-making, but also whether the model links these parts to a concept. Additionally, this visual representation can identify spurious correlations where a classification decision (buffalo) is based on a concept (grassland) that is not directly related to the class.

4. Conclusion

This paper presents a new approach for generating local post-hoc explanations based on concepts. The three-step approach proposed here successfully utilizes a hybrid LIME and TCAV-based strategy to produce concept-based explanations. While there are limitations to this approach that need to be explored, this work represents a significant step towards developing more effective and reliable models for generating local explanations. Further research will be necessary to fully explore the potential of this approach, including more rigorous experimentation and user studies. Ultimately, however, the proposed method shows promise for improving our understanding of complex machine learning models and their decision-making processes.

References

[1] Abubakar Abid, Mert Yuksekgonul, and James Zou. Meaningfully debugging model mistakes using conceptual counterfactual explanations. In *International Conference on Machine Learning*, pages 66–88. PMLR, 2022. 1

[2] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012. 3

[3] Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. Towards causal benchmarking of biasin face analysis algorithms. In *Deep Learning-Based Face Analytics*, pages 327–359. Springer, 2021. 1

[4] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6541–6549, 2017. 2

[5] Chun-Hao Chang, George Alexandru Adam, and Anna Goldenberg. Towards robust classification model by counterfactual and invariant data generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15212–15221, 2021. 1

[6] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847. IEEE, 2018. 1

[7] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[8] Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020. 1

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 2

[10] Emily Denton, Ben Hutchinson, Margaret Mitchell, Timnit Gebru, and Andrew Zaldivar. Image counterfactual sensitivity analysis for detecting unintended bias. *arXiv preprint arXiv:1906.06439*, 2019. 1

[11] Ruth Fong and Andrea Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8730–8738, 2018. 2

[12] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[13] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*, 2019. 1

[14] Badr Youbi Idrissi, Diane Bouchacourt, Randall Balestriero, Ivan Evtimov, Caner Hazirbas, Nicolas Ballas, Pascal Vincent, Michal Drozdzal, David Lopez-Paz, and Mark Ibrahim. Imagenet-x: Understanding model mistakes with factor of variation annotations. *arXiv preprint arXiv:2211.01866*, 2022. 1

[15] Saachi Jain, Hannah Lawrence, Ankur Moitra, and Aleksander Madry. Distilling model failures as directions in latent space. *arXiv preprint arXiv:2206.14754*, 2022. 1

[16] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Diffusion models for counterfactual explanations. In *Proceedings of the Asian Conference on Computer Vision*, pages 858–876, 2022. 1

[17] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019. 3

[18] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pages 2668–2677. PMLR, 2018. 1, 2

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 2, 4

[20] Karim El Mokhtari, Ben Peachey Higdon, and Ayşe Başar. Interpreting financial time series with shap values. In *Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering*, pages 166–172, 2019. 1

[21] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020. 1, 2

[22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016. 1, 2

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015. 2

[24] Jessica Schrouff, Sebastien Baur, Shaobo Hou, Diana Mincu, Eric Loreaux, Ralph Blanes, James Wexler, Alan Karthikesalingam, and Been Kim. Best of both worlds: local and global explanations with human-understandable concepts. *arXiv preprint arXiv:2106.08641*, 2021. 1, 2, 3, 4

[25] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE In-*

ternational Conference on Computer Vision, pages 618–626, 2017. 1

[26] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 24–25, 2020. 1