# Interpreting Intrinsic Image Decomposition using Concept Activations

Avani Gupta
CVIT, KCIS
IIIT-Hyderabad
avani.gupta@research.iiit.ac.in

Saurabh Saini
CVIT, KCIS
IIIT-Hyderabad
saurabh.saini@research.iiit.ac.in

P J Narayanan
CVIT, KCIS
IIIT-Hyderabad
pjn@iiit.ac.in

## ABSTRACT

Evaluation of ill-posed problems like Intrinsic Image Decomposition (IID) is challenging. IID involves decomposing an image into its constituent illumination-invariant Reflectance (R) and albedo-invariant Shading (S) components. Contemporary IID methods use Deep Learning models and require large datasets for training. The evaluation of IID is carried out on either synthetic Ground Truth images or sparsely annotated natural images. A scene can be split into reflectance and shading in multiple, valid ways. Comparison with one specific decomposition in the ground-truth images used by current IID evaluation metrics like LMSE, MSE, DSSIM, WHDR, SAW AP%, *etc.*, is inadequate. Measuring R-S *disentanglement* is a better way to evaluate the quality of IID. Inspired by ML interpretability methods, we propose Concept Sensitivity Metrics (CSM) that directly measure disentanglement using sensitivity to relevant concepts. Activation vectors for albedo invariance and illumination invariance concepts are used for the IID problem. We evaluate and interpret three recent IID methods on our synthetic benchmark of controlled albedo and illumination invariance sets. We also compare our disentanglement score with existing IID evaluation metrics on both natural and synthetic scenes and report our observations. Our code and data are publicly available for reproducibility [1].

## CCS CONCEPTS

• **Computing methodologies** → **Image-based rendering**; • **Networks** → **Network performance analysis**.

## KEYWORDS

Intrinsic Image Decomposition, ill-posed problems, evaluation techniques, Disentanglement, ML interpretability, TCAV

**ACM Reference Format:**
Avani Gupta, Saurabh Saini, and P J Narayanan. 2022. Interpreting Intrinsic Image Decomposition using Concept Activations. In *Proceedings of the Thirteenth Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'22), December 8–10, 2022, Gandhinagar, India,* Soma Biswas, Shanmuganathan Raman, and Amit K Roy-Chowdhury (Eds.). ACM, New York, NY, USA, Article 3, 9 pages. https://doi.org/10.1145/3571600.3571603


*ICVGIP'22, December 8–10, 2022, Gandhinagar, India*

ACM ISBN 978-1-4503-9822-0/22/12…$15.00
https://doi.org/10.1145/3571600.3571603

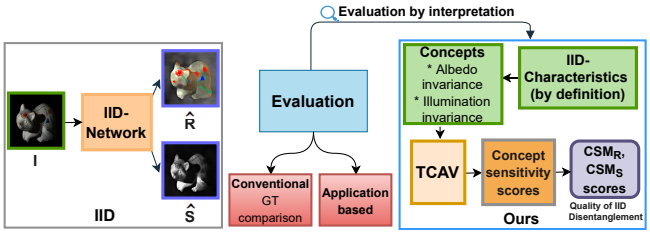


**Figure 1: Evaluation-via-interpretation:** Given a pre-trained IID network, existing evaluation techniques rely on comparison with ground truth or performance in a downstream application. We propose a novel *evaluation-via-interpretation* strategy based on learned Concept Activation Vectors (CAV) [23]. We estimate concept sensitivity scores and evaluate IID performance by measuring the quality of albedo-illumination disentanglement via our proposed Concept Sensitivity Metric (CSM).

## 1 INTRODUCTION

Image-Based Inverse-Rendering (IBIR) problems like image stylization, image harmonization, illumination estimation, palette extraction, *etc.*, are often under-constrained and ill-posed in nature. They are under-constrained as we need to estimate more output parameters than available inputs, *e.g.* style-content decomposition from a single image for stylization. These problems are also frequently ill-posed due to the underlying optical model approximations and assumptions, *e.g.* diffuse surfaces, monochromatic illumination, point light source, *etc.* As a result, performance evaluation of their solutions becomes a challenging task. The issue is exacerbated due to the lack of a proper ground truth dataset and evaluation metric. To address this issue, we propose a novel *evaluation by interpretation* technique in this paper thereby introducing a Concept Sensitivity Metric (CSM). We focus on one such problem in the paper. *i.e.*, Intrinsic Image Decomposition (IID) [27].

IID is an IBIR task that involves decomposing a given image into its constituent illumination-invariant Reflectance ($R$) and albedo-invariant Shading ($S$) components. The decomposition finds direct use in many applications, such as shadow removal [26], image colorization [34], material manipulation [6], relighting [13], and retexturing [6]. Current IID methods assume a simple Lambertian reflectance model on diffuse surfaces:

$$I = R \odot S, \tag{1}$$

where $\odot$ denotes element-wise multiplication. Due to the under-constrained nature, existing IID methods either depend on hand-crafted [5, 30, 46] or deep learned [12, 31, 32, 35] priors.

[1] https://github.com/avani17101/CSM

Performance evaluation of IID is carried out on a small number of natural images like MIT Intrinsics [20], sparse human annotation datasets like Intrinsic Images in the Wild (IIW) [5], Shading Annotation in the Wild (SAW) [25], *etc.*, or synthetic datasets like Sintel [8], As-Realistic-As-Possible (ARAP) [7], *etc.* For densely annotated GT images, evaluation is carried using dense per pixel error estimation or using quality score involving metrics like Peak Signal-to-Noise Ratio (PSNR), Local Mean Square Error (LMSE), Difference in Structural Similarity Index (DSSIM), *etc.* For sparsely annotated GT datasets, IID-specific metrics like Weighted Human Disagreement Ratio (WHDR) [5] and Average Precision (AP%) of classified shading pixel regions [25] have been proposed. Yet another way to evaluate IID solutions is via the effectiveness of the decomposed components in a downstream application. Bonneel et al. [7] propose hardcoded application scenarios like logo removal, shadow removal, texture replacement, and wrinkles attenuation on a fixed set of hand-picked 21 images to benchmark IID solutions. These evaluation strategies either require dense GT annotations which are available only for synthetic scenes [7, 8] (with exception of a few single object images from Grosse et al. [20]) or are dataset specific with sparse human annotations [5] [25]. Since multiple $R$ and $S$ pairs can result in the same image, even the "ground truth" is only one possible solution and measures on that is inadequate to evaluate the method.

To address these issues, we propose a new evaluation strategy for ill-posed problems like IID by measuring the *quality of disentanglement* between the decomposed components $R$ and $S$. We use the core IID concepts of illumination-invariance of $R$ and albedo-invariance of $S$ to measure disentanglement, without specifically relying on synthetic images or relative quality metrics computed on fixed sparsely annotated datasets. We choose an ML interpretability technique based upon Testing with Concept Activation Vectors (TCAV) [23] for this. Originally introduced for classifiers, TCAV is a post-hoc Concept-based Model Extraction (CME) [21] method that interprets a Neural Network using human understandable concepts. Specifically, TCAV quantifies the importance of a user-defined concept in the model's prediction by extracting activation vectors from a provided *concept set*. For example, for a zebra classifier one may be interested in interpreting concepts like 'striped-ness' *vs.* 'dotted-ness', which are defined by learning *Concept Activation Vectors* w.r.t. the model from user provided sets with striped and dotted textures respectively. We use as concepts two core characteristics derived from the very definition of IID, *i.e.*, illumination-invariance of $R$ and albedo-invariance of $S$. We assess disentanglement between them by measuring the model's sensitivity to these concepts in the form of *Concept Sensitivity Metrics* (CSM) (Figure 1). The CSM provides a generic framework applicable to problems other than IID using concepts relevant to them. To summarise, the main contributions our work are:

- A novel method for using ML interpretability algorithms like TCAV to measure disentanglement.
- A novel IID performance evaluation metric: Concept Sensitivity Metric (CSM) and benchmarked results on three state-of-the-art IID solutions.
- A new configurable dataset of images and corresponding generation scripts with controlled illumination and albedo variation.

## 2 RELATED WORK

### 2.1 Intrinsic Image Decomposition

As the core idea behind our proposed approach involves applying an ML interpretablity technique for IID evaluation, we discuss the relevant literature under the two sub-sections mentioned below:

***IID Methods:*** IID as modeled in Equation 1 was first proposed by Land and McCann [29]. Earlier IID solutions were mostly unsupervised optimization based approaches constrained by strong assumptions and specific auxiliary inputs like time-lapse video [32], multi-view images [13], IID using stereo images [28], IID on RGBD data [3], focal stacks [42], *etc.* Single image IID methods depend upon complex cost functions and optimization algorithms like IID by chromatic clustering [16], convex energy minimization [18], hierarchical priors [40, 41], *etc.* With the advent of Deep Learning, various supervised, semi-supervised and unsupervised Neural Network based solutions have been proposed in the literature. Initially, Bell et al. [5] and Zhou et al. [50] proposed a hybrid DL and optimization based framework for IID. Narihira et al. [37] proposed a direct $R$ and $S$ regression framework trained on synthetic Sintel dataset [8]. Li and Brown [30] introduced a relative loss function for reflectance estimation. Li and Snavely [32] learned an unsupervised IID model using time-lapse videos and consistency loss. Finally, Li and Snavely [31] and Fan et al. [14] trained multiple sequential modules supervised by hybrid synthetic, sparse, and dense datasets with appropriately designed loss functions. A fully unsupervised DL approach has been proposed [35], which poses IID as a style-transfer problem. PIE-Net [12] has a hybrid-CNN approach for addressing shading-reflectance leakages in strong illumination conditions whereas Baslamisli et al. [4] use photometric invariance and some other physics based priors in encoder-decoder architecture.

As earlier optimization based approaches are interpretable by design, we focus on recent state-of-the-art DL based IID models. Specifically, we focus on 3 IID solutions:

- Intrinsic Images by Watching the World (**IIWW**) [32] trained in a partially-supervised manner on their self-introduced Bigtime dataset consisting of time-lapse videos of natural indoor and outdoor scenes.
- CGIntrinsics (**CGIID**) [31] which does supervised training on their new synthetic dataset containing physically based renderings with GT $R$ and $S$, as well as natural scenes from IIW [5] and SAW [25] datasets.
- Unsupervised Single Image Intrinsic Image Decomposition (**USI3D**) [35] which first disentangles content from style features, then utilizes adversarial learning to separately learn $R$ and $S$ style domains and performs content preserving image translation with consistency losses for IID in an unsupervised training regime.

***Evaluation Strategies:*** Since there is lack of dense real world GT annotations for $R$ and $S$, all the above models are evaluated on synthetic images (ARAP [7], Sintel [8]), small single object scenes (MIT Intrinsics [20]) or sparse manual annotations (IIW [5], SAW [25]). Synthetic GT based evaluation is affected by synthetic-natural domain shift, whereas single object images do not capture the complexity of everyday natural scenes. Sparse manually annotated GT from IIW and SAW either provide only relative assessments or

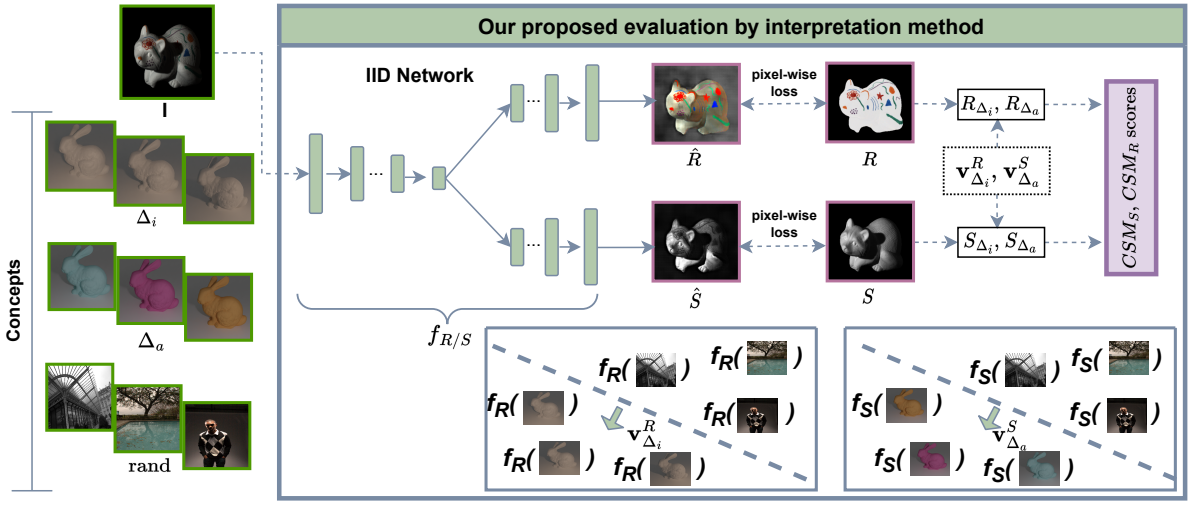


**Figure 2: Disentanglement Quality Assessment:** Given the concept sets illumination invariance $\Delta_a$ and albedo invariance $\Delta_i$ along with their negative counterpart *rand*, we learn and store Concept Activation Vectors (CAVs) $(v^S_{\Delta_a}, v^R_{\Delta_i})$ for the pretrained IID network by classifying the respective branch's concept activations $(f_S(), f_R())$ (where $f_Y$ means network output for branch $Y$) of the positive and negative concept set images. During inference, given an input image $I$, we compute its dense pixel loss by comparing predicted $(\hat{R}, \hat{S})$ with corresponding ground truth $(R, S)$ values. Based on the loss gradient's alignment along the learned CAV $(v^S_{\Delta_a}, v^R_{\Delta_i})$ directions, concept sensitivity values are estimated $(R_{\Delta_a}, R_{\Delta_i}, S_{\Delta_a}, S_{\Delta_i})$ and combined to give $CSM_S$ and $CSM_R$ scores which measure IID disentanglement.

use classification accuracies on fixed shading categories. Additionally, Bonneel et al. [6] acknowledge the issue of IID evaluation and propose to evaluate IID quality by estimating the performance in downstream image editing applications (logo removal, shadow removal, texture replacement and wrinkles attenuation) using the decomposed $R$ and $S$ components. None of these IID evaluation strategies specifically capture disentanglement quality of the decomposed $R$ and $S$, and implicitly assume that the small set of curated GT annotations/cases represents all the possible test scenarios.

***Metrics:*** On densely annotated synthetic GT images, IID quality is measured using pixel-to-pixel comparisons and metrics like Local Mean Square Error (LMSE), Mean Squared Error (MSE), and Difference in Structural Similarity Index (DSSIM). These metrics are not robust to the ambiguous nature of the IID problem ($I = R \odot S = \lambda R \odot \frac{S}{\lambda}, \quad \forall \lambda \in \mathbb{R}^+$). For sparse human annotated GTs like IIW and SAW, Weighted Human Disagreement Ratio [5] (WHDR) measures the percentage of disagreement between human assessment and model prediction weighted by the confidence of each annotation. The SAW AP% [25] is calculated based upon average precision on varying recall percentages over classification of pixels into smooth vs non-smooth shading regions. Both of these metrics only assess sparse set of pixels and ignore specific cases like multiple shadows, colored highlights, material transmissivity, *etc.* Also, they are specific to dataset which is comprised of mostly indoor scenes. Current limitations of IID metrics motivate the search for a more comprehensive and fundamental evaluation strategy which we attempt to address by our proposed approach in this paper.

## 2.2 Neural Network Interpretation

The goal of Neural Network Interpretation research is to go beyond the mere black box usage or accuracy based interpretation of Deep Learning architectures and develop an understanding of the internal workings of the learned model. Several techniques have been used for this purpose like activation maps visualization [45], saliency estimation [43], model simplification [47][39], model perturbation [15], adversarial exemplar analysis [19], *etc.* For more details on existing state of the art interpretability methods refer Linardatos et al. [33]. We focus on a specific category of post-hoc model interpretation techniques based on analysis of concepts drawing motivation from [22] which compare disentanglement approaches with concept based approaches. Zaidi et al. [48] on the other hand review various mathematical disentanglement metrics. Kim et al. [23] define model *interpretation* as a function $g : E_m \rightarrow E_h$, where $E_m$ represents the vector space of the model state and $E_h$ represents the space of high-level human-understandable concepts. *Concepts* space, $E_h$, is defined using a set of user provided samples that exemplify the desired concept well. Kim et al. [23] measure sensitivity towards *Concept Activation Vectors* (CAV) as an interpretation function ($g$) where CAV for a concept is defined as a vector in the direction of activations for the provided concept set examples. Given a few positive and negative pairs of concepts and a target class, it assigns a score to how much significance a concept has in the class prediction. While TCAV requires the concept set given by the user, ACE [17] automates the process of concept discovery by using superpixel segmentation and clustering in the activation space to get concept definitions directly from class images. Other concept based approaches like [2, 21, 24, 44] learn concepts to predict class and use them for explaining the model's predictions.

Although TCAV was initially proposed only for classification problems but in this work we extend it for IID like image generation tasks. We define two IID specific CAVs ($R$ illumination-invariance and $S$ albedo-invariance). These concepts are fundamental to the mathematically ideal modeling of the two decomposed components

and hence can be used to quantify the model's adherence to the very definition of IID. We create two synthetic concept sets for these concepts and benchmark three pre-trained IID models on our proposed Concept Sensitivity Metric (CSM) which measures the sensitivity of model towards the estimated CAVs.

## 3 APPROACH

In this section, we first provide a quick primer on the background of Testing with CAV [23], followed by our definition of IID *concept sets* for $R$ illumination-invariance and $S$ albedo-invariance and proposed IID *evaluation strategy* using our novel Concept Sensitivity Metric.

### 3.1 TCAV Primer

Testing with Concept Activation Vectors (TCAV) [23] is a Neural Network interpretation method that calculates a pre-trained model's sensitivity towards a user-defined concept by training a Concept Activation Vector (CAV) in the feature space and analysing the layers' activations direction w.r.t. the learned CAV. A concept can be any abstract high-level human-understandable category (*e.g.*, colors *vs.* non-color, striped *vs.* dotted textures, model-woman *vs.* stripes, *etc.*) and is defined in terms of a set of images.

For input image $x$ from the testset, layer $l$'s activation is given by $f_l(x)$. Given a concept of interest $C$, a set of images of that concept are taken as positive samples *vs.* a set of random images are taken as negative concept set ($C'$). A binary linear classifier (linear regression or SVM) is used to distinguish between $l$'s activations for $C$ and $C'$ (*i.e.*, between sets $f_l(x) : x \in C$ and $f_l(y) : y \in C'$). The vector representing the classifier hyperplane is stored as the CAV, $v_C^l$. Model's concept sensitivity $S_{C,l} \in [0, 1]$ for a given class sample $x$ is estimated by calculating the alignment between the learned CAV $v_C^l$ and the gradient $\nabla L_l\left(f_l(\boldsymbol{x})\right)$ of loss with respect to activation for that layer (computed via back-propagation) as:

$$S_{C,l}(\boldsymbol{x}) = \nabla L_l\left(f_l(\boldsymbol{x})\right) \cdot \boldsymbol{v}_C^l, \quad \text{where} \quad \nabla L_l\left(f_l(\boldsymbol{x})\right) = \frac{\partial L(\boldsymbol{x})}{\partial \boldsymbol{x}_{a,b}}. \quad (2)$$

TCAV has been used to analyze classifiers using a Cross-entropy (CE) loss between the predicted logit and GT class label. We use TCAV to evaluate the disentanglement of image decomposition by introducing a pixel-wise loss instead of CE. For pixel location $(a, b)$, $L$ is calculated as MSE between the predicted $\hat{R}$, $\hat{S}$ and their respective GT values. Finally, the complete TCAV sensitivity score for concept $C$ is computed as the fraction of inputs ($x$'s) in the complete concept set $X$, which were positively aligned with $v_C^l$.

$$\text{TCAV}_{C,l} = \frac{\left|\left\{\boldsymbol{x} \in X : S_{C,l}(\boldsymbol{x}) < 0\right\}\right|}{|X|} \quad (3)$$

Note that the sign is negative here because gradient is taken with respect to loss instead of logit values. Thus when $C$ has a positive influence, the loss is minimised[1].

### 3.2 IID: Evaluation by Interpretation

***IID Concepts:*** For IID evaluation-via-interpretation, we define two concepts: albedo-invariance ($C_{\Delta_i}$) and illumination-invariance ($C_{\Delta_a}$). For definition of associated concept sets ($\Delta_i$ and $\Delta_a$), we render synthetic images of objects by varying one concept while fixing the other. We use random textures, albedo maps and different illumination settings for this. For negative concept set, we randomly select images from a large dataset (unrelated to IID).

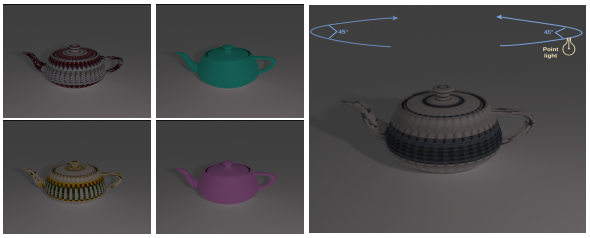

**Figure 3: Concept sets:** Left grid of images shows four samples from our $\Delta_a$ concept set with varying textures and base colors. Image on the right illustrates light source variation setting (±22 in either direction) for rendering $\Delta_i$ concept set images.

***R and S Sensitivity:*** By IID definition, in an ideal case $\Delta_a$ should only affect $\hat{R}$ and $\Delta_i$ should just affect $\hat{S}$. In other words, the sensitivity $R_{\Delta_a}$ of $\hat{R}$ towards concept $C_{\Delta_a}$ must be high and the sensitivity $R_{\Delta_i}$ towards concept $C_{\Delta_i}$ must be low. Conversely, for $\hat{S}$, the sensitivity $S_{\Delta_a}$ towards concept $C_{\Delta_a}$ must be low and the sensitivity $S_{\Delta_i}$ towards concept $C_{\Delta_i}$ must be high. In the ideal case of complete $R$-$S$ disentanglement, $R_{\Delta_a}$, $S_{\Delta_i}$ should be 1 and $R_{\Delta_i}$, $S_{\Delta_a}$ should be 0. Due to several inherent assumptions in the IID definition (diffuse surfaces, linear optics, *etc.*), total disentanglement is impossible and a measure of disentanglement will be useful.

***Concept Sensitivity Metric (CSM).*** We evaluate the model's disentanglement quality by combining the above sensitivity scores to gauge the model's performance in the two experiments separately to give Concept Sensitivity Metric (CSM) scores. CSM scores give a quantitative measure to gauge the quality of $\hat{R}$ *vs.* $\hat{S}$ disentanglement. We introduce two CSM scores: $CSM_S$ which measures $\hat{S}$ albedo invariance and $CSM_R$ which measures $\hat{R}$ illumination invariance:

$$CSM_S = \frac{R_{\Delta_a}}{S_{\Delta_a}} \quad \textit{and} \quad CSM_R = \frac{S_{\Delta_i}}{R_{\Delta_i}}. \quad (4)$$

Higher value of $CSM_S$ indicates less leakage of albedo information in $\hat{S}$. Similarly, higher value of $CSM_R$ indicates less illumination leakage in $\hat{R}$. We verify the same experimentally in section 5.

## 4 EXPERIMENTAL SETUP

### 4.1 Datasets

***Concept Sets:*** For the two IID concepts, the respective concept sets ($\Delta_a$ and $\Delta_i$) are rendered in Blender [11] in a controlled environment. We generate two types of scenes: *simple scenes* with a single object and *complex scenes* with multiple (3) objects. We setup the scene by randomly choosing from a set of frequently used standard 3D meshes[2] and placing them on a white table-top. We place a point light source of white color, 1000W power, 0.5m radius with fixed temperature $T \in \{2500, 4500, 6500\}$. Specifically, for $\Delta_a$, we render scenes with fixed viewpoint and illumination but randomly vary the base color and surface texture. For $\Delta_i$, we vary illumination by rotating the light source every 2° in ±45 left and right directions as shown in Figure 3 but keep the camera viewpoint

[1] https://github.com/tensorflow/tcav

[2] https://github.com/alecjacobson/common-3d-test-models

and base color/surface texture constant (thus getting 44 images: 22 in the left and 22 in right direction). For defining the negative concept set $C'$, we take random images from Adobe-5k-dataset [9].

We also compare the CAVs learned from our synthetically rendered concept sets against available natural scene datasets. For $\Delta_a$ concept set, we found no suitable large enough dataset with natural albedo variations, so we report on synthetic concept set only. For $\Delta_i$ we use two publicly available datasets which have natural illumination changes: *(i)* Multi-Illumination Images in the Wild dataset (**MIW**) [36]: test split which consists of 30 scenes under 25 different illuminations. *(ii)* Photometric Stereo dataset (**PS**) [1] consists of 3 objects (apple, gourd1, gourd2) under several illumination settings in all directions (approximately 100 per object). Though MIT Intrinsics dataset [20] also has scenes in varying illuminations (11), it cannot be used as concept set because at least 20 images are needed per concept for a stable CAV estimation as recommended by Kim et al. [23].

***Testsets:*** We take ARAP[7] dataset images as our input $x$ for IID networks. ARAP contains realistic synthetic renderings of both indoor and outdoor scenes. We remove the single object scenes ('Katie', 'redhead', 'skin', 'strawberries', 'toad', 'revolution') to maintain inter-scene consistency taking the remaining complex scenes to get a total of 42 scenes which have 3-4 varying illuminations.

## 4.2 Implementation details

We use our PyTorch [38] implementation of TCAV. We use the pre-trained models and the inference codes for IIWW[32], USI3D[35], and CGIID[31] from their official repositories and use their respective hyper-parameter settings. Our sensitivity computation requires only activation values of the pre-trained model and does not require full training. The hardware requirement of our framework depends on the model being analysed for CAV estimation. We tested our framework on 2 Nvidia GTX1080Ti GPUs, which were required by the largest model we analyzed (*USI3D*).

We perform multiple iterations (100) of CAV estimation experiments for robust concept definition. We decide upon the number of iterations through exhaustive experimentation as reported in section 5. In each iteration we use 100 rendered images per set with varying albedo for $C_{\Delta_a}$ and 44 images with varying illumination for $C_{\Delta_i}$. All images are resized to $256 \times 256$ dimensions. With each iteration we perform hypothesis significance testing (double sided t-test with $p = 0.01$ kept same as [23]) and average over the passing significant CAVs.

## 4.3 Experimental Details

We analyze our framework for different scenarios by enumerating over a combination of various experimental conditions:

- **Layer Selection:** Original CAV sensitivity can be evaluated for any layer of the model. In our IID adaptation, we restrict to the last layer sensitivities. There are two reasons for this design choice. First, the IID concepts which we are trying to capture are high-level abstractions which are better represented by the deeper layers. Second, different IID methods have different number of layers making the choice of comparable layers difficult across architectures *e.g.* IIWW and CGIID both have separate branches of $R$ and $S$ while USI3D has two generators of $R$ and $S$ styles plus a content encoder. Computing CAV sensitivities on the last layer makes our method more stable and architecture agnostic.
- **Scene Domain:** We analyze our technique under both synthetic and natural domains by choosing appropriate concept sets.
- **Concept Scene Complexity:** We estimate CAVs for both simple and complex scene settings with single object and multiple objects respectively.
- **Concept Albedo Complexity:** We render $\Delta_a$ concept sets with two albedo complexity settings: Simple RGB base color change and texture variation. In the first case, each object has one randomly assigned solid RGB color. In the second case, we apply a random texture map from DTD dataset [10] on each object.

Overall, we have four experimental settings for each concept: *Textured-Simple*, *Textured-Complex*, *RGB-Simple* and *RGB-Complex*. We perform comprehensive experiments by forming multiple concept sets under each of above categories. Specifically, for $C_{\Delta_a}$ concept we have 10 scenes rendered under 4 different illumination directions with 3 illumination temperatures $t \in \{2500, 4500, 6500\}$, hence $10 \times 3 \times 4 = 120$ concept sets each containing 100 albedo/texture variation images. Similarly for $C_{\Delta_i}$ we have 10 scenes $\times 3$ temperatures $\times 3$ albedos = 90 concept sets, each with 44 (22 left + 22 right) illumination direction variation images.

# 5 RESULTS AND ANALYSIS

We report our CSM disentanglement scores ($CSM_S, CSM_R$) for $C_{\Delta_a}$ and $C_{\Delta_i}$ in Table 1 after averaging over all the corresponding concepts sets. From Table 1, we find that USI3D does best disentanglement of albedo from $\hat{S}$ (best $\hat{S}$ albedo invariance thus highest $CSM_S$) while CGIID does best disentanglement of illumination information from $\hat{R}$ (highest $CSM_R$) amongst the three models. We also show illustrative qualitative results in Figure 4 which show predicted $\hat{R}$ and $\hat{S}$ from the three methods for the same scene under 2 different albedo ($A_0$ and $A_1$) and illumination ($I_0$ and $I_1$) settings.

***Performance in $\hat{S}$ albedo invariance: $CSM_S$.*** Albedo variations for the same scene are rarely observed in the training sets. Due to this, supervised methods like CGIID perform poorly on $CSM_S$ metric compared to unsupervised IIWW and USI3D. USI3D being completely unsupervised performs best, followed by IIWW which is partially unsupervised (assuming constant reflectance over time-lapse videos of varying illumination scenes). From Figure 4, USI3D has least changes in $\hat{S}$ for $\Delta a$.

***Performance in $\hat{R}$ illumination invariance: $CSM_R$.*** CGIID has significantly higher $CSM_R$ in all the four experimental settings, followed by IIWW and then USI3D as shown in Table 1 and verified from qualitative results in Figure 4 where CGIID observes least changes in $R$ for illumination variations. It also does well on real-world concept sets as seen from Table 3 and qualitative results Figure 6. The same trend is seen over complex scenes from ARAP dataset Figure 5 and MIT Intrinsics [20] (shown in supplementary pdf). This is because illumination variations are captured to some extent in existing IID datasets and hence the concept $C_{\Delta_i}$ can be learned by supervision. Thus, CGIID being a completely supervised network, performs well on $CSM_R$ metric. IIWW being trained on time-lapse videos of BigTime dataset [32] comes next, followed by USI3D which is completely unsupervised and relies on style

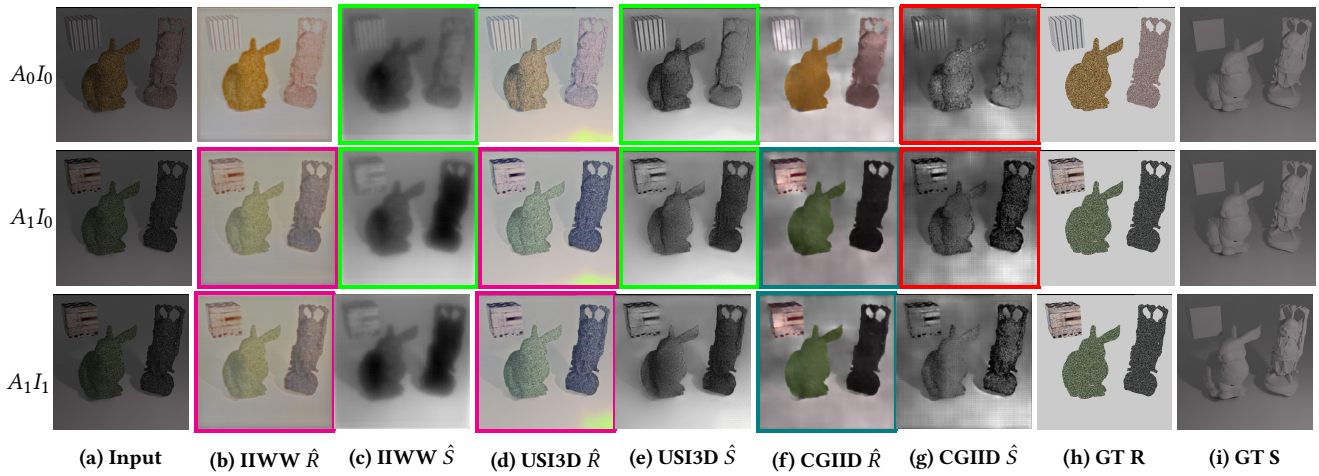


**Figure 4: Illustrative qualitative results for albedo variation (first and second rows) and illumination variation (second and third row) experiments.** $A_iI_j$ represent scene in albedo i and illumination j. For albedo variation ($A_0 \rightarrow A_1$) , USI3D [35] followed by IIWW [32] observes least changes in $\hat{S}$ (green) and thus is best in disentangling albedo information from $\hat{S}$ while CGIID [31] (magenta) is worst. For illumination variation ($I_0 \rightarrow I_1$) CGIID observes least changes in $\hat{R}$ (teal) having least leakage of shadows in $\hat{R}$ for all three rows while IIWW and USI3D have comparatively more shadow leakage (magenta). This is reflected in our $CSM_S$ and $CSM_R$ scores in Table 2 unlike in existing metrics Table 4. (Best viewed in color.)

| Model | $CSM_S \uparrow$ | | | | | $CSM_R \uparrow$ | | | | | $WHDR \downarrow$ | $SAWAP\% \uparrow$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Textured | | RGB | | | Textured | | RGB | | | | |
| | Simple | Complex | Simple | Complex | $\overline{CSM_S} \uparrow$ | Simple | Complex | Simple | Complex | $\overline{CSM_R} \uparrow$ | | |
| IIWW [32] | **2.049** | 1.286 | 0.939 | **1.732** | 1.524 | 0.857 | 0.979 | 0.936 | 0.741 | 0.878 | 20.3 | 91.87 |
| USI3D [35] | 2 | **1.943** | **2.806** | 1.669 | **2.139** | 0.648 | 0.504 | 0.38 | 0.674 | 0.552 | 18.69 | 78.69 |
| CGIID [31] | 0.753 | 1.328 | 0.606 | 0.93 | 0.909 | **1.35** | **1.231** | **1.806** | **1.79** | **1.544** | **14.8** | **97.93** |

**Table 1: Disentanglement quality:** Quality of disentanglement for albedo variation (as measured by $CSM_S$) and illumination variation (as measured by $CSM_R$). USI3D performs best in $CSM_S$ metric and CGIID the worst. The trend is reversed for the $CSM_R$ metric. Note: Best is **bold**, second best is underlined.

| Temp | Model | $CSM_S \uparrow$ | | | | | $CSM_R \uparrow$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Textured | | RGB | | $\overline{CSM_S} \uparrow$ | Textured | | RGB | | $\overline{CSM_R} \uparrow$ |
| | | Simple | Complex | Simple | Complex | | Simple | Complex | Simple | Complex | |
| 2500 | IIWW [32] | 1.735 | 1.112 | 0.894 | 1.426 | 1.292 | 1.021 | 1.098 | 0.994 | 0.757 | 0.968 |
| | USI3D [35] | **4.138** | **2.802** | **4.11** | **2.643** | **3.423** | 0.449 | 0.406 | 0.275 | 0.646 | 0.444 |
| | CGIID [31] | 0.901 | 1.379 | 0.554 | 1.051 | 0.971 | **1.166** | **1.412** | **1.52** | **1.66** | **1.44** |
| 4500 | IIWW [32] | 1.995 | 1.544 | 0.955 | **1.877** | 1.593 | 0.805 | 0.979 | 0.886 | 0.771 | 0.86 |
| | USI3D [35] | **2.352** | **2.248** | **2.697** | 1.54 | **2.209** | 0.646 | 0.497 | 0.385 | 0.645 | 0.543 |
| | CGIID [31] | 0.745 | 1.334 | 0.706 | 0.9 | 0.921 | **1.347** | **1.282** | **1.858** | **1.862** | **1.587** |
| 6500 | IIWW [32] | **2.5** | 1.248 | 0.965 | **1.969** | **1.671** | 0.763 | 0.879 | 0.93 | 0.693 | 0.816 |
| | USI3D [35] | 1.091 | **1.285** | **2.063** | 1.253 | 1.423 | 0.942 | 0.74 | 0.512 | 0.727 | 0.73 |
| | CGIID [31] | 0.611 | 1.278 | 0.558 | 0.837 | 0.821 | **1.624** | **1.295** | **2.147** | **1.865** | **1.733** |

**Table 2: Disentanglement quality in different temperatures:** On an average a similar trend is followed across temperatures.

distributions of $R$ and $S$. The same trend is seen in MIT Intrinsics [20] (shown in qualitative results in supplementary video).

**Effect of different temperature settings.** We show results on different $T$ in Table 2. For $CSM_S$, the same trend is observed for T = 2500 and 4500: USI3D>IIWW»CGIID, while for T=6500, IIWW is slightly better than USI3D (IIWW>USI3D>CGIID). With an increase in T, its occurrence gets rare in training sets. Thus, USI3D being an unsupervised method, does significantly better than its partially

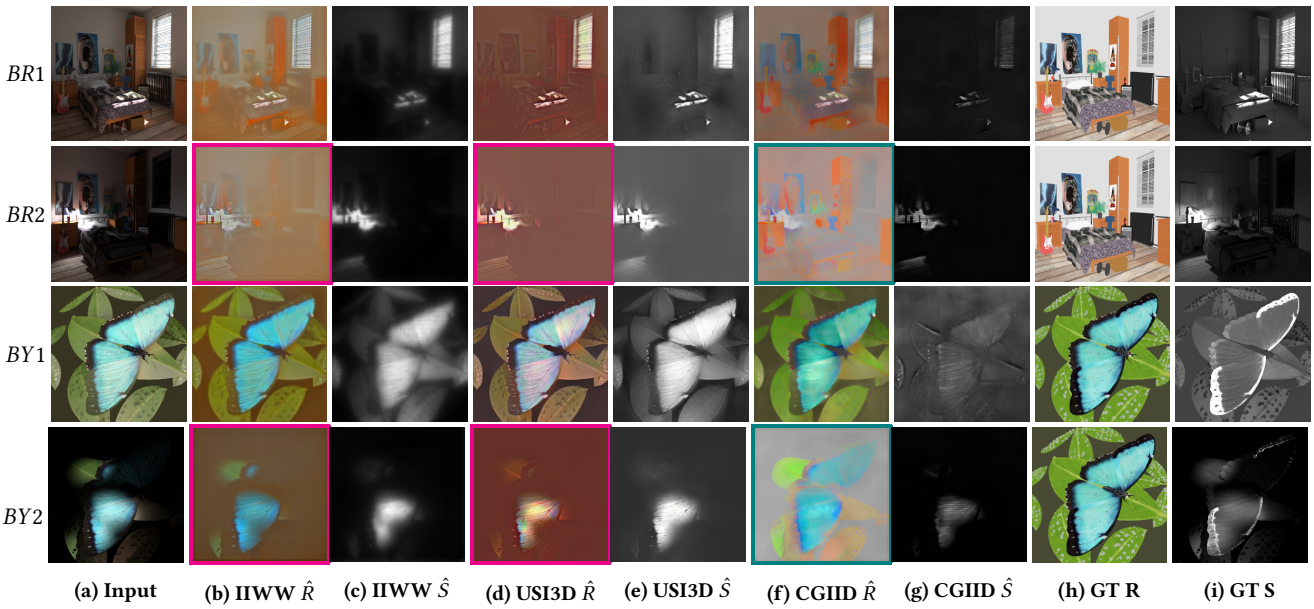


**Figure 5: Illumination changes in ARAP dataset scenes:** The first and second rows are bedroom scene in different illuminations (BR1, BR2) while third and fourth are butterfly in different illuminations (BY1, BY2). For both the scenes CGIDD has least leakage of illumination in $\hat{R}$. In BR1 and BR2, the shadow information is leaked more in $\hat{R}$ for both IIWW and USI3D (magenta), while it is leaked lesser for CGIDD (green). From $BY1 \longrightarrow BY2$, butterfly's top wing has a drastic illumination change. CGIDD predicts this top wing's albedo correctly to huge extent(green) while IIWW and USI3D have that illumination directly leaked in $\hat{R}$ (magenta)

| Model | MIW [36] | | | PS [1] | | | $\overline{CSM_R} \uparrow$ |
|---|---|---|---|---|---|---|---|
| | $R_i \downarrow$ | $S_i \uparrow$ | $CSM_R \uparrow$ | $R_i \downarrow$ | $S_i \uparrow$ | $CSM_R \uparrow$ | |
| IIWW [32] | 0.391 | 0.296 | 0.757 | 0.409 | 0.501 | 1.225 | <u>0.991</u> |
| USI3D [35] | 0.173 | 0.133 | 0.751 | 0.75 | 0.612 | 0.816 | 0.784 |
| CGIID [31] | 0.373 | 0.462 | **1.239** | 0.061 | 0.587 | **9.623** | **5.431** |

**Table 3: Natural concept datasets:** Similar to synthetic case in Table 1, CGIID significantly performs the best in $CSM_R$ followed by IIWW and USI3D on both natural image datasets. This shows no significant shift between the two domains for CAV computation.

| Img | Model | MSE↓ | | LMSE↓ | | D-SSIM↓ | | LPIPS↓[49] | |
|---|---|---|---|---|---|---|---|---|---|
| | | R | S | R | S | R | S | R | S |
| $A_0I_0$ | IIWW [32] | 0.025 | 0.021 | 0.006 | 0.004 | 0.254 | **0.365** | 0.336 | 0.429 |
| | USI3D [35] | **0.013** | 0.058 | 0.006 | 0.006 | **0.188** | 0.431 | **0.185** | **0.396** |
| | CGIID [31] | 0.021 | **0.02** | 0.006 | **0.003** | 0.28 | 0.421 | 0.364 | **0.396** |
| $A_1I_0$ | IIWW [32] | 0.035 | **0.041** | 0.011 | 0.011 | 0.261 | **0.456** | 0.362 | 0.439 |
| | USI3D [35] | **0.014** | 0.07 | 0.006 | **0.01** | **0.162** | 0.495 | **0.153** | **0.406** |
| | CGIID [31] | 0.062 | **0.041** | 0.006 | 0.014 | 0.327 | 0.559 | 0.388 | 0.443 |

**Table 4: Limitation of standard IID metrics:** For two exemplar images $A_0I_0$ and $A_1I_0$ in Figure 4, USI3D exhibits best performance for $\hat{R}$ although it contains shadow leakages. Whereas for the $\hat{S}$ component, the metrics contradict each other.

supervised (IIWW) and supervised (CGIID) counterparts. For T = 6500, IIWW has slightly better $CSM_S$, the reason being its training on illumination varying BigTime dataset having temperatures in that range (most natural images are near 6500 T). For $CSM_R$, same overall ordering of models is observed across temperatures: CGIID>IIWW>USI3D. Qualitative results for the same are given in supplementary pdf.

**Effect of number of CAV iterations**. We experiment with the number of iterations for stable CAV verification by t-testing as pointed in subsection 3.1 and find that any iterations above 80 work well (Figure 7). We thus take 100 as our number of cav iterations.

**Comparison with existing metrics: R-S disentanglement.** Individual comparisons of $\hat{R}$ and $\hat{S}$ with GT only consider how close $\hat{R}$, $\hat{S}$ are to GT and do not consider disentanglement. $\hat{R}$ and $\hat{S}$ closeness to GT does not guarantee disentanglement since the reconstruction can be good enough but entain illumination-albedo leakages. As shown in Table 4, $\hat{R}$ gets better MSE, LMSE, D-SSIM and LPIPS [49] values but still has shadow leakages because it resembles GT the most, except for pixels where shadows are there. USI3D on the other hand, achieves best performance in $\hat{R}$ in terms of existing metrics on MIT Intrinsics, ARAP as well as our synthetic concept sets(given in Supplementary pdf), but it has clear illumination leakages in $\hat{R}$ for which its score is penalised by our method which gauges disentanglement.

WHDR and SAW AP% are not designed for measuring R-S disentanglement by their very definition, which can be seen from Table 1 where they don't align with $CSM_R$ and $CSM_S$ along with qualitative results as shown in Figure 4, 5, 6. Ideally R, and S must be disentangled by the definition of IID but must also resemble GT. For example, though CGIID is best in terms of disentangling illumination information from $\hat{R}$, closeness of $\hat{R}$ to GT achieved best by USI3D is important as well (Figure 4 Table 4). Hence, our method must be combined with existing GT comparison metrics for the most robust IID evaluation.

**Cross-dataset performance** USI3D has significantly lower performance on SAW AP% metric which measures quality of $\hat{S}$, but it

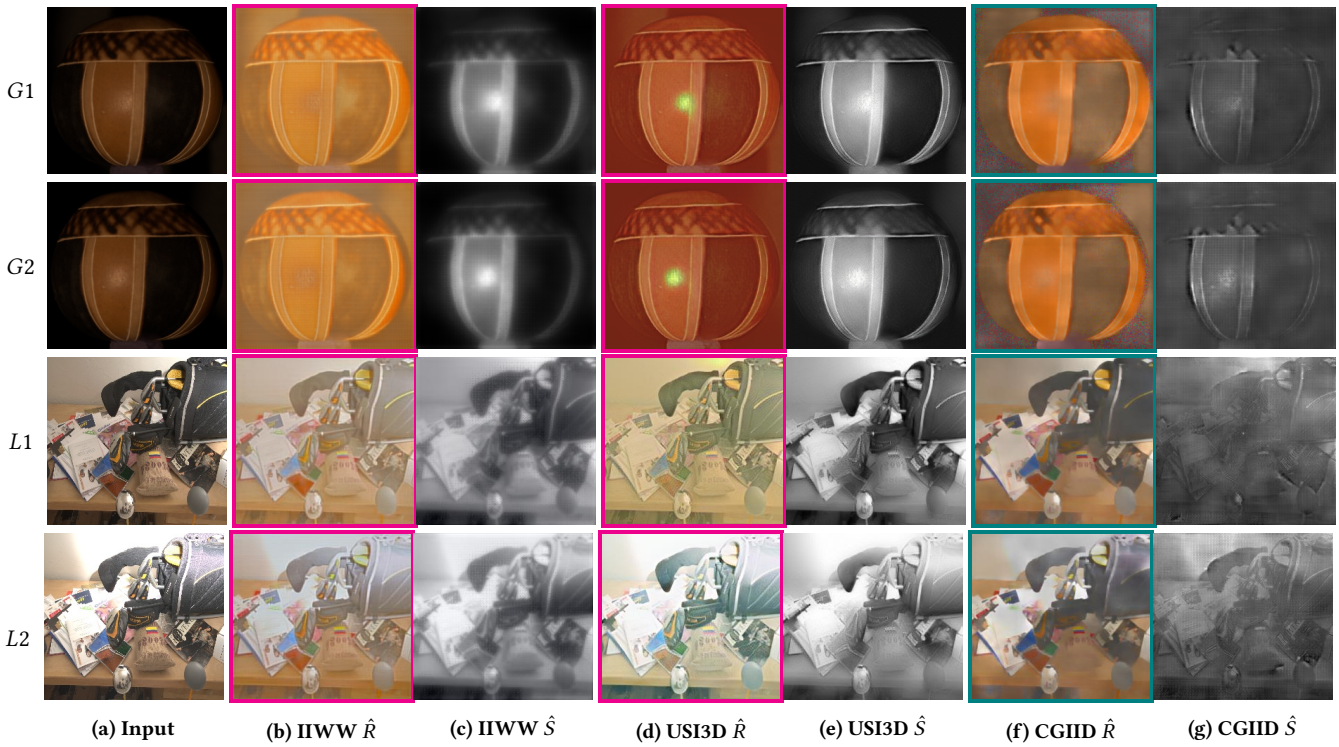


**Figure 6: Real-world illumination change results:** First two rows are images from Gourd scene (G) which belongs to PS [1] dataset, last two rows are scene lobby (L) from MIW[36]. For G1 and G2, $\hat{R}$ illumination variance is most in USI3D (circular luminant) followed by IIWW while it is least by CGIID. Though CGIID is unable to predict good $\hat{R}$ for the dark background of G1 and G2, its foreground's $\hat{R}$ is good. In lobby scene (L1 and L2) USI3D has global intensity changes in $\hat{R}$ and IIWW also has a more changes compared to CGIID which has lesser changes (as seen in black bag on right). Thus CGIID's $\hat{R}$ is less sensitive to $\Delta_i$ followed by IIWW and then USI3D (which performs worst).

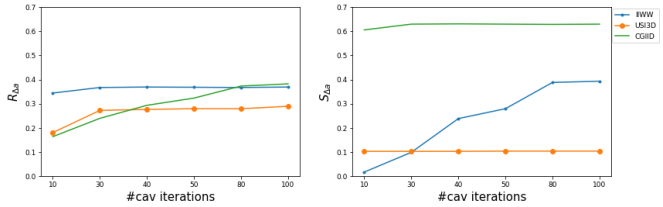


**Figure 7: Sensitivity scores with number of CAV iterations** for RGB-Simple concepts $C_{\Delta_a}$. Both sensitivity scores $R_{\Delta_a}$ and $S_{\Delta_a}$ plateau after 80 iterations for all three models. Note: $S_{\Delta_a}$ for USI3D is constantly low since $C_{\Delta_a}$ is in-significant for its $\hat{S}$.

exhibits best performance in terms of pixel-wise comparisons on MIT Intrinsics[20], ARAP dataset[7], and our synthetic concept set (shown in supplementary pdf). This highlights the issue of cross-dataset performance evaluation inconsistency for the standard IID metrics. On the other hand our proposed $CSM_R$ metric performs consistently and exhibits the same trend in performance: CGIID > IIWW > USI3D (see $\overline{CSM_R}$ in Table 1 and Table 3)

***Limitations of our concept sensitivity based approach.*** Models can be sensitive to concepts in ways that are not desirable. For example, a model might predict noisy R which keeps on changing (as seen from black foggy artifacts in CGIID predictions Figure 4) and get a high $R_{\Delta_a}$ score. Similarly, artifacts in $\hat{S}$ might lead to high sensitivity. In such rare cases, usually R and S both have noise and taking the ratios cancels out and gives a lower score to the model.

## 6 CONCLUSION

We presented Concept Sensitivity Metric, a framework that adapts an ML interpretability method, to evaluate the quality of IID based on its definition. The $CSM_R$ and $CSM_S$ metrics evaluate the disentanglement of the recovered reflectance and shading. These metrics overcome several shortcomings of the current IID evaluation strategies. They are consistent over real-world and synthetic scenes and have lesser dependence on the evaluation set as we use model's sensitivity towards concepts rather than direct pixel-to-pixel comparison with ground truth annotations.

Since these metrics measure the quality of the output and can provide additional terms to the loss being minimized to improve the IID calculations like in a fine-tuning step. We intend to work on this in the future. The use of metrics defined for interpretability in a loop to improve the performance on the original problem has wide scope of applicability.

The approach underlying Concept Sensitivity Metric have wider potential application beyond the IID problem. Choosing appropriate concepts and their activations, CSM can be used to evaluate results of image harmonization, style transfer, image enhancement, *etc.*

# REFERENCES


[1] Neil Alldrin, Todd Zickler, and David Kriegman. 2008. Photometric stereo with non-parametric and spatially-varying reflectance. In *Computer Vision and Pattern Recognition (CVPR)*.

[2] David Alvarez Melis and Tommi Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. *Neural Information Processing Systems (NIPS)* (2018).

[3] Jonathan T Barron and Jitendra Malik. 2013. Intrinsic scene properties from a single rgb-d image. In *Computer Vision and Pattern Recognition (CVPR)*.

[4] Anil S Baslamisli, Yang Liu, Sezer Karaoglu, and Theo Gevers. 2021. Physics-based shading reconstruction for intrinsic image decomposition. *Computer Vision and Image Understanding* 205 (2021), 103183.

[5] Sean Bell, Kavita Bala, and Noah Snavely. 2014. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)* 33, 4 (2014), 1–12.

[6] Nicolas Bonneel, Balazs Kovacs, Sylvain Paris, and Kavita Bala. 2017. Intrinsic decompositions for image editing. In *Computer Graphics Forum*, Vol. 36. Wiley Online Library, 593–609.

[7] Nicolas Bonneel, Balazs Kovacs, Sylvain Paris, and Kavita Bala. 2017. Intrinsic Decompositions for Image Editing. *Computer Graphics Forum (Eurographics State of The Art Report)* (2017).

[8] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. 2012. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision (ECCV)*. Springer.

[9] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. 2011. Learning Photographic Global Tonal Adjustment with a Database of Input / Output Image Pairs. In *Computer Vision and Pattern Recognition (CVPR)*.

[10] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. 2014. Describing Textures in the Wild. In *Computer Vision and Pattern Recognition (CVPR)*.

[11] Blender Online Community. 2018. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam. http://www.blender.org

[12] Partha Das, Sezer Karaoglu, and Theo Gevers. 2022. PIE-Net: Photometric Invariant Edge Guided Network for Intrinsic Image Decomposition. In *Computer Vision and Pattern Recognition (CVPR)*.

[13] Sylvain Duchêne, Clement Riant, Gaurav Chaurasia, Jorge Lopez-Moreno, Pierre-Yves Laffont, Stefan Popov, Adrien Bousseau, and George Drettakis. 2015. Multi-view intrinsic images of outdoors scenes with an application to relighting. *ACM Transactions on Graphics* (2015), 16.

[14] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. 2018. Revisiting deep intrinsic image decompositions. In *Computer Vision and Pattern Recognition (CVPR)*.

[15] Ruth C Fong and Andrea Vedaldi. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *International Conference on Computer Vision (ICCV)*. 3429–3437.

[16] Elena Garces, Adolfo Munoz, Jorge Lopez-Moreno, and Diego Gutierrez. 2012. Intrinsic images by clustering. In *Computer graphics forum*, Vol. 31. Wiley Online Library, 1415–1424.

[17] Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. 2019. Towards Automatic Concept-based Explanations. In *Neural Information Processing Systems (NIPS)*.

[18] Tom Goldstein and Stanley Osher. 2009. The split Bregman method for L1-regularized problems. *SIAM journal on imaging sciences* 2, 2 (2009), 323–343.

[19] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[20] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. 2009. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *International Conference on Computer Vision (ICCV)*. IEEE, 2335–2342.

[21] Dmitry Kazhdan, Botty Dimanov, Mateja Jamnik, Pietro Liò, and Adrian Weller. 2020. Now you see me (CME): concept-based model extraction. *arXiv preprint arXiv:2010.13233* (2020).

[22] Dmitry Kazhdan, Botty Dimanov, Helena Andres Terre, Mateja Jamnik, Pietro Liò, and Adrian Weller. 2021. Is disentanglement all you need? comparing concept-based & disentanglement approaches. *arXiv preprint arXiv:2104.06917* (2021).

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

[24] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *International Conference on Machine Learning*. PMLR, 5338–5348.

[25] Balazs Kovacs, Sean Bell, Noah Snavely, and Kavita Bala. 2017. Shading annotations in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6998–7007.

[26] Vivek Kwatra, Mei Han, and Shengyang Dai. 2012. Shadow removal for aerial imagery by information theoretic intrinsic image analysis. In *2012 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 1–8.

[27] MCJJ L EH. 1971. Lightness and retinex theory. *J. Opt. Soc. Am.* 61, 1 (1971), 1–11.

[28] Pierre-Yves Laffont, Adrien Bousseau, and George Drettakis. 2012. Rich intrinsic image decomposition of outdoor scenes from multiple views. *IEEE transactions on visualization and computer graphics* 19, 2 (2012), 210–224.

[29] Edwin H Land and John J McCann. 1971. Lightness and retinex theory. *Josa* 61, 1 (1971), 1–11.

[30] Yu Li and Michael S Brown. 2014. Single image layer separation using relative smoothness. In *Computer Vision and Pattern Recognition (CVPR)*.

[31] Zhengqi Li and Noah Snavely. 2018. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *European Conference on Computer Vision (ECCV)*.

[32] Zhengqi Li and Noah Snavely. 2018. Learning intrinsic image decomposition from watching the world. In *Computer Vision and Pattern Recognition (CVPR)*.

[33] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris B. Kotsiantis. 2021. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* 23 (2021).

[34] Xiaopei Liu, Liang Wan, Yingge Qu, Tien-Tsin Wong, Stephen Lin, Chi-Sing Leung, and Pheng-Ann Heng. 2008. Intrinsic colorization. In *ACM SIGGRAPH Asia 2008 papers*. 1–9.

[35] Yunfei Liu, Yu Li, Shaodi You, and Feng Lu. 2020. Unsupervised learning for intrinsic image decomposition from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3248–3257.

[36] Lukas Murmann, Michael Gharbi, Miika Aittala, and Fredo Durand. 2019. A Multi-Illumination Dataset of Indoor Object Appearance. In *2019 IEEE International Conference on Computer Vision (ICCV)*.

[37] Takuya Narihira, Michael Maire, and Stella X Yu. 2015. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *International Conference on Computer Vision (ICCV)*. 2992–2992.

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Neural Information Processing Systems (NIPS)*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[39] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[40] Saurabh Saini and PJ Narayanan. 2019. Semantic hierarchical priors for intrinsic image decomposition. *arXiv preprint arXiv:1902.03830* (2019).

[41] Saurabh Saini and P. J. Narayanan. 2018. Semantic Priors for Intrinsic Image Decomposition. In *British Machine Vision Conference (BMVC)*.

[42] Saurabh Saini, Parikshit Sakurikar, and P. J. Narayanan. 2016. Intrinsic image decomposition using focal stacks. In *Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*.

[43] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296* (2017).

[44] Anirban Sarkar, Deepak Vijaykeerthy, Anindya Sarkar, and Vineeth N. Balasubramanian. 2022. A Framework for Learning Ante-hoc Explainable Models via Concepts. *Computer Vision and Pattern Recognition (CVPR)* (2022), 10276–10285.

[45] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[46] Li Shen, Ping Tan, and Stephen Lin. 2008. Intrinsic image decomposition with non-local texture cues. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

[47] Mike Wu, Michael Hughes, Sonali Parbhoo, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. 2018. Beyond sparsity: Tree regularization of deep models for interpretability. In *AAAI conference on artificial intelligence*, Vol. 32.

[48] Julian Zaidi, Jonathan Boilard, Ghyslain Gagnon, and Marc-André Carbonneau. 2022. Measuring Disentanglement: A Review of Metrics. *IEEE transactions on neural networks and learning systems* PP (2022).

[49] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Computer Vision and Pattern Recognition (CVPR)*.

[50] Tinghui Zhou, Philipp Krahenbuhl, and Alexei A Efros. 2015. Learning data-driven reflectance priors for intrinsic image decomposition. In *International Conference on Computer Vision (ICCV)*.