Rough bibliographical notes on intrinsic images, equivariance and relighting

By D.A. Forsyth, a cut and paste job from many documents for ICVSS-2022

Human surface color

It has been known since at least 1867 that human reports of surface color are much more like reports of albedo than of reflected light [45]; rough algorithms for achieving this date to at least 1874 [46]. A review of early theories for human vision appears in [11]; of more recent ideas in [41]. Humans can recover rich material properties from images [1, 81, 68]. Illusions are common [2].

Color constancy

Recovering surface color in the presence of unknown colored light is sometimes known as color constancy. Early algorithms include [35]; reviews in [3]; more recent are [5, 4, 94, 101, 38, 9, 26, 95]

Evaluation

Direct evaluation methods include: WHDR [12]; scores on the images of [43] (but there are very few images in unrealistic illumination [6]); and scores on SINTEL frames (from [23]), as in [25] (but this rendered data is quite unlike real images as in [58], section 2).

Multiimage data Emerging papers

The WHDR evaluation framework was put in place by [12], who constructed a dataset (Intrinsic Images in the Wild or IIW) consisting of human judgements which compare the absolute lightness at pairs of points in real images. Each pair is labelled with one of three cases (first lighter; second lighter; indistinguishable) and a weight, which captures the certainty of labellers. One evaluates by computing a weighted comparison of algorithm predictions with human predictions; the comparison is known as the weighted human disagreement ratio (WHDR). Predictions were originally by testing ratios of estimated albedo against a standard threshold [12]. Other authors test against a threshold chosen using validation data (eg. rescaling of Retinex in [70]), or test differences in estimated log-albedo (eg [70]). The choice of predictor is significant. Differences in log-albedo are scale invariant, but this predictor may perform poorly over the full range of albedos. Two quite similar dark albedos will have the same difference in logs as two quite different light albedos. Differences in albedo are not scale invariant, and this means that the scale on which the algorithm reports albedo and the test thresholds are fungible. Some authors fix threshold, and learn scale; others fix scale and choose threshold using validation data.

Methods

Early methods do not see any form of training data, but more recently both CGI data and manual annotations of relative lightness (labels) have become available. We organize methods into four classes based on what kind of data is used and how it is used. Methods that use data could use: ground truth data for real images (but it is hard to produce data by experiment, and so only very small quantities of real albedo and shading data are available, e.g. [43, 91]); ground truth data for CGI images (but CGI images present problems, below); simulated data from abstract spatial models (our approach); or statistical summaries of ground truth data. We lump together methods that see ground data for real images and those that see CGI, and order by the likely informativeness of the data. This gives four categories:

- No-ground-truth methods (N-methods) use no labels, albedo or shading for any image;
- Stats-only methods (S-methods) see statistical summaries of CGI albedos and shading, but no other data;
- Paradigm methods (P-methods) use synthetic training data produced by abstract spatial models, but no other data; and
- Ground-truth methods (G-methods) use CGI albedos, CGI shadings, real albedos, real shadings, or human lightness labels.

There is a standard WHDR test-train split (20% test and 80% train) introduced by [70]. The choice of scale and threshold significantly affects reported WHDR (see table 1 of [70]). Table 1 shows reported WHDR's for a large selection of methods, using the best rescaled value known as appropriate. Generally, the more data a method is shown, the better the WHDR; but relatively few G-methods and no S- or N-methods are better than our method.

No-ground-truth methods: (N-methods) In 1959, Edwin Land described procedures that estimated albedo at an image location by accumulating comparisons [55, 56]. Land modelled images as shaded Mondrians — albedos were modelled as piecewise constant patches of color and shading as a smooth field — and concluded that albedo displays large (but no small) image gradients, and that shading has small (but no large) gradients. This **Retinex** assumption results in a class of methods (**Retinex-like** methods) that: compute image gradients; recover albedo gradients from the image gradients (typically, by testing gradient magnitude); then recovering the albedo from the albedo gradients (typically, by a form of integration). The Retinex assumption or variants underly numerous algorithms for recovering albedo, which typically differ by how the albedo gradients are identified and by how albedo is recovered from putative gradients (which are not directly integrable) [67, 48, 47, 18, 22, 53, 31, 97, 60, 59, 104, 90, 20, 34, 21, 19]. The strategy is naturally generalized by (a) writing cost functions or priors that capture the properties of albedo and shading then (b) using an optimization procedure to find albedo and shading that are most consistent with the image and also most like the models [24, 39, 82, 83]; user intervention helps [59, 19]; as does using more than one shading component [85]. Coupling to shape models appears to significantly improve shading and reflectance estimation [7].

N-methods can be trained from data, by showing the method indirect supervisory information. Aligned views of the same real scene under distinct illuminants offer strong cues to intrinsic image decomposition, exploited in [96, 63, 54]. Alternatively, one can use aligned CGI renderings of the same scene [58] (an N-method because the method does not see albedo, just multiple images). [66] show how to exploit these cues to learn a method that, at inference time, can be applied to a single view. [103] show that it is enough to partially align images of real scenes (by matching sections of frames).

Stats-only methods: (S-methods) see only statistical models of albedo (resp. shading), much like the original Retinex assumption. [65] use albedo and shading CGI renderings to build autoencoders. These are used to impose albedo (resp. shading) structure on the inferred components of the input image; the components must also compose to make the image. This method obtains the current SOTA WHDR for any method that doesn't see any ground truth (an extremely strong 18.69%, Table 1).

Paradigm methods: (P-methods, this paper) see samples from abstract statistical models of albedo and shading during training. The key difference between N-methods and P-methods is that P-methods see samples from models (rather than, say, energy functions; priors; etc.). This means that the models can be of complicated form, and inference can be relegated to a network.

Ground-truth methods: (G-methods) see albedo or shading of training images, or labels. With even a few ground truth images are available, local regression strategies have been successful [91]. The recent literature strongly emphasizes directly supervised convolutional neural network (CNN) based models. One option is to [69] regress lightness differences against image features using IIW data. [107] smooth pairwise lightness comparisons (learned using WHDR data) to albedo and shading fields using a fully connected CRF. Recent methods emphasize direct supervision using CGI rendering of scene models [25, 69, 61] However, models trained exclusively on rendered scenes do not do well on real images (eg [58]; section 2). This is likely because rendered images are insufficiently "like" real images in some important ways. Competitive modern methods are trained using a training portion of the IIW dataset, then evaluated on a the test portion. [33] obtain the best current WHDR of 14.45% in this way, but their method produces strange colors in albedo images, making its applicability in computational photography questionable and qualitative comparison unhelpful. [17] use a similar approach, but different network architectures, to obtain a mean WHDR of 17.18% with strong qualitative results; we use this method for qualitative comparison. There is good evidence that relatively little supervision is required, and that self-supervision can be successful. [49] apply a learned renderer to decompositions of unlabelled data to obtain a residual loss that improves performance. [27] show that a form of bootstrapping (augment training data with the results of previous models) is effective in improving performance.

Flattening: WHDR scores can be improved by postprocessing, because most

Table 1: *Summary comparison to recent high performing supervised (above) and unsupervised (below) methods, all evaluated on the standard IIW test set; sources indicated. We distinguish between training with IIW and threshold selection using IIW. WHDR values computed for Retinex use the most favorable scaling, using the rescaling experiments of [70]. For our method, we report the held-out threshold value of WHDR. We report two figures for [15], because we found two distinct figures in the literature. Key: *: method uses IIW training data to set scale or threshold* ONLY*. +: [65] build models of albedo and shading from CGI, but do not use them for direct supervision. a: [103] use patches of registered images from MegaDepth.*

Class	Method	Source	IIW labels	CGI labels	Flattening	Test WHDR
N	*Zhao *et al.* '12 [104]	[70]	N	N	N	26.4
	*Shen and Yeo '11 [83]	[70]	N	N	N	26.1
	Yu and Smith '19[103]	ibid	N	N	N	21.4 (a)
	Retinex (rescaled; color/gray)	[70]	N	N	N	19.5*/18.69*
	*Bell *et al* '14 [12]	[70]	N	N	Y	18.6
	Liu *et al* '20[65]	ibid	N	Y+	N	18.69
	Bi *et al* '15 [15]	ibid	N	N	Y	18.1
	Bi *et al* '15 [15]	[17]	N	N	Y	17.69
S	Liu *et al* '20[65]	ibid	N	Y+	N	18.69
P	Our best		N	N	N	16.86*
G	Shi *et al.* '17[86]	[17]	N	Y	N	54.44
	Zhou *et al* '15[106]	[17]	Y	N	Y	19.95
	*Narihira *et al*[70]	ibid	N	N	N	18.1
	Bi *et al* '18 [17]	ibid	N	Y	Y	17.18
	Zhou *et al* '15[107]	ibid	Y	N	Y	15.7
	Li and Snavely '18[62]	ibid	Y	Y	Y	14.8
	Fan *et al* '18[33]	ibid	Y	N	Y	14.45

methods produce albedo fields with very slow gradients, rather than piecewise constant albedos. [16] demonstrate the value of “flattening” albedo (see also [71]); [17] employ a fast bilateral filter [8] to obtain significant improvements in WHDR.

0.1 Invariance and Equivariance

A function $\phi : \mathbf{x} \in X \to \mathbf{y} \in Y$ is equivariant under the action of a group G if there are actions of G on X and Y such that $\phi(g \circ \mathbf{x}) = g \circ \phi(\mathbf{x})$. An alternative statement of the equivariance property will be convenient. Equivariance means that we can choose a convenient coordinate system in which to evaluate $\Phi(f)$ at $\mathbf{p}$. We have that, for *any* $g \in G$,

$$(g^{-1} \circ \Phi \circ g)(f)(\mathbf{p})$$

does not depend on g. In turn, this supplies a formal construction of an equivariant operation Ψ_{eq} out of any operation Ψ: we could simply average over G, to have

$$\Psi_{\text{eq}}(f) = \left[\int_{g \in G} (g^{-1} \circ \Psi \circ g)(f) dg\right] / \left[\int_{g \in G} dg\right],$$

assuming that the integrals can be constructed, etc. Unfortunately, for most group actions of interest there are very few equivariant mappings that we can evaluate in practice, so there is no reason to construct the integral. If the mapping is per pixel – for example, $\Phi : I(x, y) \to I^2(x, y)$ – it is equivariant, but such mappings are seldom of interest. For other mappings, evaluating $\Phi(f)$ at the point u, v requires knowing f in some window $\mathcal{S}_{u,v}$ that depends on u, v and is larger than a single pixel. Because we know the image only within some viewport on the image plane, we cannot evaluate the mapping for any u, v such that any part of $\mathcal{S}_{u,v}$ lies outside the viewport. Avoiding this problem (for example, by modelling the image as a function on the torus or working with complete spherical images) leads to a rich theory rooted in harmonic analysis [30, 40]. Padding the image is not a solution, because padding means that the process used to evaluate $\Phi(f)$ for u, v close to the boundary is different from that for u, v near the center. Further, the problem can be avoided for some finite group actions [29], and there is good evidence that well-known feature representations are approximately equivariant [57].

There is good evidence that imposing equivariance properties improves models. Imposing permutation equivariance results in better performing learned set-to-set mappings [44]. Functions of point clouds can be equivariant, and [78] show performance improvements from an E(n) equivariant construction of a graph neural network on point clouds. An E(3) equivariant construction for neural interatomic potentials appears in [10]. A general theory for graph neural networks is in [73]. Ignoring equivariance considerations in image-to-image mapping because the theory of group actions doesn't apply is unwise. For many very interesting image-to-image mappings, the estimate at a pixel should not depend on where the pixel is in the image. For example, if Φ maps images to albedos, then the albedo depends on the physical object being viewed, so that if – say – we move the viewport to the left, the albedo should move to the right but not otherwise change. [37] show that a simpler version of our averaging construction produces significant improvements in albedo estimates.

Averaging

We wish to model an image-to-image mapping that we expect naturally has an equivariance property under some group G which acts on the image plane (for example, a map from image to albedo, or from dark image to bright image, should be equivariant under at least rotation, translation and scale). The workhorse of image-to-image mapping is the U-net [77], an image mapper that is flexible as to the size of the input image. U-nets are defined on sampled images. A U-net will not accept an image sampled on a grid with too few samples, because the subsampling processes in the encoder will produce a data block that is empty.

Without loss of generality, we choose some D and always apply our U-net to a $D \times D$ grid (a *tile*). We model an image as a function on the unit square $\mathbb{U} = [0, 1] \times [0, 1]$ and a U-net as an object that will map any image tile, sampled on a $D \times D$ grid, to another function defined on a $D \times D$ grid. Assume we have trained a U-net to implement this mapping in the usual way; write Φ_U to represent this U-net. There is no prospect that the U-net will actually be equivariant, because training procedures do not impose equivariance; the architecture does not guarantee it; and interesting mappings that are formally equivariant are not available anyhow. Fig. **??** illustrates an example where overlapping crops given to the UNet model result in different estimations in the overlapping region.

However, a relaxed version of the procedure to obtain an equivariant mapping from any mapping is extremely interesting. Write S for the *s*ampling operator that maps a function on $\mathbb{U}$ to sampled version of that function on a $D \times D$ and R for a *r*econstruction operator that maps a $D \times D$ sampled grid to a continuous function on $\mathbb{U}$. Write $\mathcal{R}_{\mathbf{p}} = \left\{ g \in G | g^{-1}(\mathbb{U}) \in \mathbb{U} \& g(\mathbf{p}) \in \mathbb{U} \right\}$ – for the set of group operations that takes some window $\mathbf{p} \ni W$ in $\mathbb{U}$ to $\mathbb{U}$. We consider

$$\Phi_{u,\mathrm{eq}}(f)(\mathbf{p}) = \left[\left(\int_{g \in \mathcal{R}_{\mathbf{p}}} w(g)(g^{-1} \circ R \circ \Phi_u \circ g)(f\big|_{g^{-1}(\mathbb{W})})(\mathbf{p}) \right) dg \right] / \left[\int_{g \in \mathcal{R}_{\mathbf{p}}} w(g) dg \right]$$

Here $w(g)$ is a weighting function; for the moment, assume this is one everywhere. Notice this does not result in an equivariant mapping because we cannot average over all group operations – the ones that lead to windows outside $\mathbb{U}$ are omitted. Furthermore, this averaging process is not meaningful if the mapping we are trying to model is not equivariant, because then averaging over G or parts of it is not helpful. The averaging process has important and interesting properties. The estimate of the mapped value at location $\mathbf{p}$ is an ensemble estimate obtained by averaging over many different estimators

$$\Phi_{u,g}(f)(\mathbf{p}) = \left[(g^{-1} \circ R \circ \Phi_u \circ g)(f\big|_{g^{-1}(\mathbb{W})}) \right] (\mathbf{p})$$

(which estimates the value of the mapped f at point $\mathbf{p}$). The ensemble estimate may have reduced variance. The estimators are different, because the U-net sees a different image window for each g in the average. However, training practices mean the estimators should have zero mean (where the random element is the choice of window).

The U-net will be trained with a large number of distinct image crops, and the loss will require that each predicted value be close to the true value. Assuming that the training data is extremely large, the U-net will have seen many distinct windows surrounding a particular pixel, and will be trained to predict the same value for each. The random element of the estimate at a particular pixel is the choice of window containing that pixel that is presented to the U-net. We can expect that training will result in a U-net that has zero mean error.

Zero mean error at each pixel is not the same as error that has no spatial structure. We expect that the error at different locations in the output of the U-net is correlated over some range of scales, because many pairs of output units have overlapping receptive fields. This means the error could take the form of a moderately sized, spatially slow, but structured, error field (Fig. **??**).

An ensemble estimate can control this class of error if we can force down the variance at each location. This occurs if the error produced by each of the estimators $\Phi_{u,g}(f)(\mathbf{p})$ in the average is "sufficiently independent" and if we do not average in estimators with large variance. As section **??** demonstrates, this can be achieved in practice. If we have some reliable method of identifying estimators with large variance, the weighting function can be used to down weight them. As section **??** demonstrates, this can be achieved in practice.

A network should not change prediction if the input image is shifted or scaled. In other words, an ideal method will report the same estimation for the same location in a scene, however that location is viewed. We know of no crisp theoretical framework to impose this criterion. The theory of group actions does not exactly apply to transformations of the input image such as shifting, cropping, scaling or even rotating because almost all transformations of this form involve information being gained or lost at the boundary of the image [**?**].

GAN reading

You have to read [42]. There is some discussion of the effects of features, etc in []. Sauer et al argue that using a projected set of features results in a match in the projected space [79]. Little is known about what happens when there isn't a saddle point. Exponential model averaging is known to be a very effective way to control cycling in GANs [102].

Image manipulation: StyleGAN [51, 52, 50] is currently de facto state-of-the-art for editing generated images, likely because its mapping of initial noise vectors to style codes which control entire feature layers produces latent spaces that are heavily disentangled and so easy to manipulate. Recent editing methods include [93, 98, 84, 108, 28, 76], with a survey in [99]. The architecture can be adapted to incorporate spatial priors for authoring novel and edited images [64, 92, 32].

Reshading and relighting

Relighting discussion in slides is from [13, 36] Insertion rendering from [14]; an improvement, not yet on arxiv, is entitled "Image-based Object Insertion using Persistent and Transient Decomposition" (by Anand Bhattad, Brian Chen, Stephan R. Richter, David A. Forsyth) and should appear there shortly. There is a background on diffuse interreflections and their mathematics in [36].

Image Relighting: Shih et al. show that matching to time-lapse video together with an example-based color transfer scheme can relight outdoor scenes [87]. For scenes, there are workshop tracks (*eg.* [**?**, **?**]), challenges [**?**, **?**] and datasets [**?**, **?**]. Existing image relighting work learns image mappings – pure image mappings in [**?**], depth guided in [**?**], using wavelets in [**?**] shadow priors in [**?**]. In all these cases, methods are learned with paired data of the same scene under different illuminations, available in the VIDIT dataset [**?**] and the MIE dataset [**?**]. VIDIT data is CGI, and emphasizes point light sources with strong shadows, which are uncommon in indoor scenes. Pairing is necessary to ensure that these methods preserves scene characteristics [**?**, **?**]. Date augmentation with relighting improve two patch matching tasks [**?**]. Methods can learn to create outdoor shadows [**?**] and soft attached shadows for objects that have been inserted into indoor scenes [**?**, **?**].

Image Relighting using StyleGAN: [92] uses StyleGAN to relight faces but require three-dimensional morphable face model. In contrast, StyLitGAN does not require any 3D model of the scene. [100] uses semantic label attributes "indoor lighting" and "natural lighting" to train a binary classifier to find directions in latent space that represent them, but cannot produce diverse relighting and requires a search in decoder layers to apply relight edits without changing layout of the scene.

Face Relighting Methods mostly use carefully collected supervisory data from light-stages [88, 105, 72, 80, 75]. ShadeGAN [74] and Volux-GAN [89] uses a volumetric rendering approach to learn the underlying 3D structure of the face and the illumination encoding. Volux-GAN also requires image decomposition obtained from [75] that is trained using a carefully curated light-stage data.

References

[1] Adelson, E.H.: On seeing stuff: the perception of materials by humans and machines. In: Proceedings of the SPIE, vol. 4299, Human Vision and Electronic Imaging VI (2001) 1

[2] Adelson, E.: Lightness perception and lightness illusions. In: The new cognitive neurosciences (2000) 1

[3] Barnard, K., Cardei, V., Funt, B.: A comparison of computational color constancy algorithms. I: Methodology and experiments with synthesized data. IEEE Transactions in Image Processing (2002) 1

[4] Barron, J.T.: Convolutional Color Constancy. In: International Conference on Computer Vision (2015) 1

[5] Barron, J.T., Malik, J.: Color constancy, intrinsic images, and shape estimation. ECCV (2012) 1

[6] Barron, J.T., Malik, J.: Shape, illumination, and reflectance from shading. IEEE transactions on pattern analysis and machine intelligence (2014) 1

[7] Barron, J.T., Malik, J.: Shape, Illumination, and Reflectance from Shading. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(8), 1670–1687 (2015) 3

[8] Barron, J.T., Poole, B.: The fast bilateral solver. In: Proceedings of the European Conference on Computer Vision (2016) 5

[9] Barron, J.T., Tsai, Y.T.: Fast Fourier Color Constancy. In: Computer Vision and Pattern Recognition (2017) 1

[10] Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J., Kornbluth, M., Molinari, N., Smidt, T., Kozinsky, B.: E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials (2022) 5

[11] Beck, J.: Surface color perception. Cornell University Press (1972) 1

[12] Bell, S., Bala, K., Snavely, N.: Intrinsic images in the wild. ACM Trans. on Graphics (SIGGRAPH) (2014) 1, 4

[13] Bhattad, A., Forsyth, D.A.: Enriching stylegan with illumination physics (2022). https://doi.org/10.48550/ARXIV.2205.10351, https://arxiv.org/abs/2205.10351 8

[14] Bhattad, A., Forsyth, D.A.: Cut-and-paste neural rendering (2020). https://doi.org/10.48550/ARXIV.2010.05907, https://arxiv.org/abs/2010.05907 8

[15] Bi, S., Han, X., Yu, Y.: An L1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition. ACM Transactions on Graphics **34**(4), 78–78:12 (Jul 2015) 4

[16] Bi, S., Han, X., Yu, Y.: An l 1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition. ACM Transactions on Graphics (TOG) (2015) 5

[17] Bi, S., Kalantari, N.K., Ramamoorthi, R.: Deep hybrid real and synthetic training for intrinsic decomposition. In: Eurographics Symposium on Rendering (2018) 3, 4, 5

[18] Blake, A.: Boundary conditions for lightness computation in mondrian world. Computer Vision, Graphics and Image Processing **32**, 314–327 (1985) 2

[19] Bousseau, A., Bousseau, A., Paris, S., Durand, F.: User-assisted intrinsic images. In: ACM Transactions on Graphics (TOG). p. 130. ACM (Dec 2009) 2, 3

[20] Brainard, D.H., Freeman, W.T.: Bayesian color constancy. JOSA A (1997) 2

[21] Brainard, D., Wandell, B.: Analysis of the retinex theory of color vision. J. Opt. Soc. America - A **3** (1986) 2

[22] Brelstaff, G., Blake, A.: Computing lightness. Pattern Recognition Letters **5**(2), 129–38 (1987) 2

[23] Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: ECCV (2012) 1

[24] Chang, J., Cabezas, R., Fisher III, J.W.: Bayesian Nonparametric Intrinsic Image Decomposition. In: European Conference on Computer Vision (2014) 3

[25] Chen, Q., Koltun, V.: A simple model for intrinsic image decomposition with depth cues. In: ICCV (2013) 1, 3

[26] Cheng, D., Prasad, D.K., Brown, M.S.: Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution. JOSA A (2014) 1

[27] Cheng, L., Zhang, C., Liao, Z.: Intrinsic image transformation via scale space decomposition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018) 3

[28] Chong, M.J., Lee, H.Y., Forsyth, D.: Stylegan of all trades: Image manipulation with only pretrained stylegan. arXiv preprint arXiv:2111.01619 (2021) 7

[29] Cohen, T.S., Welling, M.: Group equivariant convolutional networks. In: ICML (2016) 5

[30] Cohen, T., Geiger, M., Weiler, M.: A general theory of equivariant cnn's on homogeneous spaces. In: NeurIPS (2019) 5

[31] Elad, M., Kimmel, R., Shaked, D., Keshet, R.: Reduced complexity retinex algorithm via the variational approach. J Vis Commun Image R **14**(4), 369–388 (2003) 2

[32] Epstein, D., Park, T., Zhang, R., Shechtman, E., Efros, A.A.: Blobgan: Spatially disentangled scene representations. arXiv preprint arXiv:2205.02837 (2022) 7

[33] Fan, Q., Yang, J., Hua, G., Chen, B., Wipf, D.: Revisiting deep intrinsic image decompositions. In: CVPR (2018) 3, 4

[34] Farenzena, M., Fusiello, A.: Recovering intrinsic images using an illumination invariant image. pp. III: 485–488 (2007) 2

[35] Forsyth, D.A.: A novel algorithm for color constancy. IJCV (1990) 1

[36] Forsyth, D.A., Bhattad, A., Asthana, P., Zhong, Y., Wang, Y.: Sirfyn: Single image relighting from your neighbors (2021). https://doi.org/10.48550/ARXIV.2112.04497, https://arxiv.org/abs/2112.04497 8

[37] Forsyth, D., Rock, J.: Intrinsic image decomposition using paradigms. TPAMI (In Press), https://www.computer.org/csdl/journal/tp/5555/01/09573351/1xH5D2WNbEc 6

[38] Gao, S., Han, W., Yang, K., Li, C., Li, Y.: Efficient Color Constancy with Local Surface Reflectance Statistics. In: European Conference on Computer Vision. pp. 158–173. Springer, Cham, Cham (Sep 2014) 1

[39] Gehler, P.V., Rother, C., Kiefel, M., Zhang, L., Scholkopf, B.: Recovering Intrinsic Images with a Global Sparsity Prior on Reflectance. In: Advances in neural information processing systems. pp. 765–773 (2011) 3

[40] Gerken, J.E., Aronsson, J., Carlsson, O., Linander, H., Ohlsson, F., Petersson, C., Persson, D.: Geometric deep learning and equivariant neural networks. In: arxiv (2021), https://arxiv.org/abs/2105.13926 5

[41] Gilchrist, A.: Seeing Black and White. Oxford University Press (2006) 1

[42] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems 27. Curran Associates, Inc. (2014), http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf 7

[43] Grosse, R., Johnson, M.K., Adelson, E.H., Freeman, W.T.: Ground truth dataset and baseline evaluations for intrinsic image algorithms. In: International Conference on Computer Vision. pp. 2335–2342 (2009) 1, 2

[44] Hartford, J., Graham, D.R., Leyton-Brown, K., Ravanbakhsh, S.: Deep models of interactions across sets. In: ICML (2018) 5

[45] Helmholtz, H.: Treatise on Physiological Optics. Thoemmes (2000), orig. publication 1866 1

[46] Hering, E.: Outlines of a theory of the light sense (1964), translated from the German of 1874 by L.M Hurvich and D. Jameson 1

[47] Horn, B.K.: On lightness. MIT AI Memo 295 (1973) 2

[48] Horn, B.K.: Determining lightness from an image. Computer graphics and image processing (1974) 2

[49] Janner, M., Wu, J., Kulkarni, T.D., Yildirim, I., Tenenbaum, J.B.: Self-supervised intrinsic image decomposition. In: NIPS (2017) 3

[50] Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. Advances in Neural Information Processing Systems **34** (2021) 7

[51] Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 7

[52] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: Proc. CVPR (2020) 7

[53] Kimmel, R., Elad, M., Shaked, D., Keshet, R., Sobel, I.: A variational framework for retinex. International Journal of computer vision **52**(1), 7–23 (2003) 2

[54] Laffont, P.Y., Bazin, J.C.: Intrinsic decomposition of image sequences from local temporal variations. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (December 2015) 3

[55] Land, E.: Color vision and the natural image: Part i. PNAS **45**(1), 115–129 (January 1959) 2

[56] Land, E.: Color vision and the natural image: Part ii. PNAS **45**(4), 636–644 (April 1959) 2

[57] Lenc, K., Vedaldi, A.: Understanding image representations by measuring their equivariance and equivalence. Int J Comput Vis **127**(456–476) (2019) 5

[58] Lettry, L., Vanhoey, K., Gool, L.V.: Deep unsupervised intrinsic image decomposition by siamese training. CoRR **abs/1803.00805** (2018), `http://arxiv.org/abs/1803.00805` 1, 3

[59] Levin, A., Weiss, Y.: User Assisted Separation of Reflections from a Single Image Using a Sparsity Prior. IEEE TPAMI **29**(9), 1647–1654 (Jan 2007) 2, 3

[60] Levin, A., Zomet, A., Weiss, Y.: Separating reflections from a single image using local features. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. (2004) 2

[61] Li, Z., Snavely, N.: CGIntrinsics - Better Intrinsic Image Decomposition Through Physically-Based Rendering. ECCV **11207**(4), 381–399 (2018) 3

[62] Li, Z., Snavely, N.: Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 371–387 (2018) 4

[63] Li, Z., Snavely, N.: Learning intrinsic image decomposition from watching the world. In: Computer Vision and Pattern Recognition (CVPR) (2018) 3

[64] Ling, H., Kreis, K., Li, D., Kim, S.W., Torralba, A., Fidler, S.: Editgan: High-precision semantic image editing. Advances in Neural Information Processing Systems **34** (2021) 7

[65] Liu, Y., Li, Y., You, S., Lu, F.: Unsupervised learning for intrinsic image decomposition from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 3, 4

[66] Ma, W.C., Chu, H., Zhou, B., Urtasun, R., Torralba, A.: Single image intrinsic decomposition without a single intrinsic image. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018) 3

[67] McCann, J.J., McKee, S.P., Taylor, T.H.: Quantitative studies in retinex theory a comparison between theoretical predictions and observer responses to the "color mondrian" experiments. Vision research (1976) 2

[68] Motoyoshi, I., Nishida, S., Sharan, L., Adelson, E.H.: Image statistics and the perception of surface qualities. Nature (2007) 1

[69] Narihira, T., Maire, M., Yu, S.X.: Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In: Proceedings of the IEEE international conference on computer vision (2015) 3

[70] Narihira, T., Maire, M., Yu, S.X.: Learning lightness from human judgement on relative reflectance. In: Proceedings of the IEEE CVPR (2015) 1, 2, 4

[71] Nestmeyer, T., Gehler, P.V.: Reflectance adaptive filtering improves intrinsic image estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6789–6798 (2017) 5

[72] Nestmeyer, T., Lalonde, J.F., Matthews, I., Games, E., Lehrmann, A., Borealis, A.: Learning physics-guided face relighting under directional light (2020) 8

[73] N.Keriven, Peyre, G.: Universal invariant and equivariant graph neural networks. In: NeurIPS (2019) 5

[74] Pan, X., Xu, X., Loy, C.C., Theobalt, C., Dai, B.: A shading-guided generative implicit model for shape-accurate 3d-aware image synthesis. In: Advances in Neural Information Processing Systems (NeurIPS) (2021) 8

[75] Pandey, R., Escolano, S.O., Legendre, C., Haene, C., Bouaziz, S., Rhemann, C., Debevec, P., Fanello, S.: Total relighting: learning to relight portraits for background replacement. ACM Transactions on Graphics (TOG) **40**(4), 1–21 (2021) 8

[76] Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2287–2296 (2021) 7

[77] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015) 6

[78] Satorras, V.G., Hoogeboom, E., Welling, M.: E(n) equivariant graph neural networks. In: Arxiv (2021), https://arxiv.org/abs/2102.09844 5

[79] Sauer, A., Chitta, K., Müller, J., Geiger, A.: Projected gans converge faster. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 17480–17492. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper/2021/file/9219adc5c42107c4911e249155320648-Paper.pdf 7

[80] Sengupta, S., Curless, B., Kemelmacher-Shlizerman, I., Seitz, S.M.: A light stage on every desk. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021) 8

[81] Sharan, L., Li, Y., Motoyoshi, I., Nishida, S., Adelson, E.H.: Image statistics for surface reflectance perception. J. Opt. Soc. Am. A **25**(4), 846–865 (Apr 2008) 1

[82] Shen, J., Yang, X., Jia, Y., Li, X.: Intrinsic images using optimization. In: Computer Vision and Pattern Recognition. pp. 3481–3487. IEEE (2011) 3

[83] Shen, L., Yeo, C.: Intrinsic images decomposition using a local and global sparse representation of reflectance. CVPR (2011) 3, 4

[84] Shen, Y., Yang, C., Tang, X., Zhou, B.: Interfacegan: Interpreting the disentangled face representation learned by gans. IEEE transactions on pattern analysis and machine intelligence (2020) 7

[85] Sheng, B., Li, P., Jin, Y., Tan, P., , Lee, T.Y.: Intrinsic image decomposition with step and drift shading separation. TPAMI (2020) 3

[86] Shi, J., Dong, Y., Su, H., Yu, S.X.: Learning nonlambertian object intrinsics across shapenet categories. In: CVPR (2017) 4

[87] Shih, Y., Paris, S., Durand, F., Freeman, W.T.: Data-driven hallucination of different times of day from a single outdoor photo. ACM Trans. Graph. (proc. SIGGRAPH Asia) **32**(6) (2013) 8

[88] Sun, T., Barron, J.T., Tsai, Y.T., Xu, Z., Yu, X., Fyffe, G., Rhemann, C., Busch, J., Debevec, P., Ramamoorthi, R.: Single image portrait relighting. ACM Transactions on Graphics (Proceedings SIGGRAPH) (2019) 8

[89] Tan, F., Fanello, S., Meka, A., Orts-Escolano, S., Tang, D., Pandey, R., Taylor, J., Tan, P., Zhang, Y.: Volux-gan: A generative model for 3d face synthesis with hdri relighting. arXiv preprint arXiv:2201.04873 (2022) 8

[90] Tappen, M.F., Freeman, W.T., Adelson, E.H.: Recovering intrinsic images from a single image. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(9), 1459–1472 (2005) 2

[91] Tappen, M.F., Adelson, E.H., Freeman, W.T.: Estimating Intrinsic Component Images using Non-Linear Regression. In: Computer Vision and Pattern Recognition. pp. 1992–1999. IEEE (2006) 2, 3

[92] Tewari, A., Elgharib, M., Bharaj, G., Bernard, F., Seidel, H.P., Pérez, P., Zollhofer, M., Theobalt, C.: Stylerig: Rigging stylegan for 3d control over portrait images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6142–6151 (2020) 7, 8

[93] Voynov, A., Babenko, A.: Unsupervised discovery of interpretable directions in the gan latent space. arXiv preprint arXiv:2002.03754 (2020) 7

[94] van de Weijer, J., Gevers, T.: Edge-based color constancy. IEEE Transactions on Image Processing (2007) 1

[95] van de Weijer, J., Schmid, C.: Using high-level visual information for color constancy. Computer Vision pp. 1–8 (2007) 1

[96] Weiss, Y.: Deriving intrinsic images from image sequences. In: ICCV (2001) 3

[97] Weiss, Y.: Deriving intrinsic images from image sequences. In: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001 (2001) 2

[98] Wu, Z., Lischinski, D., Shechtman, E.: Stylespace analysis: Disentangled controls for stylegan image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12863–12872 (2021) 7

[99] Xia, W., Zhang, Y., Yang, Y., Xue, J.H., Zhou, B., Yang, M.H.: Gan inversion: A survey. arXiv preprint arXiv: 2101.05278 (2021) 7

[100] Yang, C., Shen, Y., Zhou, B.: Semantic hierarchy emerges in deep generative representations for scene synthesis. International Journal of Computer Vision (2020) 8

[101] Yang, K.F., Gao, S.B., Li, Y.J.: Efficient illuminant estimation for color constancy using grey pixels. In: Computer Vision and Pattern Recognition. pp. 2254–2263. IEEE (2015) 1

[102] Yazıcı, Y., Foo, C.S., Winkler, S., Yap, K.H., Piliouras, G., Chandrasekhar, V.: The unusual effectiveness of averaging in gan training. arXiv preprint arXiv:1806.04498 (2018) 7

[103] Yu, Y., Smith, W.A.: Inverserendernet: Learning single image inverse rendering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019) 3, 4

[104] Zhao, Q., Tan, P., Dai, Q., Shen, L., Wu, E., Lin, S.: A closed-form solution to retinex with nonlocal texture constraints. IEEE TPAMI **34**(7), 1437–1444 (2012) 2, 4

[105] Zhou, H., Hadap, S., Sunkavalli, K., Jacobs, D.W.: Deep single-image portrait relighting. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7194–7202 (2019) 8

[106] Zhou, T., Krahenbühl, P., Efros, A.A.: Learning data-driven reflectance priors for intrinsic image decomposition. In: International Conference on Computer Vision. pp. 3469–3477. IEEE (2015) 4

[107] Zhou, T., Krahenbuhl, P., Efros, A.A.: Learning data-driven reflectance priors for intrinsic image decomposition. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (December 2015) 3, 4

[108] Zhu, J., Shen, Y., Zhao, D., Zhou, B.: In-domain gan inversion for real image editing. In: Proceedings of European Conference on Computer Vision (ECCV) (2020) 7