IDTransformer: Transformer for Intrinsic Image Decomposition

Partha Das[1,3] Maxime Gevers[2,3] Sezer Karaoglu[1,3] Theo Gevers[1,3]

University of Amsterdam, The Netherlands[1] Concordia University, Canada[2]
3DUniversum, Amsterdam, The Netherlands[3]

Abstract

The aim of intrinsic image decomposition (IID) is to recover reflectance and the shading from a given image. As different combinations are possible, IID is an under constrained problem. Previous approaches try to constrain the search space using hand crafted priors. However, these priors are based on strong imaging assumptions and fall short when these do not hold. Deep learning based methods learn the problem end-to-end from the data. But these networks lack any explicit information about the image formation model.

In this paper, an IID transformer approach (IDTransformer) is proposed by learning photometric invariant attention, derived from the image formation model, integrated in the transformer framework. The combination of invariant features in both a global and local setting allows the network to not only learn reflectance transitions, but also to group similar reflectance regions, irrespective of the spatial arrangement. Illumination and geometry invariant attention is exploited to generate the reflectance map, while illumination invariant and geometry variant attention is used to compute the shading map.

Enabling physics-based explicit attention allows the network to be trained on a relatively small dataset. Ablation studies show that adding invariant attention improves the performance. Experiments on the Intrinsic In the Wild dataset shows competitive results with competing methods. The project page with the code is available at https://morpheus3000.github.io/IDTransformer.web/.

1. Introduction

The apparent colour of an object can be defined as the combination of the object's material colour (reflectance/albedo) and the geometry and scene illumination (shading). The inversion of this process, where the reflectance and the shading are recovered from a given image, is defined as Intrinsic Image Decomposition (IID). The use of the separated components are beneficial to downstream tasks like object recognition [21], semantic segmentation [5], geometry estimation [22] or object recolouring [34] and relighting [46]. However, as only the image is given, the IID problem is under-constrained. Previous approaches try to solve the IID problem with explicit priors like associating gradient change patterns to reflectance and shading [26], piece-wise consistency, or reflectance parsimony [3, 43]. Different modalities of explicit priors are also explored to integrate implicit physical information in the decomposition process, like depth [2] and textures [20]. These methods enforce assumptions about the real world and hence may fall short when the assumptions are violated. On the other hand, various image formation based invariants [19, 21] are explored for tasks like colour invariant pose estimation [37] and object recognition [19].

Also CNN-based methods are proposed in combination with large datasets. This approach mitigates the need for explicit priors and enables end-to-end learning directly from the image data. However, CNN receptive fields are local and they are unable to learn global relationships, which are useful cues for the problem of IID (like reflectance changes). Transformers [48] address the problem of local receptive fields and are successfully applied to various vision tasks like object detection [11, 16], semantic segmentation [31] and dense depth prediction [39]. With this new attention paradigm, global relationships between input tokens (e.g. image patches) are learned. This is beneficial for the problem of IID where global cues may provide important cues to the network allowing to enforce fundamental IID constraints like reflectance consistency and shading smoothness.

Priors allow for the use of explicit physics-based image formation models, while deep learning based methods enable the use of flexible models to learn from image datasets

Contact: p.das@uva.nl/partha.das.pdt@gmail.com

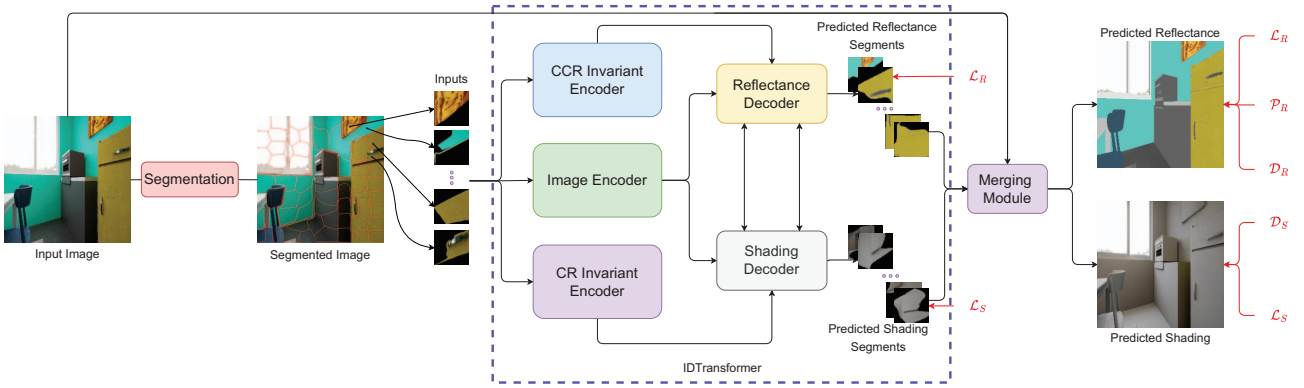

Figure 1. Overview of the proposed IDTransformer. The network takes an image as input. Then, the image is segmented into regions, or segments, of homogeneous reflectance. These segments are then fed into the global and local attention layers. The attention layers exploit physics based invariants, CCR & CR. The (homogeneous reflectance) pixels within the segments are compared. In addition, similarity is computed among the segments. This results in a latent space where similarly coloured segments are closer. The encoded features are then fed into component specific joint decoders. The predicted segments are collected by a learnable merger module that predicts the final reflectance and shading. The segment predictor is trained using the MSE loss ($\mathcal{L}_R$ and $\mathcal{L}_S$) between the predicted and ground truth segments. The final prediction is trained using the MSE loss on the full reconstructed reflectance and shading, and the perceptual loss ($\mathcal{P}_R$) and dissimilarity metric on the reflectance and the shading maps ($\mathcal{D}_R$ and $\mathcal{D}_S$).

(i.e. recovering the image formation process by learning from the data). Attempts are made to combine the frameworks [14, 15]. The authors propose to combine invariance with the flexibility of a deep learning model. However, the approach uses invariants as a pre-processed input to the network. This limits the expressibility of invariant priors, since the cues provided by neighbourhoods are selected during the preprocessing step. This also makes the network unable to exploit any useful global cues that these invariant features should have. So far, transformers have ignored the use of explicit physics-based image formation models as a guidance in their architecture. Further, the dot-product employed by transformers are instance specific, i.e., the attention output depends on the incoming instance.

Therefore, in this paper, we propose an IID transformer approach (IDTransformer) by learning photometric invariant attention, derived from an image formation model, integrated in the transformer framework. Image patches are given as input by segmenting the image into regions (i.e. superpixels) containing homogeneous reflectance using a non-learnable segmentation method [47, 1]. Hence, the superpixel segments are composed of approximately homogeneous reflectance (albedo) regions but these regions may vary in illumination conditions (shading). In this way, the segments have approximately uniform albedo which is beneficial as a starting point to process them by the IDTransformer to yield intrinsic image decomposition. In fact, the segments are combined based on a physics-based invariant attention model. Pixel comparisons within the segments allow learning a local comparison and allow for refining the reflectance boundaries that may not necessarily coincide with the segment boundaries due to photometric effects. Then, comparing each of these segments with other segments allows the IDTransformer network to learn similar segments. Fig 1 visualises the proposed architecture. In summary, the contributions of this paper are as follows:

- Segmentation is used to divide the image into regions (i.e. superpixels) containing homogeneous reflectances. Segments are used as tokens in the Transformer.
- An IID transformer approach (IDTransformer) is proposed based on physics based invariant attention mechanism for global and local attention.
- Our approach enables a new avenue of research to integrate the image formation process (priors) into (flexible) Transformer models.

2. Related works

The seminal work of [26] pioneered the use of intrinsic specific priors, using image gradients as a guide to the IID problem. For example, larger gradients are associated with reflectance changes and smaller ones correspond to shading changes. Other priors are also explored such as texture [50], reflectance sparsity [20] and depth [27]. [3] combines several constraints such as a piece-wise constant reflectance and a smooth shading assumption to guide the IID problem. Additional modalities such as infrared images [12] and surface normals [23] are explored to further constrain the search space. More implicit constraints such as user annotated priors [7, 36, 9] and multi-frame inputs [49, 33] are also proposed. However, these priors are based on strong imaging assumptions and fall short if these do not hold.

Deep learning approaches are proposed [35, 44] where the IID problem is modelled as an end-to-end process parameterised by deep neural networks. This is made possible by large datasets [7, 29, 10, 53, 40] which model the IID problem as a data distribution. Edge maps [18], depth [17, 25] and surface normals [32] are studied as additional inputs to the network to constrain the search space and guide the decomposition problem. [4] extends the image formation model to include illumination effects such as shadows and ambient lighting. However, these methods are based on CNNs, focusing on local receptive fields. Therefore, global cues are not taken into account unless explicitly supported by the training data.

Transformers [48], on the other hand, focus on learning global relations. They are successfully applied to various computer vision tasks such as object detection [11], depth prediction [39], and semantic segmentation [31]. Furthermore, the global attention property of a transformer is suitable for the IID problem. [52] applies transformers to the IID problem. However, they rely on very large datasets [30] for the transformer to learn the relationship, without any explicit physics-based guidance or formulation. The network also eschews local relations in favour of only global ones.

On the other hand, several physics-based invariants are proposed, such as Colour Ratios [19] (CR) and Cross Colour Ratios [21] (CCR). These are illumination and geometry invariant neighbourhood descriptors that are useful cues for reflectance recovery. Recent work [6, 14, 13, 15] explores the use of such invariants for IID. However, they use them as an input prior, which only takes local neighbourhoods into account. This limits the expressibility of the descriptors.

Superpixels are also explored for IID [24, 45]. [24] uses superpixels on aerial hyperspectral imagery outdoors, which has more globally consistent illumination, compared to an indoor scene where there might be strong illumination effects (shadows, etc.). [45] uses superpixels with spherical harmonics (SH) to model reflectance and shading, respectively, applying only to single objects. The meanshift based superpixels is also unstable for overlapping colours or shadows. It also enforces uniformity per super pixel, resulting in reflectance leakage in shading for finely textured surfaces. In contrast, this proposed work uses local and global attention to work around such a deficiency.

In contrast to existing work, in this paper, an integrated approach is proposed combining (1) physics-based invariance, (2) local learning of CNNs, and (3) global learning of transformers. These components are unified in a new framework using a novel attention mechanism. Specifically, the transformer framework is used to propose a new illumination and/or geometry invariant attention mechanism integrating both local and global cues together with physics-based cues. This allows the IID problem to be formulated as an end-to-end paradigm.

	Invariances	
Algorithm	Illumination	Geometry
Colour Ratio [19] (CR)	✓	×
Cross Colour Ratio [21] (CCR)	✓	✓

Table 1. Photometric invariant properties for CR and CCR. Due to a flat geometry assumption, CR is geometry variant when objects are curved. CCR is robust to this and hence fully invariant to both illumination and geometry. CCR is useful as a reflectance descriptor, while CR can provide geometric cues. Comparing CR with CCR provides useful cues to recover the shading.

3. Methodology

3.1. Invariant Descriptors

Consider the Lambertian image formation model [42]:

$$I = m(\vec{n}, \vec{l}) \int_{\omega} e(\lambda)\ \rho_b(\lambda)\ f(\lambda)\ \mathrm{d}\lambda\ , \tag{1}$$

where, I is the image; $\vec{n}$ represents the surface normal and $\vec{l}$ the illuminant direction, forming the parameters of m which is a function of the object geometry and illuminant interaction. λ is the incoming wavelength of the illuminant ω; e is the spectral power distribution of the illuminant, ρ_b the object reflectance (albedo), and f is the spectral camera sensitivity function.

In discrete RGB pixel domain, we obtain:

$$C_{p_1} = m(\vec{n_{p_1}},\ \vec{l_{p_1}})\ e^{C_{p_1}}(\lambda)\ \rho^{C_{p_1}}(\lambda)\ , \tag{2}$$

where p_1 denotes the pixel and C is the RGB channel.

Based on the image formation modelled by (2), invariant descriptions such as Colour Ratios (CR) are proposed by [19]. However, the method assumes flat surfaces, which makes CR illumination invariant, but geometry variant. Instead, [21] proposes the Cross Colour Ratio (CCR) which is both illumination and geometry invariant. CCR provides reflectance descriptors that is useful for recovering the albedo. Table 1 provides an overview of the invariants.

These invariants are computed over neighbouring pixels. Applying these invariants to local neighbours will result in edges (i.e. transitions). However, using longer range (global) neighbours provides comparison across descriptors. Previous methods, exploiting invariants for IID [6, 13, 14, 15] used these as pre-processed priors but only focusing on local neighbourhoods. In contrast, in this paper, invariants are modelled as a part of the learning process, whereby the invariance is exploited both locally and globally as an attention mechanism.

3.2. Self attention

Consider segments ($\sum_{i=N}^{1} \mathcal{S}_i$) obtained by a superpixel segmentation algorithm [47, 1]. Each $\mathcal{S}$ consists of pixels with approximately homogeneous reflectances. Hence, each segment contains homogeneous reflectance (albedo) but may vary in illumination conditions (shading). However, due to various photometric effects, the segments may not always correspond to uniform reflectances. Learning a relationship between the pixels within the segments may ensure that they have the same reflectance. At the same time, by learning relationships between the segments themselves in a one-to-many comparison, global cues are also exploited. This overcomes the locality limitation of previous approaches.

3.3. Invariant Attention

Standard transformers divide an image into arbitrary patches with w, h as the width and height of the patch. In this paper, patches correspond to segments i.e. superpixels with approximately uniform albedo (reflectance) but with possibly varying shading (illumination). Hence, segments form the basis of the intrinsic decomposition process. Specifically, segments are separated by using a tightly fitting bounding box ($\mathcal{S}_{w,h}$) and used as input to the transformer. Then, these segments are unrolled and converted into queries (Q), keys (K) and values (V). An attention score is obtained as follows:

$$\mathcal{A}(Q, K) = Softmax(\frac{QK^T}{\sqrt{d_k}}), \tag{3}$$

where d_k is the embedding dimension and $\mathcal{A}$ is the alignment score. The final attention is obtained using a dot product with the alignment score and V, defined by:

$$attn(\mathcal{A}, V) = \mathcal{A} \cdot V. \tag{4}$$

where Q and K are the transformed image pixels. For IID, the network must learn to distinguish between reflectance and shading properties completely from the data itself. The dot product with V (4) then aligns the values from (3) to the most related single input. However, the dot product only depends on the instance of the input values. To remedy this, two methods are proposed: (1) The interaction between Q and K is replaced by an invariance function. This allows the network to use individual invariant features across n neighbours to obtain an initial attention. In this way, the image formation model is integrated in the attention. (2) The final dot product is replaced by a learnable layer that takes into account the invariant features of n neighbours, while also being instance-independent.

$\mathcal{S}_{w,h}$ is unrolled into a vector $\vec{V} \in \mathbb{R}^n$, where $n = w \times h$, denoting all the pixels within the segments. A $n \times n$ matrix is created from the vector, where each element of the matrix denotes an one-to-many relationship for a given pixel:

$$\mathbb{N} = \vec{V} \times \vec{V}^T, \tag{5}$$

where $\mathbb{N} \in \mathbb{R}^{n \times n}$ and the diagonal of the matrix represent pixels paired with themselves. This matrix models both short-range and long-range relationships. Various invariants (such as CCR and CR) are applied across all pixel pairs as follows:

$$\mathbb{C} = \mathbb{I}_{inv}(\mathbb{N}_{ij}), \tag{6}$$

where $\mathbb{C} \in \mathbb{R}^{n \times n}$, $i = 1, n$ & $j = 1, n$ and $\mathbb{I}_{inv}$ is a two neighbourhood invariance function. This results in a diagonal matrix with invariances on the upper right. The bottom right are the same invariances with their signs reversed. Each element in the row of the matrix provides the invariance with all other pixels.

Given a pair of pixels, not all neighbourhoods may be equally important. In fact, for cases such as the same reflectance region, some of the immediate scales of the local neighbourhood can be merged together as they would have the same descriptor. Conversely, a local neighbourhood should have the same invariances for the same reflectance region. This information is incorporated into the model in the form of a dynamic weighting of the different neighbourhoods. This allows the network to give higher weights to those invariances that contribute more to the recovery of IID. This is modelled as follows:

$$\mathbb{W} = \mathcal{F}(x, y), \tag{7}$$

where $\mathcal{F}$ is a function parameterised by a linear layer that compares the neighbourhood of a pixel to output a weight indicating how important it is compared to the other pixels. Here, x and y are the two candidate pixels in a neighbourhood whose weight is to be calculated. And $\mathbb{W} \in \mathbb{R}^{n \times n}$.

The invariances and weights are combined to obtain the final attention for the pixels as follows:

$$\mathcal{A}_s = \mathbb{W} \odot \mathbb{C}, \tag{8}$$

where $\mathcal{A}_s \in \mathbb{R}^{n \times n}$ is the attention for the n-th pixel of segments $\mathcal{S}_{w,h}$, depending on $n-1$ neighbours for each row. The final attention for each pixel is obtained by summing along each row:

$$\vec{\mathcal{A}} = \sum_{j=1}^{n} \mathbb{C} \odot \mathcal{F}(x, y) \tag{9}$$

where $\vec{\mathcal{A}} \in \mathbb{R}^n$ and j is the column index. One such attention is obtained for each segment, the final local attention yields $\vec{\mathcal{A}}_l \in \mathbb{R}^{b \times s \times n \times c}$, where b is the image batch, s the number of segments, c the feature channels and n the unrolled pixels.

Attention within a segment results in a local attention map. To enable global attention between segments, each of the segments is tokenised. This results in a vector of $\mathcal{B}_g \in \mathbb{R}^{b\times s\times 1\times 1\times c}$, where each segment is reduced to a single token. This allows one-to-many attention within the segments on a global scale. $\mathcal{S}_g$ then similarly runs through (9) to obtain the global attention $\vec{\mathcal{A}}_g \in \mathbb{R}^{b\times s\times 1\times c}$. The final global and local attention is obtained by summing up $\vec{\mathcal{A}}_g$ and $\vec{\mathcal{A}}_l$.

3.4. Segment-based Learning

The attention-enhanced segments contain information about the similarity not only between the pixels they contain, but also between the other segments. This allows the network to bring the segments that are perceptually similar closer together, while also doing the same within the segments in the feature space. This segment-based learning allows the network to tessellate a given image into non-overlapping regions. For example, in an indoor scene, two opposite walls might belong to the same reflectance values. But they might not always be next to each other in an image. There could be furniture or objects of different colours between them. In traditional learning, where the whole image is captured, the features of non-wall objects may "leak" into the features associated with the wall colour. Similarly, simply dividing the image into blocks could result in a significant overlap of non-wall objects. Dividing the image into superpixel segments avoids this. Similarly colored segments are therefore placed closer together in the feature space than dissimilar ones. To enforce this decoupling, it is proposed to learn only at the patch level, rather than recombining the patches before passing them to the decoder.

More concretely, the input image is first segmented using a superpixel segmentation method (in the experiments we use SLIC [1]) to obtain segments with approximately uniform reflectances. These segments are non-overlapping and are more likely to be separate object boundaries, compared to other pixel-wise grouping or a standard division into rectangular patches. Since the input and output to the network are segments, the output segments can be merged to obtain the full image. However, the predicted intrinsic component images, due to non-overlapping segments, will not have a smooth transition between regions. This is remedied using a merger network, which takes the segmented images and computes a complete image.

3.5. Network architecture

The network consists of a transformer encoder and a component-specific convolutional decoder. The input to the network are the segments (superpixels) of the input image. These segments are concatenated, resulting in a volume $\mathcal{S} \in \mathbb{R}^{b\times s\times w\times h\times c}$, where b is the batch size of the input image, s is the number of segments, w, h and c are the width, height and channels of the segments respectively. An image encoder based on the standard transformer attention also provides global and local colour cues for the invariant paths. The components are described below.

Invariant Attention Layer: The proposed attention (9) is implemented following the architecture of a transformer. However, it uses the invariant functions for neighbourhood relations. The similarity function is changed to a learned function parameterised by a neural network. This allows the network to dynamically weight the different ranges of invariants. Computing the neighbourhood for every pixel in an image is intractable. Therefore, the input image is first divided into patches of size $n \times n$ and the neighbours are computed across the patches. The patches are first tokenised and passed through a layer norm, followed by the attention described by (9).

Encoder: The input image is passed through different paths to obtain component-specific features. The illumination and geometry invariance of the CCRs are exploited for reflectance. Similarly, the illumination invariance but geometry variance of the CRs is exploited for shading. Separate pathways are provided for both types of attention, allowing the network to learn specialised feature spaces. For each encoder, the invariant attention layer is repeated 3 times sequentially. Separate pathways are created for local and global attention. The local and global attentions are then summed to obtain the final attention. The image encoder pathway uses a standard transformer layer with the same 3 block configuration as the invariant encoder to maintain spatial parity. The image encoder provides a colour and illumination function to support the invariant features to recover the intrinsic components. Three separate encoders are used in the network: (1) CCR encoder ($\mathcal{F}_{CCR}$), (2) CR encoder ($\mathcal{F}_{CR}$), and (3) image encoder ($\mathcal{F}_{img}$). All three encoders take the input image and build an independent feature space based on attention type.

Decoder: The bottlenecks $\mathcal{F}_{CCR}$ and $\mathcal{F}_{CR}$ provide invariant specific features, while $\mathcal{F}_{img}$ provides colour and illumination features independently. However, $\mathcal{F}_{CCR}$ is closely related to reflectance change features, while $\mathcal{F}_{CR}$ encodes some geometric information. Thus, $\mathcal{F}_{img}$ and $\mathcal{F}_{CCR}$ are concatenated for the reflectance decoder pathway, while $\mathcal{F}_{img}$ and $\mathcal{F}_{CR}$ are concatenated for the shading decoder pathway. Furthermore, according to the Lambertian image formation model, reflectance and shading are mutually exclusive, so there are useful contrast cues that could be used by the component to better enforce the image formation model. As a result, an interconnected decoder

module is added on top of the bottlenecks. Each of the decoder blocks consists of a convolution, followed by a batch norm and a Relu nonlinearity layer. The interconnection between the decoder paths, one each for reflectance and shading, allows the network to learn the intrinsic components jointly. In total, 5 decoder blocks are stacked for each of the decoder paths, resulting in reflectance and shading. The inputs to the decoders are the segments $\mathcal{S} \in \mathbb{R}^{b \times s \times w \times h \times c}$. The decoders output the appropriate segments for each component. This allows the network to output segments with similar component properties, while also fine-tuning the predictions within the segments, even if the segments do not correspond to the true reflectance boundary.

3.6. Losses

The network is trained in a supervised manner. The predicted reflectance and shading are compared with the ground truth IID components using an MSE loss. The gt reflectance and shading are decomposed into segments using the same masks that are used to process the input image. A reconstruction loss is also included to regularise the predicted components. This provides the dense segment-wise monitoring for the IID components, which is collected as the IID loss ($\mathcal{L}_{iid}$). The network is trained using the Adam optimiser with a learning rate of $1e-5$ and beta of 0.5 and 0.999. The network is trained for 300 epochs until convergence.

For the learnable merger module, the same component wise losses are trained, namely, the MSE loss for the reconstructed reflectance and shading. Additionally, to encourage perceptually consistent and sharper textures, a perceptual and dssim loss are included as $\mathcal{P}_R$, $\mathcal{D}_R$ and $\mathcal{D}_S$, respectively. The final loss term to minimise for the merger module thus becomes:

$$\mathcal{L} = \mathcal{L}_{iid} + \lambda_p * \mathcal{P}_R + \lambda_d * (\mathcal{D}_R + \mathcal{D}_S) \qquad (10)$$

where λ_p and λ_d are weighting terms for the perceptual and dissimilarity losses. They are empirically set to 0.05 and 0.4, respectively. The network is trained for 400 epochs, with a learning rate of $2e-4$ and the Adam optimiser.

4. Experiments & results

4.1. Datasets

The proposed network needs dense supervision. Hence, a dataset with dense reflectance and shading ground truth is required. While large scale datasets [29, 30, 40] do exist, they are often not realistic. The dataset proposed by [15], although smaller, consists of realistic and physically based ray traced scenes. The dataset provides 5791 samples with the corresponding dense reflectance and shading ground truth. The train and test sets consist of 4632 and 1159 samples respectively. For a real-world case, our method is also finetuned on the IIW [7] dataset. This dataset consists of sparse human judgement for the reflectance and is only used as a finetuning experiment to show the potential of the proposed attention mechanism in real-world, unconstrained settings.

	Reflectance			Shading		
	MSE	LMSE	DSSIM	MSE	LMSE	DSSIM
Img only	0.0271	0.0216	0.1657	0.0268	0.0228	0.1781
Img + CCR	0.0044	0.0063	0.0463	0.0027	0.0044	0.0484
Img + CCR + CR	**0.0034**	**0.0054**	**0.0444**	**0.0017**	**0.0033**	**0.0437**

Table 2. Ablation study on the different types of invariant attention. Default transformer attention performs the worse. Adding the component specific invariant attentions and pathways improves the performance. This validates the proposed invariant attention's usefulness.

4.1.1 Ablation Study

In this experiment, the influence of the proposed attention is studied. The CCR and CR invariant attention are first disabled, resulting in a baseline transformer network that only has a single encoder with the reflectance and shading as the output. Following this, the CCR invariant encoder is enabled to see the influence of only the reflectance descriptors. Finally, the CR invariant encoder is also added, arriving at the proposed IDTransformer architecture. All the experiments are performed on the same dataset and evaluated with the same test split. The results are presented in table 2. A visual comparison with the proposed architecture and image only configuration is shown in fig 2.

From the results, it is shown that without the invariant attention, the performance is decreased. Adding the invariant attention allows the network to improve performance on both the intrinsic components. Further, adding the CR invariant further improves the performance over only using the CCR invariance. This is because while the CCR provides useful reflectance cues, the shading is dependent on the reflectance being correct. However, adding the CR invariance on top of the CCR invariance strengthens the shading decomposition. This, through the joint decoder, provides useful cues to the reflectance and vice versa.

4.1.2 Intrinsics In the Wild

In this experiment, the performance of the network on a real-world dataset is given. The network is first pretrained on the synthetic dataset and then finetuned on the IIW dataset for 6 epochs. The IIW dataset comes with sparse annotations. Hence, the original losses used to train the network cannot be used. Instead the ordinal loss is used to finetune on the dataset. Table 3 shows the results. Fig 3 and 4 shows visual comparisons.

The results show that using a segment-based approach already improves over using square patches. While the net-

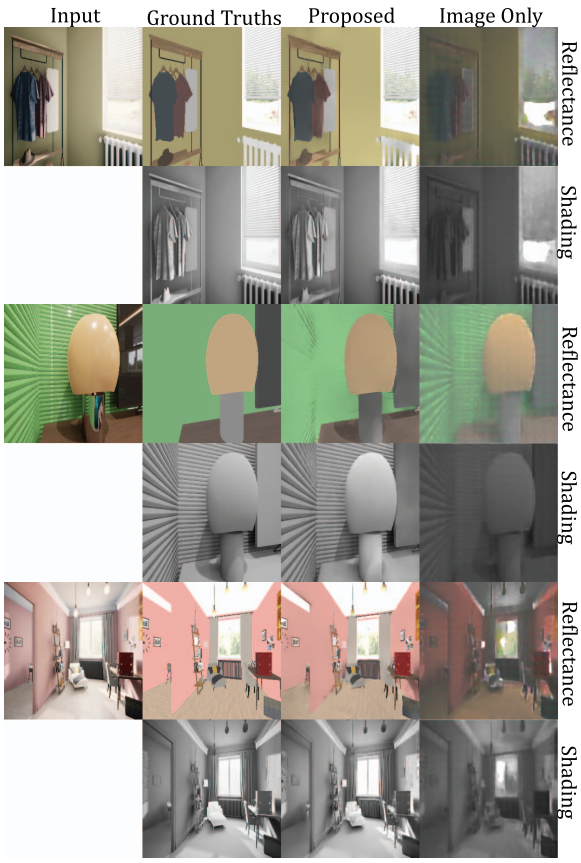

Figure 2. Visual comparison of the proposed method (IDTransformer) against the network consisting of an image-only-transformer. It is shown that the addition of the invariants helps the network to improve the decomposition. For example, in the middle example, adding the invariant results in a flatter reflectance, compared to the image-only-configuration where shwoing illumination pattern leakages in the reflectance.

Methods	WHDR (mean)
Nestmeyer *et al*. [38]	19.5
Bi *et al*. [8]	17.7
Sengupta *et al*. [41]	16.7
Li *et al*. [28]	15.9
CGIntrinsics [29]	14.8
GLoSH [51]	14.6
Fan *et al*. [18]	14.5
SIGNet [15]	13.9
IRISFormer [52]	13.1
Baseline	19.1
Intrinsics Transformer	18.7

Table 3. Baseline comparison for the IIW dataset.

work doesn't achieve SotA performance, the current performance is obtained by using a comparably simpler network without any specialised layers, apart from a modified transformer attention. No purpose built losses are needed either, as compared to other baselines. Moreover, visually, the network is able to distinguish photometric effects better than the other baselines. For example, in the textures in the teacup are better preserved by the proposed network, while artefacts on the corners of the walls in the bedroom is free from discolouration as compared to the baselines.

5. Conclusion

In this paper, a physics-based invariant attention mechanism has been proposed for the task of intrinsic image decomposition. The illumination and geometry invariant property of CCR and illumination invariant and geometry variant property of CR is exploited by the attention to guide the network towards improved intrinsic component recovery. The invariants are also exploited in a global and local stages using a transformer framework to recover the intrinsic components. Finally, a new learnable similarity function has been used to solve the instance specific learning of the dot product used in standard transformer formulations.

An ablation study was performed to show that the addition of the invariants improves the performance of the network. Visually, the network has shown to be able to better disentangle photometric effects, compared to other baselines, while being trained on a smaller dataset. Our approach shows the possibility to integrate the image formation process (priors) into (flexible) Transformer models.

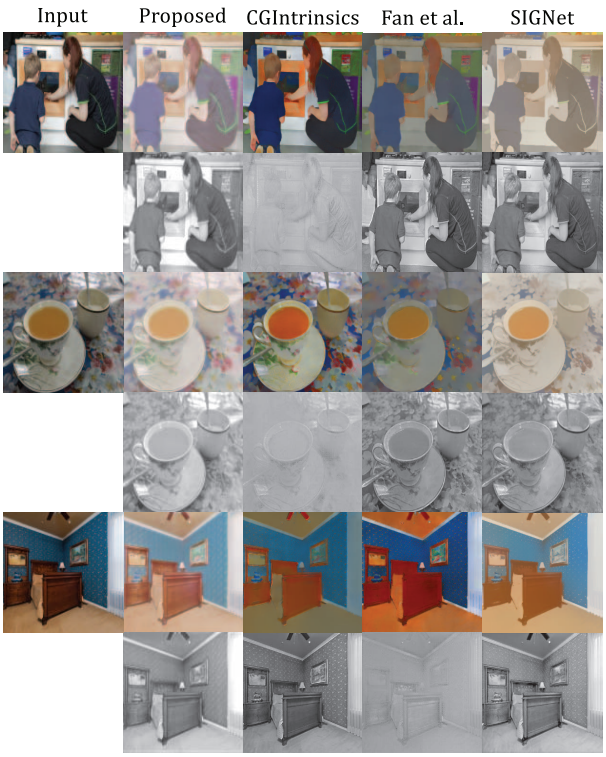

Figure 3. Visual comparison on the IIW dataset. The proposed method is able to predict more consistent reflectance and shading that are closer to the original image colour. Existing methods exhibit strong color biases.

Figure 4. The proposed method is shown to be able to handle illumination transfers properly and preserve the underlying reflectance. Existing methods show textural deficiencies.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012. 2, 4, 5

[2] J. T. Barron and J. Malik. Intrinsic scene properties from a single rgb-d image. In *CVPR*, 2013. 1

[3] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. *IEEE TPAMI*, pages 1670–1687, 2015. 1, 2

[4] Anil S. Baslamisli, Partha Das, Hoang-An Le, Sezer Karaoglu, and Theo Gevers. Shadingnet: Image intrinsics by fine-grained shading decomposition. *IJCV*, 129:2445–2473, 2021. 3

[5] A. S. Baslamisli, T. T. Groenestege, P. Das, H. A. Le, S. Karaoglu, and T. Gevers. Joint learning of intrinsic images and semantic segmentation. In *ECCV*, 2018. 1

[6] Anil S. Baslamisli, Yang Liu, Sezer Karaoglu, and Theo Gevers. Physics-based shading reconstruction for intrinsic image decomposition. *Comput. Vis. and Image Understanding*, pages 1–14, 2020. 3

[7] S. Bell, K. Bala, and N. Snavely. Intrinsic images in the wild. *ACM TOG*, 2014. 2, 3, 6

[8] Sai Bi, Xiaoguang Han, and Yizhou Yu. An l1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition. *ACM TOG*, 34(4), July 2015. 7

[9] N. Bonneel, K. Sunkavalli, J. Tompkin, D. Sun, S. Paris, and H. Pfister. Interactive intrinsic video editing. *ACM TOG*, pages 197:1–197:10, 2014. 2

[10] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012. 3

[11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020. cite arxiv:2005.12872. 1, 3

[12] Ziang Cheng, Yinqiang Zheng, Shaodi You, and Imari Sato. Non-local intrinsic decomposition with near-infrared priors. In *ICCV*, October 2019. 2

[13] Partha Das, Sezer Karaoglu, and Theo Gevers. Intrinsic image decomposition using physics-based cues and cnns. *Computer Vision and Image Understanding*, 223:103538, 2022. 3

[14] Partha Das, Sezer Karaoglu, and Theo Gevers. Pie-net: Photometric invariant edge guided network for intrinsic image decomposition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 19758–19767. IEEE, 2022. 2, 3

[15] Partha Das, Sezer Karaoglu, Arjan Gijsenij, and Theo Gevers. Signet: Intrinsic image decomposition by a semantic and invariant gradient driven network for indoor scenes. *CoRR*, abs/2208.14369, 2022. 2, 3, 6, 7

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1

[17] J. T. Barron E. Shelhamer and T. Darrell. Scene intrinsics and depth from a single image. In *ICCV*, 2015. 3

[18] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf. Revisiting deep intrinsic image decompositions. In *CVPR*, 2018. 3, 7

[19] G. D. Finlayson. Colour object recognition. Master's thesis, Simon Fraser University, 1992. 1, 3

[20] P. V. Gehler, C. Rother, M. Kiefel, L. Zhang, and B. Schölkopf. Recovering intrinsic images with a global sparsity prior on reflectance. In *NeurIPS*, 2011. 1, 2

[21] T. Gevers and A. Smeulders. Color-based object recognition. *PR*, pages 453–464, 1999. 1, 3

[22] Paul Henderson and Vittorio Ferrari. Learning single-image 3d reconstruction by generative modelling of shape, pose and shading. *International Journal of Computer Vision*, 2019. 1

[23] J. Jeon, S. Cho, X. Tong, and S. Lee. Intrinsic image decomposition using structure-texture separation and surface normals. In *ECCV*, 2014. 2

[24] Xudong Jin and Yanfeng Gu. Superpixel-based intrinsic image decomposition of hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 55(8):4285–4295, 2017. 3

[25] S. Kim, K. Park, K. Sohn, and S. Lin. Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In *ECCV*, 2016. 3

[26] E. H. Land and J. J. McCann. Lightness and retinex theory. *J. of Optical Society of America*, pages 1–11, 1971. 1, 2

[27] Kyong Joon Lee, Qi Zhao, Xin Tong, Minmin Gong, Shahram Izadi, Sang Uk Lee, Ping Tan, and Stephen Lin. Estimation of intrinsic image sequences from image+depth video. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *ECCV*, pages 327–340, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. 2

[28] Zhengqin Li, Mohammad Shafiei, R. Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. *CVPR*, pages 2472–2481, 2020. 7

[29] Z. Li and N. Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *ECCV*, 2018. 3, 6, 7

[30] Zhengqin Li, Ting Yu, Shen Sang, Sarah Wang, Mengcheng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh B. Gundavarapu, Jia Shi, Sai Bi, Hong-Xing Yu, Zexiang Xu, Kalyan Sunkavalli, Milo Haan, Ravi Ramamoorthi, and Manmohan Chandraker. Openrooms: An open framework for photorealistic indoor scene datasets. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7186–7195, 2021. 3, 6

[31] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *International Conference on*

Computer Vision and Pattern Recognition (CVPR), 2022. 1, 3

[32] Jundan Luo, Zhaoyang Huang, Yijin Li, Xiaowei Zhou, Guofeng Zhang, and Hujun Bao. Niid-net: Adapting surface normal knowledge for intrinsic image decomposition in indoor scenes. *IEEE Transactions on Visualization and Computer Graphics*, 26(12):3434–3445, 2020. 3

[33] Y. Matsushita, K. Nishino, K. Ikeuchi, and M. Sakauchi. Illumination normalization with time-dependent intrinsic images for video surveillance. *IEEE TPAMI*, pages 1336–1347, 2004. 2

[34] A. Meka, M. Zollhöfer, C. Richardt, and C. Theobalt. Live intrinsic video. *ACM TOG*, 2016. 1

[35] T. Narihira, M. Maire, and S. X. Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *ICCV*, 2015. 3

[36] T. Narihira, M. Maire, and S. X. Yu. Learning lightness from human judgement on relative reflectance. In *CVPR*, pages 2965–2973, June 2015. 2

[37] S. K. Nayar and R. M. Bolle. Reflectance based object recognition. *IJCV*, pages 219–240, 1996. 1

[38] Thomas Nestmeyer and Peter V. Gehler. Reflectance adaptive filtering improves intrinsic image estimation. *CoRR*, abs/1612.05062, 2016. 7

[39] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021. 1, 3

[40] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *International Conference on Computer Vision (ICCV) 2021*, 2021. 3, 6

[41] Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W. Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. *CoRR*, abs/1901.02453, 2019. 7

[42] S. Shafer. Using color to separate reflection components. *Color Research and App.*, pages 210–218, 1985. 3

[43] L. Shen, P. Tan, and S. Lin. Intrinsic image decomposition with non-local texture cues. In *CVPR*, 2008. 1

[44] J. Shi, Y. Dong, H. Su, and S. X. Yu. Learning non-lambertian object intrinsics across shapenet categories. In *CVPR*, 2017. 3

[45] Jian Shi, Yue Dong, Xin Tong, and Yanyun Chen. Efficient intrinsic image decomposition for rgbd images. In *Proceedings of the 21st ACM Symposium on Virtual Reality Software and Technology*, VRST '15, page 17–25, New York, NY, USA, 2015. Association for Computing Machinery. 3

[46] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. *CoRR*, abs/1704.04131, 2017. 1

[47] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013. 2, 4

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. 1, 3

[49] Y. Weiss. Deriving intrinsic images from image sequences. In *ICCV*, 2001. 2

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

[51] Hao Zhou, Xiang Yu, and David W. Jacobs. Glosh: Global-local spherical harmonics for intrinsic image decomposition. In *ICCV*, October 2019. 7

[52] Rui Zhu, Zhengqin Li, Janarbek Matai, Fatih Porikli, and Manmohan Chandraker. Irisformer: Dense vision transformers for single-image inverse rendering in indoor scenes. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022. 3, 7

[53] Yongjie Zhu, Jiajun Tang, Si Li, and Boxin Shi. Derendernet: Intrinsic image decomposition of urban scenes with shape-(in)dependent shading rendering. *CoRR*, abs/2104.13602, 2021. 3