Quantifying Explainability with Multi-Scale Gaussian Mixture Models

Anthony Rhodes
Intel Labs
anthony.rhodes@intel.com

Yali Bian
Intel Labs
yali.bian@intel.com

İlke Demir
Intel Labs
ilke.demir@intel.com

Abstract

With the increasing complexity and influence of machine learning models, the development of model explanation techniques has recently gained significant attention, giving rise to the field of Explainable Artificial Intelligence (XAI). Although there exists vast literature on XAI methods, they are usually compared with human evaluations, model-dependent metrics, or distribution shifts. In the present work, we introduce a novel explainability comparison metric, eXplainable Multi-Scale Gmm Distance (XMGD). XMGD provides a principled probabilistic framework for analyzing and quantifying any model or dataset similarity through the lens of explainability. Through experimental results, we demonstrate several critical advantages of XMGD over alternative saliency comparison metrics, including improved robustness and the ability of XMGD to illuminate fine-grain saliency comparison distinctions.

1. Introduction

As the responsible AI ecosystem matures, more tools are proposed for model interpretability, giving birth to the field of Explainable Artificial Intelligence. Despite the recent growth in XAI techniques, the utility of some XAI tools is nevertheless still sometimes opaque, and there are few algorithms today that adequately provide a means to objectively assess and analyze different modalities of explainability.

Within the domain of XAI, image-based explanations are evaluated and compared by standard image similarity metrics, by distribution similarity, by human evaluations, or by change in model accuracy when the input pixels are modified according to the explanation. Although these metrics may answer different use cases, they tend to depend on input samples, model selections, humans, or domains. In the current work, we introduce an explainability-based comparison metric, Explainable Multi-Scale GMM Distance (XMGD), that can be used to compare model saliencies in a statistically-principled way.

Following the guidelines provided by [4], we aim to provide a saliency evaluation metric that encompasses an "intuitive scale", i.e., one that utilizes only pixels and a small number of parameters, encapsulates both spatial and distributional understanding, and is sufficiently robust to saliency noise. Our method leverages flexible Gaussian Mixture Models (GMMs) to learn a high-order probability distribution in the input pixel space at multiple scales for each saliency comparison. We then calculate the 2-Wasserstein distance between these two GMMs [17] – one pair for each input scale – to quantify saliency similarity.

XMGD provides three distinct advantages over conventional visual-XAI metrics. First, XMGD is less sensitive to individual input/pixel saliency intensities and thus more robust as an explainable similarity measure than alternative metrics, because we fit a high-order distribution (*cf.* KL-Divergence which performs a discrete, pixel-level calculation). Additionally, XMGD is not sensitive to dataset size and it therefore does not suffer from poor convergence properties due to small datasets. XMGD metric also operates directly on saliency maps, without the need for manipulating the input or the model. Finally, XMGD can enhance explainability across a large number of diverse use cases, including saliency comparisons for individual images, entire datasets, or cross-model analysis.

We use XMGD to compare detector saliencies while aligning with expectations based on model similarities. We evaluate XMGD by comparing to other XAI metrics; in terms of diversity of distances, conformity to model-based metrics, and exposure of similar data and models.

2. Related Work

Today, there exists a large number of XAI techniques, spanning visualization tools such as saliency maps, counterfactual explanations, model uncertainty, feature attributions, and rule-based explanations. In the current work, we emphasize saliency methods, a widely adopted and valuable approach for interpreting neural networks.

Saliency maps depict the distribution of salient pixels in an input image, attention map, or visual representation. Saliency images can be compared in pixel space using techniques such as SSIM [31], IoU [19], and RMSE, or they can be compared in distribution space, e.g., KL divergence [15]

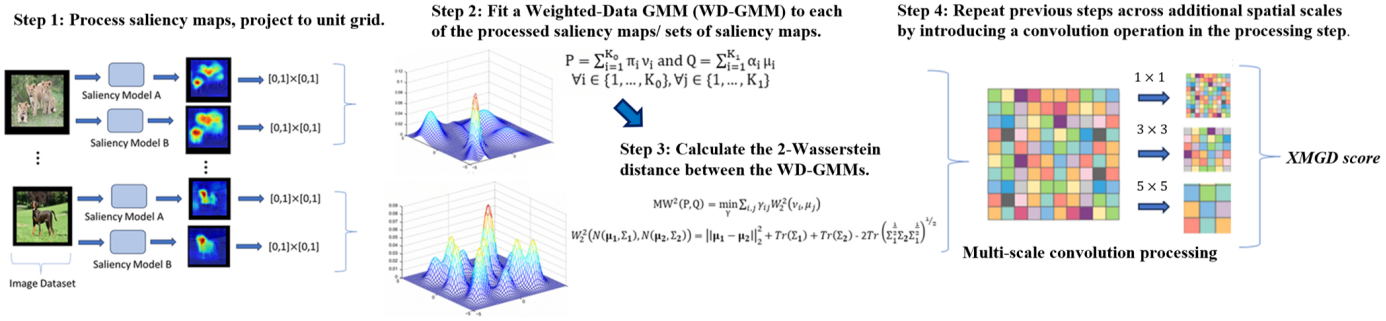

Figure 1. **XMGD Algorithm Schematic.** Different datasets or models are processed with different XAI methods. Then, a WD-GMM is fit to each set of saliency maps. Finally, the previous step is repeated for each spatial scale obtained by convolution.

or EMD [21]. These metrics, similar to image comparison, serve for specific comparisons in structure, area, pixel, and distribution; thus they do not provide comprehensive similarity independent of content, artifacts, or domain.

In addition, as their main use case is to understand model responses, they can also be evaluated with respect to the model they explain. Several metrics have been proposed to this end, including Insertion Correlation (IC) [10], Average Drop (AD) [5], Average Drop in Deletion (ADD) [14], and Insertion/Deletion Area Under Curve (IAuC/DAuC) [18]. Each of these methods rely on eliminating the salient or non-salient pixels in an input image and then measuring the induced change in model accuracy on the manipulated image. Although these metrics are useful for tying model explanations to models, their evaluation requires access to the underlying model; moreover, they are mostly sample and image size dependent, and their computational time is dependent on the model inference time.

3. Explainable Multi-Scale GMM Distance

Our method provides an XAI-forward metric to quantify the dis/similarity between different ML models and datasets. Given two input saliency maps (or aggregations of saliency maps) representing different models or datasets, XMGD can be proceeds in four steps (Fig. 1): (i) We preprocess the saliency maps and project them to the 2D unit grid. (ii) Next, we fit a Weighted-Data GMM (WD-GMM) to each of the two processed saliency maps. (iii) Then, we calculate the 2-Wasserstein distance between the WD-GMMs. (iv) Finally, to generate a fine-grain saliency comparison metric capturing spatial properties, we repeat the above steps across additional spatial scales by introducing a convolution operation. The final *XMGD score* consists of a weighted sum of these multi-scale distances.

3.1. XMGD Formulation

Given two sets of saliency images, or aggregations of saliency images where a mean, per-pixel saliency is calculated across the set of images, $S^I \supset s^i : i \leq |I|$ and $S^J \supset s^j : j \leq |J|$ we apply min-max normalization so that $S^I, S^J \in [0,1]^{w \times h}$.

Next, we train a data-weighted GMM [8] to fit each saliency image. We first transform each saliency image into a corresponding "dataset" in the input pixel space. Concretely, we convert each pixel in the saliency image to a unit grid: $[0,1] \times [0,1]$, and then determine each per-pixel normalized saliency value as a data weight. For example, if pixel $x_{mn} \in S^I$ has normalized saliency score S^I_{mn}, we map this weight to the corresponding pixel in the unit grid.

The formal weighted-GMM problem has been solved in [9], where the authors demonstrate that the solution for the weighted-GMM is equivalent to "duplicated point GMM", where one can simply duplicate data points in correspondence with data weights and solve using classical Expectation-Maximization. We duplicate data points in this fashion by introducing a tunable binning parameter b, so that $n_{ij} = \lfloor \frac{s_{ij}}{b} \rfloor$, where n_{ij} denotes the number of data point duplications for point x_{ij}. With our dataset corresponding to spatial saliency defined over the unit grid with duplicated points, we fit a GMM using the EM algorithm for each of the two input saliency images S^I and S^J, which we will denote as P and Q.

Then, given two GMMs,

$$P = \sum_{i=1}^{K_0} \pi_i \nu_i \text{ and } Q = \sum_{i=1}^{K_1} \alpha_i \mu_i \quad (1)$$
$$\forall i \in \{1, ..., K_0\}, \forall j \in \{1, ..., K_1\}$$

we calculate *mixed* 2-Wasserstein distance similar to [17]:

$$MW^2(P,Q) = \min_{\gamma} \sum_{i,j} \gamma_{ij} W_2^2(\nu_i, \mu_j) \quad (2)$$

where 2-Wasserstein distance is expressed in closed form:

$$W_2^2(N(\mu_1, \mathbf{\Sigma}_1), N(\mu_2, \mathbf{\Sigma}_2)) = ||\mu_1 - \mu_2||_2^2 + Tr(\mathbf{\Sigma}_1) + Tr(\mathbf{\Sigma}_2) - 2Tr\big((\mathbf{\Sigma}_1^{\frac{1}{2}} \mathbf{\Sigma}_2 \mathbf{\Sigma}_1^{\frac{1}{2}})^{\frac{1}{2}}\big) \quad (3)$$

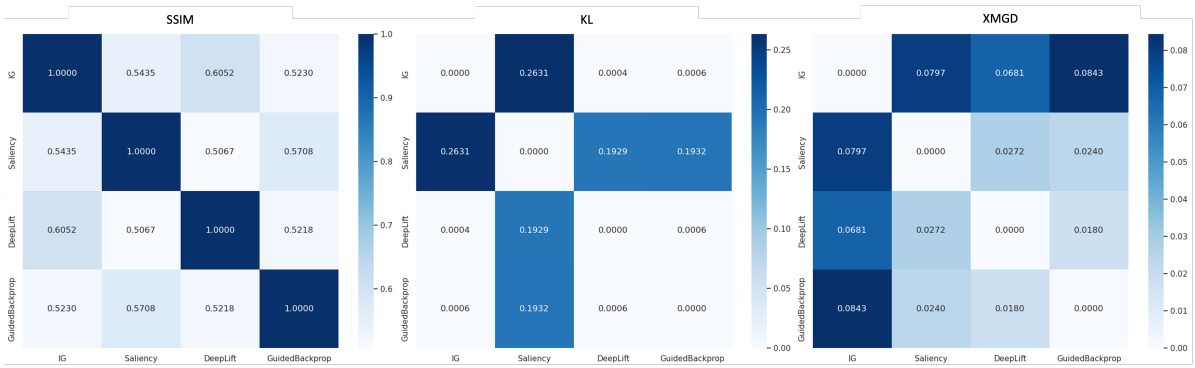

Figure 2. **Saliency Comparison**. Aggregated saliency maps over 7 detectors are compared by SSIM, KL, and XMGD (our metric).

3.2. Solving XMGD

In Eqn. 2, γ can be solved using discrete optimal transport [24]. Note that efficient Python solvers for both GMM fitting [30] and optimal transport [7] are available.In our case, one XMGD comparison takes approximately one minute using an Intel Xeon^{TM} processor for saliency images of size $(256, 256)$.

3.3. XMGD Parameterization

Lastly, in order to generate a more fine-grained saliency comparison metric capturing spatial properties, we compute $MW^2(P, Q)$ across l spatial scales, using $l-1$, $2D$ convolutions with kernel sizes $\{C_w \times C_w,\ C_{2\times w-1} \times C_{2\times w-1}, \dots\}$ on saliency maps, respectively. The final XMGD score is computed by $\frac{1}{l}\sum_l MW_l^2(P_l, Q_l)$ as the average of the distances from all levels. The number of levels l, sizes of convolution kernels w, number of bins b, and number of Gaussians for the GMM fit can be tunable per domain, dataset, and image size. For our domain, we set $l = 3$, so in addition to the original saliency images, we applied convolutions with kernel sizes 5×5 and 3×3 to the original saliency images when calculating XMGD.

4. Experimental Settings

For our evaluations, we employ the FaceForensics++ [20] (FF++) dataset containing 1000 real videos and 5000 corresponding deepfakes generated using DeepFakes [1] (DF), Face2Face [29] (F2F), Face Shifter [16] (FSh), Face Swap [2] (FSw), and Neural Textures [28] (NT). We create saliency maps using several widedly-adopted, gradient-based XAI methods: Saliency [23], Guided Backpropagation [25], Integrated Gradients [26], and DeepLIFT [22]. These saliency methods are employed across seven deepfake detectors, namely MesoNet [3], MesoInception4 [3], ResNet [11], Xception [6], Inception [27], MobileNet [12], and SqueezeNet [13], culminating in nearly 1 million saliency maps. To relate saliency maps back to the model explanations, we compute IC [10], AD [5], and ADD [14] metrics, with respect to these models. Finally, we use SSIM and KL Divergence as baseline saliency similarity metrics to compare against XMGD.

5. XMGD Evaluation

To evaluate XMGD, we compare its representativeness in terms of (i) diversity of distances, (ii) relevance to model-based metrics, and (iii) preserved correlations of detectors/generators. We also discuss its "free" benefits.

In Tab. 1, we compute model-dependent metrics to set a baseline expectation about metric performance. IC and AD removes non-salient pixels and computes the average increase and drop in accuracy. ADD removes salient pixels and computes the average drop in accuracy. Higher IC and ADD and lower AD defines a better saliency map. IG seems to perform the best for this domain on all three metrics.

	IG	Saliency	DeepLift	GuidedBP
IIC↑	**0.612**	0.358	0.561	0.518
AD↓	**-0.082**	0.344	0.289	0.229
ADD↑	**1.064**	-0.070	0.201	-0.120

Table 1. **Model-dependent Saliency**. We compare IIC, AD, and ADD on four saliency methods, where IG performs the best.

Fig. 2 shows experimental results for generated saliencies aggregated over the afmorementioned detector models; we then pairwise compare the similarity of these aggregated saliency maps across the four XAI methods. When compared with baseline saliency comparison methods KL Divergence and SSIM, XMGD demonstrates the most differentiated levels of distances between distributions, supporting (i) (also observed in the first row of Fig. 3). Comparing Tab. 1 and Fig. 2 for (ii), we observe similar differentiation of IG, by both XMGD and three model-based metrics. This observation is critical, as it demonstrates that XMGD can capture model-based insights *without* the need for input manipulation and additional inference.

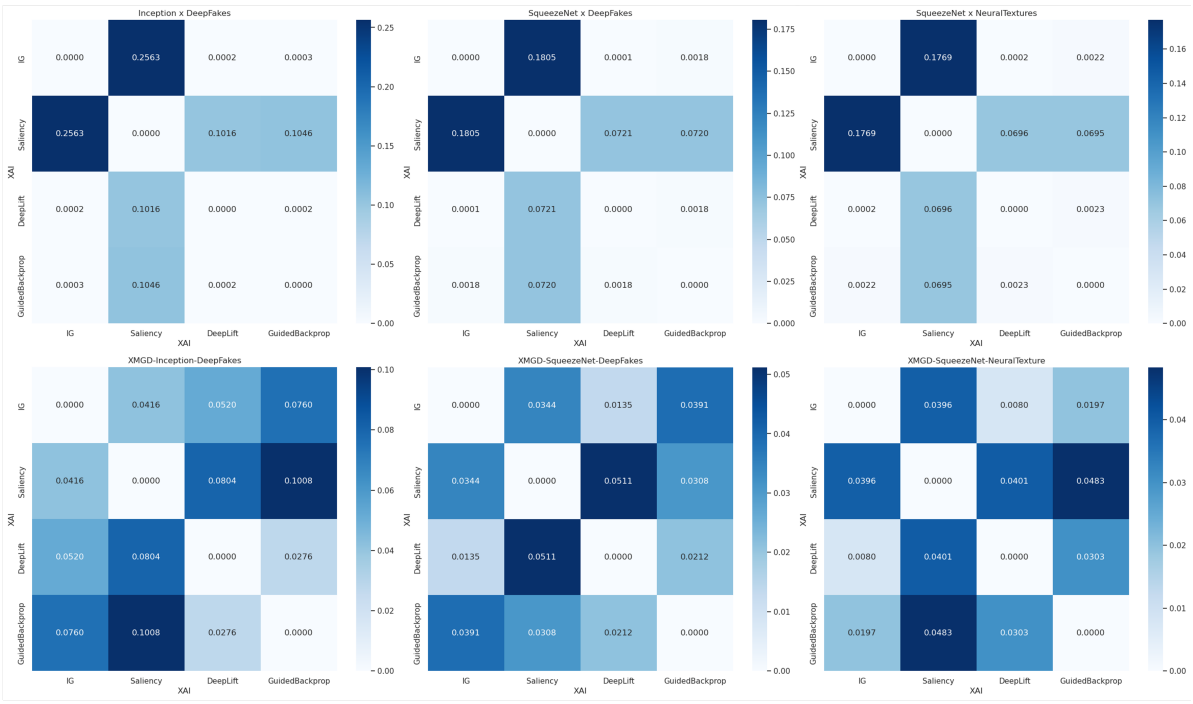

Figure 3. **XMGD vs. KL Comparison**. Single detector and generator saliency maps generated by four XAI methods are compared by KL Divergence (top row) and XMGD (ours, bottom row), where XMGD shows higher correlation when generator or detector is preserved.

To capture (iii), we select two different detectors and two different generators: SqueezeNet and Inception, and Deepfakes and NeuralTextures, respectively. Then, we compare the saliency maps of single-generator trained models of these combinations by KL and by XMGD in Fig. 3. While KL-based comparisons are minimally informative (as most comparison scores are nearly equal across many of the XAI methods), XMGD comparisons by constrast yield more meaningful differentiantions. We dive deeper and compute pair-wise correlations of XMGD scores in the second row of Fig. 3, assuming that same detector and same generator comparisons will surface as high correlations. Deepfakes+SqueezeNet (mid) and Deepfakes+Inception (left) yields 0.813 correlation, Deepfakes+SqueezeNet (mid) and NeuralTextures+SqueezeNet (right) gives 0.810 correlation, whereas the pair without common components yields only 0.742 correlation, as a supporting point for (iii) with room for more exploration.

XMGD is furthermore less sensitive to individual pixel/input saliency features, thus it is more robust as a similarity measure for explainability. This property is due to the fact that XMGD fits a high-order probability distribution, i.e., a GMM, unlike KL Divergence, which relies on discrete, pixel-level calculations. In addition, XMGD is less sensitive to dataset size – particularly for small datasets – as it overcomes poor convergence properties associated with Fréchet Distance-based measures [17]).

6. Discussion

As XMGD is a multi-scale metric, it should normalize the effects of different receptive fields of detectors on saliencies. We observe that it indeed normalizes Inception's similarity and distinguishes sharp saliencies better. We propose investigating ERF and saliency dependency using XMGD for more supporting analysis.

We also propose exploring saliency maps of temporal detectors or transformer-based models, then investigating differences between consecutive saliency maps if they indeed form temporal properties, revealing how to best capture these properties. We end our exploration with asking, can we extend XMGD to temporal saliency distributions?

7. Conclusions

We propose a model-free, generalizable, saliency comparison metric, XMGD, that can be used in principle to compare models, datasets, and XAI methods through an explainability lens. XMGD leverages multi-scale spatial information with flexible, high-order probability distributions to render a robust and informative similarity measure for XAI applications. We demonstrate several key advantages of XMGD over traditional saliency comparison algorithms through large-scale experiments with deepfake detection.

References

[1] Deepfakes. https://github.com/deepfakes/faceswap. Accessed: 2020-03-16. 3

[2] Faceswap. https://github.com/MarekKowalski/FaceSwap. Accessed: 2020-03-16. 3

[3] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, Dec 2018. 3

[4] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:740–757, 2016. 1

[5] 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, 2018. 2, 3

[6] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[7] Marco Cuturi, Laetitia Meng-Papaxanthos, Yingtao Tian, Charlotte Bunne, Geoff Davis, and Olivier Teboul. Optimal transport tools (ott): A jax toolbox for all things wasserstein. *arXiv Preprint arXiv:2201.12324*, January 2022. 3

[8] Bingwei Ge, Fatma Najar, and Nizar Bouguila. Data-weighted multivariate generalized gaussian mixture model: Application to point cloud robust registration. *Journal of Imaging*, 9(9), 2023. 2

[9] Israel Dejene Gebru, Xavier Alameda-Pineda, Florence Forbes, and Radu Horaud. Em algorithms for weighted-data clustering with application to audio-visual scene analysis. *IEEE PAMI*, 38(12):2402–2415, 2016. 2

[10] Tristan Gomez, Thomas Fréour, and Harold Mouchère. Metrics for saliency map evaluation of deep learning explanation methods. In Mounîm El Yacoubi, Eric Granger, Pong Chi Yuen, Umapada Pal, and Nicole Vincent, editors, *Pattern Recognition and Artificial Intelligence*, pages 84–95, Cham, 2022. Springer International Publishing. 2, 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3

[12] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017. 3

[13] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size. *arXiv:1602.07360*, 2016. 3

[14] H. Jung and Y. Oh. Towards better explanations of class activation mapping. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1316–1324, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society. 2, 3

[15] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951. 1

[16] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019. 3

[17] Lorenzo Luzi, Carlos Ortiz Marrero, Nile Wynar, Richard G Baraniuk, and Michael J Henry. Evaluating generative networks using gaussian mixtures of image features. In *WACV*, pages 279–288, 2023. 1, 2, 4

[18] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018. 2

[19] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 1

[20] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. Faceforensics++: Learning to detect manipulated facial images. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 3

[21] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40:99–121, 2000. 2

[22] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3145–3153. JMLR.org, 2017. 3

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 3

[24] Justin Solomon. Optimal transport on discrete domains. *AMS Short Course on Discrete Differential Geometry*, 2018. 3

[25] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 3

[26] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org, 2017. 3

[27] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[28] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph.*, 38(4), July 2019. 3

[29] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2Face: Real-time Face Capture and Reenactment of RGB Videos. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2016. 3

[30] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014. 3

[31] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 1