IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Digital Object Identifier 10.1109/ACCESS.2024.0429000

G-SemTMO: Tone Mapping with a Trainable Semantic Graph

ABHISHEK GOSWAMI[1], ERWAN BERNARD[2], ARU RANJAN SINGH[1], WOLF HAUSER[2], FREDERIC DUFAUX[3], (Fellow, IEEE) and RAFAL MANTIUK[4], (Member, IEEE)

[1]University of Warwick, UK (e-mail: {abhishek.goswami, aru.singh}@warwick.ac.uk)
[2]DxO Labs, France (e-mail: {ebernard, whauser}@dxo.com)
[3]L2S, Université Paris-Saclay, CNRS, CentraleSupélec, France (e-mail: frederic.dufaux@centralesupelec.fr)
[4]University of Cambridge, UK (e-mail: rkm38@cam.ac.uk)

Corresponding author: Abhishek Goswami (e-mail: abhishek.goswami@warwick.ac.uk).

ABSTRACT A Tone Mapping Operator (TMO) is required to render images with a High Dynamic Range (HDR) on media with limited dynamic capabilities. TMOs compress the dynamic range with the aim of preserving the visually perceptual cues of the scene. Previous literature has established the benefits of TMOs being semantic-aware and understanding the content in the scene to preserve cues better. Expert photographers analyze the semantic and contextual information of a scene and decide tonal transformations or local luminance adjustments. This process can be considered a manual analogy to tone mapping. In this work, we draw inspiration from an expert photographer's approach and present a Graph-based Semantic-aware Tone Mapping Operator, G-SemTMO. We leverage semantic information as well as the contextual information of the scene in the form of a graph capturing the spatial arrangements of its semantic segments. Using Graph Convolutional Network (GCN), we predict intermediate parameters called Semantic Hints and use these parameters to apply tonal adjustments locally to different semantic segments in the image. In addition, we also introduce LocHDR, a dataset of 781 HDR images tone mapped manually by an expert photo-retoucher with local tonal enhancements. We conduct ablation studies to show that our approach, G-SemTMO[a], can learn both global and local tonal transformations from a pair of input linear and manually retouched images by leveraging the semantic graphs and produce better results than both traditional and learning based TMOs. We also conduct ablation experiments to validate the advantage of using GCN.

[a]Code and dataset to be published with the final version of the manuscript.

INDEX TERMS Deep Learning, Graph Convolutional Networks, Semantic Awareness, HDR Tone Mapping Operators.

I. INTRODUCTION

TONE mapping operators compress the dynamic range of an image, trying to preserve its aesthetic and visual quality. The problem of finding a balance between dynamic range compression and aesthetic quality predates digital image processing. Renaissance painters created high-fidelity paintings with a limited dynamic range of pigments while maintaining the contextual cues of the scene. In the era of analog photography, photo-retouchers reproduced high dynamic range content on limited dynamic media by locally adjusting exposure and contrast [1]. Artists naturally took regions of semantic similarity and image saliency into account in order to reproduce the visual cues of the scene. Therefore, while a TMO maps the luminance values from a linear image to its output, it helps if it is also aware of the content in the scene. The importance of TMOs being aware of the semantic context of a scene has been established in literature [2]. The research problem addressed in this paper lies in the question, how can we use contextual semantic information explicitly in the tone mapping pipeline? We hypothesize that ideally a TMO should analyse an image like an expert photographer, generate an abstract understanding of the scene and modify the image locally based on the abstract semantic information. *How do photographers analyse a scene while retouching?* Parsing a scene is essential for aesthetically modifying an image. Learning-based semantic segmentation networks assign fine-grained labels to pixels and generate a semantic map for an image [3]. Unlike fine semantic segmentation, photographers parse scenes at a coarser level. We conducted interviews with expert photographers and photo retouchers to understand the tonal adjustment process. It is a two-step pro-

cess. First, the expert identifies photographically important objects (semantic classes or labels) in the scene. A coarse map is thus created with individual segments signifying regions of semantic similarity. For each region of interest the expert considers its semantic class, neighbouring classes, the histograms and local tonal deviations to create an estimate of tonal adjustment to apply. Finally, the expert uses a tool such as radial or gradual filter or a brush to apply the estimated adjustment locally to the individual region.

We aim to mimic this two-step process using a data-driven method. The spatial arrangement of the semantic classes and their neighborhood adjacency can be represented in form of nodes in a connected graph. Drawing inspiration from the expert's process, abstract features can be extracted for each region by combining the spatial arrangement, their global attributes, such as the luminance histogram, and local tonal deviations. We call these abstracted semantic features *'Semantic Hints'*. Due to the nature of data representation, GCN provides a powerful medium of leveraging local and neighbourhood information to compute the semantic hints. For each node, a GCN can leverage direct and indirect neighbourhood relations through connected edges and extract node specific semantic hints. These hints are used to apply the actual tonal adjustment locally using a fully connected network. In summary, in this work, we propose:

- a tone mapping operator which learns the tonal transformation as a function of semantic and contextual information of the image.
- a GCN to exploit the semantic information from the spatial arrangement of semantic segments in the image and predict aforementioned semantic hints.
- to exploit the hints in conjunction with the semantic features from the linear image to predict a tone mapped image aesthetically and perceptually close to a retouched version as generated by an expert photographer.
- LocHDR, a locally enhanced dataset of 781 HDR images tone mapped manually by an expert photo-retoucher.

II. RELATED WORK

The term ''tone mapping" is used to describe a broad range of techniques, often solving different problems. Therefore, it is important to position our research in that broader scope. The three main application areas of tone mapping are computer graphics, HDR video/television, and photography. In computer graphics tone mapping is used in the final stages of the rendering pipeline to simulate either a camera or the eye. Tone-mapping in graphics is often intended to bring a cinematographic look by simulating lens softness, or flare [4]. Alternatively, it could be used to mimic the appearance of scene as it would be perceived by the eye, for example, by simulating night vision [5], [6]. Another important application is HDR video and television, where color graded HDR content needs to be mapped to a display that may offer lower dynamic range and brightness than the reference HDR display used for color grading [7]. This paper focuses on the application of tone mapping in photography, where the goal is to produce images of certain aesthetics from linear (RAW) images captured by a camera sensor. All three application areas are not mutually exclusive, however, their input and aims are distinct.

The early tone mapping techniques for photography relied on heuristics or rules often inspired by photographic practices, intended to reproduce images of good contrast on displays of limited dynamic range [8], [9]. Later methods were guided by optimization, which attempted to find the best reproduction by minimizing a perceptual difference between the original and reproduced images [10], [11]. It has been acknowledged that the goal of finding the ''best'' tone mapped image is ill-posed as tone mapping is a subjective process with one-to-many possible mappings. Enhancement is often based on style of the individual and even then, can be inconsistent from one result to another. Mustafa et al. [12] reinforce this hypothesis and show that a style vector distilled from a ground-truth pair of raw and stylised images is required as a conditional parameter to find the closest colour mapping for each image. Since the main goal of photographic tone mapping is reproducing loosely defined image aesthetics, the problem is an excellent fit for machine learning techniques, which can learn from a large number of examples, without the need for well-defined rules. More recently, machine learning was introduced to tone mapping as a tool to learn mapping from RAW/HDR/linear images to their desired reproduction from a large dataset of training examples [13].

Tone mapping can be considered as a regression problem, in which the goal is to learn a function mapping from input HDR, RAW or linear images to the desired tone mapped images, usually provided by a large dataset of input/output examples. Such regression could be realized by standard techniques, such as LASSO (least absolute shrinkage and selection operator) or GPR (gaussian process regression) [13], by finding nearest-neighbors in a dataset of reference images [14], using a fully connected neural network to learn the coefficients of the quadratic polynomial basis functions [15], or learning simple brightness adjustment for semantic segments [2]. More recent methods involve a combination of fully connected and convolutional neural networks to extract both local and global (contextual) features from images [16], or the use of Generative Adverserial Networks (GANs) [17] to enhance region of shadows for darkened images [18]. Another popular choice is encoder-decoder architecture, based on convolutional neural networks [19], [20]. More recent unsupervised learning approach [21] or unpaired adverserial training-based operators [22] have also attempted to find the best tone mapping result. One common feature in all these methods is that the input to the regression typically combines local features, such as pixel color and its neighborhood and global features, such as image statistics, contextual or semantic information. Our method expands on this concept by explicitly modeling a trainable semantic graph of the image, which guides the tone mapping process.

All the aforementioned learning based methods implicitly use semantic information in different forms to improve tone

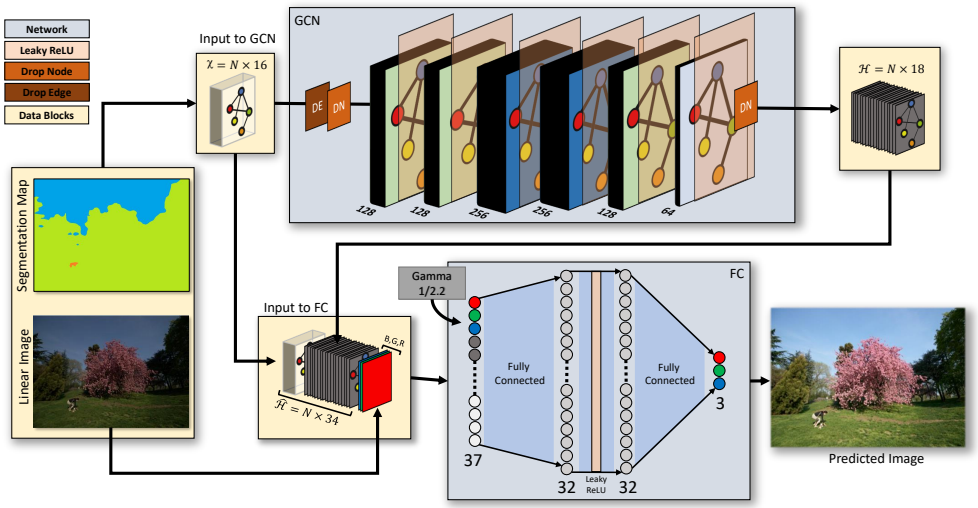


FIGURE 1. G-SemTMO has 4 data blocks and 2 network blocks. We obtain a connected graph of *N* semantic nodes from the linear image and its semantic map. Input feature matrix $\mathcal{X}$ and the node adjacency matrix from the graph is forwarded to the semantic hints module, GCN. The GCN uses graph convolutions to leverage scene context from the spatial arrangement of semantic segments and provide semantic-aware hints. It has 6 graph convolutional layers (128-128-256-256-128-64 channels) followed by an activation layer of Leaky-ReLU. *DropEdge* [23] and Node dropouts are used to prevent over-fitting. The GCN outputs latent semantic hints $\mathcal{H}$ with 18 hints per node. The latent features or hints highlight semantic relationship and guide the tone mapping in a data-driven manner. Broadcasted features $\mathcal{X}$ and $\mathcal{H}$ stacked together ($\hat{\mathcal{R}}$) with the input linear RGB create the final data block which is forwarded to the final network block FC. The FC block acts as a tone mapping module which uses the semantic hints from the GCN as a guide to tone map the linear RGB inputs. The FC has 2 fully connected layers (32-32 channels) with an activation Leaky-ReLU layer between the two. A gamma curve of 2.2 is applied to the input of the FC and the output is the tone mapped image.

mapping. However, we realise that semantic awareness is not limited to learning local or global attributes based on semantic categories. It also involves analysing the context under which the semantic categories are observed. Hence, we explicitly analyse semantic information through a graph of connected semantic segments. GCN helps us pass information between nodes in the graph [24] and learn local adjustments based on contextual information. A comparative study of graph neural networks and its applications [25] lists the domain of computer vision and image sciences where GCN has been applied for image classification [26], [27], segmentation [28] and reasoning [29] or image denoising [30]. However, to the best of our knowledge, our work is the first attempt to apply GCN as a model of trainable image semantics for tone mapping. Although digital images have a regular grid-like structure, their semantic segmentation maps combined with attributes per segment leads to an irregular data structure fit for graph-based representation. Training a GCN to learn contextual information from semantic categories and how it affects tonal modification, requires a dataset of input-and-retouched image pairs. MIT Adobe FiveK [13] dataset provides 5000 RAW images and their retouched versions created manually by 5 expert photographers. This dataset has been used to learn expert retouching styles, most notably for HDRNET [16].

III. SEMANTIC-AWARE TONE MAPPING

We propose a neural network architecture that is trained over the pairs of RAW linear and expert retouched images. The network learns to generate latent hints based on the semantic content of the image and adjust tone mapping based on this semantic information, as illustrated in Fig. 1. In the following subsections we describe our new learning-based pipeline. The pipeline has broadly two modules: a *Semantic Hints Module* and a *Tone Mapping Module*. The semantic hints module drives the semantic awareness of the TMO and generates the aforementioned hints. The application module works as a *n*-dimensional lookup table and learns a mapping as a function of the aforementioned hints. Before going further, it is necessary to delve into the notion of semantic awareness.

A. INTRODUCING SEMANTIC AWARENESS

To introduce semantic awareness, we incorporate the semantic features of a scene, based on the different labels obtained using a semantic segmentation algorithm, *e.g.* the color and luminance statistics per semantic label. We also incorporate the contextual understanding of the scene through a graph representing the neighborhood and spatial arrangements of the semantic labels in the segmentation map. We hypothesize that, along with the semantic features, the node-level neighborhood semantic information guides the image enhancement while retouching images.

Fig. 2 shows two images *A* and *B* from the Adobe FiveK [13] dataset, both manually retouched by expert E. We use FastFCN semantic segmentation algorithm [3] and merge the labels to coarser bins as suggested in SemanticTMO [2]. Although visually both images have a similar composition, a closer examination reveals the difference in semantic labels and their neighbors. The *water* semantic segment is surrounded by *sky* and *mountain* in image *B*, whereas in image *A* the *vegetation* and *city* are also neighbors to *water*. For

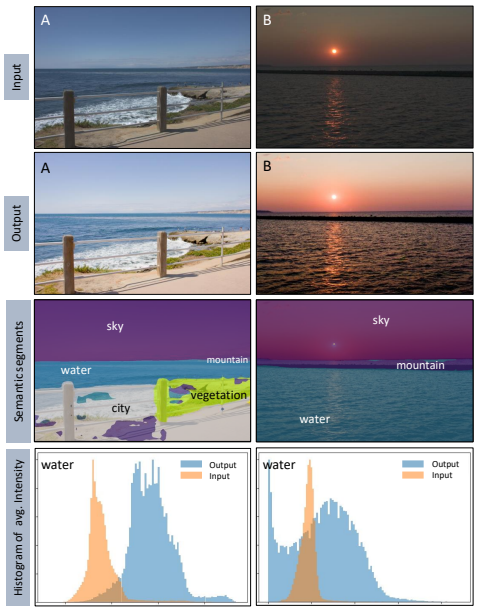


FIGURE 2. Understanding semantic awareness. *Row 1*: Gamma corrected input images A (*a1824*) and B (*a1892*). *Row 2*: Images manually retouched by expert E from MIT Adobe FiveK [13]. *Row 3*: Coarse semantic segments – fine labels obtained via FastFCN [3] segmentation and merged as per SemanticTMO [2]. *Bottom*: Input and output average intensity histograms for the '*water*' semantic segment. Histograms show markedly different output distribution for relatively similar input distribution.

both images, we plot the intensity histograms of the *water* segments for both the gamma-corrected input image and output image modified by the expert. The input histograms have a similar narrow distribution, although visibly shifted to the left for image *B* due to the overall low light. However, the output histograms show a very different distribution. The two segments receive different tonal adjustments despite having the same semantic label. This prompts us to conclude that the tonal adjustments are not just a function of semantic-based priors, but also of the local neighborhood of the semantic labels and their attributes, such as the intensity distribution or label information. Hence, we propose a learning-based tone mapping algorithm which leverages spatial semantic information, as well as the contextual information in the form of a graph capturing the spatial arrangements of the segments.

B. SEMANTIC HINT MODULE

A semantic segmentation network creates a map which divides an image into regions of semantic consistency. The segmentation map can be represented as a connected graph in which each node corresponds to a semantic segment and an edge is inserted when two semantic segments are neighbors in the map. This representation mimics the way a photographer may analyze the semantic information in an image.

Formally, an input image I with linear color values and with n semantic segments can be represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ are n nodes corresponding to the semantic segments, and $\mathcal{E}$ are the edges, represented as an adjacency matrix, such that $\mathcal{E}_{i,j} = 1$ if the segments corresponding to the nodes i and j are neighbours to each other. A GCN [24] is trained to learn a function on the graph $\mathcal{G}$. More specifically, it takes, for each node in the graph, an input feature vector $\mathbf{x}_i, i \in n$, summarised in a $n \times d$ feature matrix $\mathcal{X}$, where d is the number of features defining the semantic node. The GCN produces a node-level $n \times f$ output feature matrix $\mathcal{H}$, where f is the number of output features per node.

In our pipeline, the GCN takes an $n \times 16$ input feature matrix and produces an $n \times 18$ output feature matrix, called *semantic hints* $\mathcal{H}$. The input features include: the one-hot-encoded labels of the semantic segments (with 9 semantic classes, see Sec. III-D2), median and standard deviation for each R, G and B channel, and the median luminance value, all computed for the pixels belonging to the corresponding semantic segment.

Each layer l of the GCN can be represented as a function:

$$Y^{(l+1)} = \sigma\left(\mathcal{E}Y^{(l)}W^{(l)}\right), \tag{1}$$

where $Y^0 = \mathcal{X}$, $Y^{(L)} = \mathcal{H}$, and L is the last layer. $\mathcal{E}$ is the edge representation in form of an adjacency matrix, W^l is the weight matrix of layer l of the GCN and $\sigma(\cdot)$ is a non-linear activation function which, in our case, is Leaky-ReLU.

C. TONE MAPPING MODULE

The tone mapping module is a Fully Connected (FC) network, a 3D lookup table to map each input linear RGB pixel to the output display-encoded RGB pixel. Supplemental inputs allow this function to be local and semantics aware: the contextual information in form of $n \times 18$ semantic hints $\mathcal{H}$ from the GCN is passed in addition to the spatial information from the $n \times 16$ input feature matrix $\mathcal{X}$. The combined semantic information $\hat{\mathcal{H}}$ from the resulting $n \times 34$ matrix is spatially arranged with the input linear image such that each pixel in the image corresponds to 37 values: the 3 RGB channels and a 34-element semantic hint-feature vector. Consequently the FC trains over this 37 channel data to learn a mapping function:

$$f(I_{R,i,j}, I_{G,i,j}, I_{B,i,j}, \hat{h}_{1,k}, \hat{h}_{2,k}, ...\hat{h}_{34,k}) = O_{c,i,j}, \tag{2}$$

where k is the semantic segment corresponding to pixel (i,j). We train to minimise the L_1 difference in pixel values for all pixel positions $\{i,j\}$ and colour channels between the predicted, O, and reference, R, images:

$$\mathcal{L} = \sum_{i,j} \sum_{c \in \{R,G,B\}} \left| R_{c,i,j} - O_{c,i,j} \right|, \tag{3}$$

where both R and O are gamma-encoded RGB images in ITU-Rec.709 color space [31].

D. THE IMPLEMENTATION DETAILS

1) Preparing the image dataset

MIT-Adobe FiveK dataset [13] provides a set of 5000 high resolution RAW images and their manually retouched versions provided by 5 expert photographers (A, B, C, D, E). Prior work on image enhancement uses retouched versions created by expert C [13]–[15]. Gharbi et al. [16] use all

5 expert versions for their HDRNET but point out the inconsistencies among the expert retouches. In particular, they mention that expert B is more self-consistent and easier to learn for the network.

We initially choose expert E based on our subjective aesthetic preference of retouched results. However, we show in Sec. V-C that our architecture can learn irrespective of the choice of expert photographer. We observe that the dataset contains a significant number of images with large portion of saturated pixels in the RAW images. Adobe Lightroom software reconstructed those pixels to non-unique colors in the retouched images. As such saturated pixels may lead to inconsistent learning, we filter images with high number of saturated pixels before training. More specifically, we consider pixels as saturated where any of the RGB channels' value is above a normalised tonal value of .99. We filter out images with more than 3% of saturated pixels, this threshold being set empirically. This provides us with 4205 16-bit linear color images and their retouched versions for our training. We use the 'as-shot' white balance applied by the camera while exporting the linear images. For training, we resize images to the resolution of 100×100 pixels.

2) Preparing the input features

The next step is to generate input feature space for each image-graph representation. Global attributes and overall visual cues such as the average luminance or standard deviation of intensity values could inform decision on image enhancement. Based on this idea, Yan et al. [15] use both global and contextual feature descriptors for their image enhancement. We use similar attributes corresponding to each semantic region of the image. First, we use FastFCN semantic classifier [3] pre-trained over ADE20K dataset [32] to generate segmentation maps. ADE20K provides a dataset with 150 annotated labels which results in a very fine-grained semantic breakdown of an image. However, we realise that, in the use-case of digital photography, the semantic abstraction which drives decision on image edits is not as fine-grained. Therefore, we merge the fine labels to a coarser semantic abstraction based on the work of Goswami et al. [2]. The 9 coarse labels – *sky, mountain (terrain), vegetation, water, human subject, non-living subject, city, indoor-room* and *others* fit the use-case of digital photography better. The segmentation maps are generated at full resolution and consequently resized to match the training image resolution of 100. The spatial arrangement of the segments are stored in the edge descriptor $\mathcal{E}$ in Pytorch coordinate format (COO) for the GCN. Furthermore, we compute attributes for each segmented region: the median and standard deviation of RGB values, the median luminance and the 9-class one-hot encoded semantic labels for each semantic node.

3) GCN and semantic hint generation

The GCN-based Semantic Hint module has 6 graph convolutional layers generating $128, 128, 256, 256, 128$ and 64 latent features respectively. Each convolutional layer is followed by a Leaky-ReLU activation. To prevent overfitting the model, we use dropout layers before the first convolutional layer and after the last convolutional layer with probability of 0.2 and 0.5 respectively. Furthermore, we apply a DropEdge [23] with a probability of 0.2 before the first dropout.

4) Prediction using FC

The FC module is a function which learns the tone mapping from the input RGB values and semantic hint-feature vector. Specifically, we define it as a function $f : \mathbb{R}^{37} \rightarrow \mathbb{R}^3$. During training, the input to the FC is a 2D array 10000×37 containing all pixels in the image and their corresponding hints and features. We observe that applying a power of $1/2.2$ to the input of the FC improves its ability to learn non-linearity. The FC has two fully connected hidden layers with 32 neurons each separated by a Leaky-ReLU activation function. The output of the FC is the predicted non-linear RGB value. Due to the design of pixel prediction, the inference can be obtained on a high resolution image instead of 100×100.

5) Blending

The predicted output RGB values show visible inconsistencies at the border of semantic regions due to 1) the difference in tone mapping function across regions and 2) lack of smooth transition and segmentation precision of the FastFCN algorithm. In order to incorporate pixel precision, we utilise a shared alpha matting technique [33] and draw inspiration from the semantic framework idea of Goswami et al. [2] which involves stacking normalised fuzzy segmentation maps of each semantic region and blending the tonal modification.

To create the framework, we obtain n binary maps from a segmentation map containing n unique labels. Shared matting [33] converts each binary map into an alpha map using a trimap obtained by morphologically dilating the segment in the binary map with a disk of the radius 25 pixels. A bilateral filter (pixel neighborhood diameter $d = 50$ and color parameter $\sigma = 30$) is applied to each alpha map to remove discontinuities stemming from the morphological operations. The alpha maps are stacked along the z-axis and normalised to complete the semantic framework *(S)*. The FC is used to infer n images, one for each semantic hint where the same hint is used for all pixels. Stacking the n images similarly provides an image framework *(F)*. The weighted summation of the two frameworks provides us the blended image result.

$$O_{blended} = \sum_{i}^{n} S_i \cdot F_i \tag{4}$$

6) Training procedure

We use 4000 resized images out of the selected 4205 to train our network and keep 106 images for validation and 99 for inference. The weights and biases are optimized by minimising the loss defined in Eq. 3. The weights are further regularised with a weight decay of $5e-4$. We optimize the network parameters using ADAMW solver [34]. We train in

FIGURE 3. ***Ablation comparisons.*** **We present tone mapped results from the ablation studies for 3 images from the FiveK dataset: *a4886, a4986* and *a5000*. *Left to right –Ablation 1*: 3D LUT Global tone mapping, *Ablation 2*: 3D LUT Local tone mapping with semantic-specific information and Proposed G-SemTMO which considers semantic information as well as the contextual information from spatial arrangement of semantic labels using graph convolutions. The manually retouched version of each produced by expert E [13] is also included. The HyAB and PSNR objective metric scores for each tone mapped image validates the advantage of graph-based learning over the other ablation studies.**

batch size of 1 due to the variable structure of the graphs and the learning rate is scheduled to vary with the epoch. We train for 250 epochs with a learning rate of 10^{-3} between epoch 0 and 75, of 10^{-4} between epoch 75 and 150, and 10^{-5} from 150 onwards. We implement our architecture using PyTorch [35] and PyTorch Geometric [36] on an Nvidia RTX2060 GPU. The training takes about 24 hours.

IV. ABLATION STUDIES

To analyze the importance of each component of our method, we conduct two ablation experiments in addition to the proposed G-SemTMO. We observe in literature that tone mapping approaches work better than existing methods when explicit semantic information is provided as input [2]. Our hypothesis is that it can be improved further when contextual semantic information is supplied in conjunction to the learning pipeline. We designed our ablation experiments to incrementally modify the sophistication of semantic information introduced to the learning pipeline as follows:

Ablation 1: 3D LUT-Global mapping *Fully connected neural network (FC) without any semantic information.*

We utilise the FC architecture of our Tone Mapping Module to learn the mapping from linear RAW images to expert retouched images. No additional semantic information is provided and GCN is not used.

Ablation 2: 3D LUT-Local Semantic mapping *FC with semantic information.*

We train the image pairs over the FC architecture similar to Ablation 1. But for every pixel semantic information is added. A vector of size 19 is provided – 3 colour channel values and semantic-specific input features of size 16 similar to the input to GCN (refer to Sec. III-B). GCN is not used.

Ablation 3: Graph based Semantic mapping *G-SemTMO.*

This ablation consists of the full architecture with GCN, as explained in Sec. III. GCN uses semantic-specific input features to provide semantic hints to FC.

We conduct hyper-parameter optimisation during training. Schedulers are used to vary the learning rate for training to achieve generalisation. The models for ablation studies are finally trained using the same hyper-parameters. They are trained on 3000 training image pairs and validated on 20 image pairs for 250 epochs. To report test results, we compute the mean pixel HyAB perceptual colour distance [37] and the PSNR for the prediction results of 99 test images. We used HyAB rather than CIE DeltaE as it was shown to better capture luminance differences.

Observations: Fig. 3 presents 3 images from the FiveK dataset [13] tone mapped by the networks from the ablation studies and their respective HyAB color distance and PSNR scores. Based on subjective assessment, we conclude that our proposed graph-based learning produces results much closer to the ground truth for the selected images. Fig. 4 illustrates the training and validation loss curves across the three studies. The curves confirm our hypothesis that enriching the feature space with contextual semantic information improves the performance of the model. Across the three ablation studies, we observe that the model with the full semantic information results in lower training and validation loss. Fig. 5 plots the

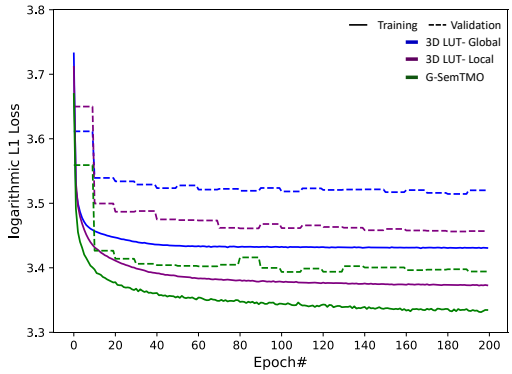


FIGURE 4. Training and validation curves for the ablation study. Introduction of semantic graph improves performance of the model.

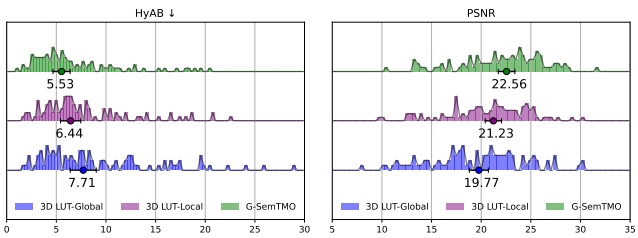


FIGURE 5. The histograms of HyAB and PSNR scores for the 3 ablation studies are presented. The histograms correspond to score distribution over 99 test images. The median of the distribution is plotted with a solid circle with a confidence interval of 95%.

histogram of HyAB and PSNR objective scores across 99 test images for each ablation study. Additionally, we plot the median for each histogram with its confidence interval of 95%. We observe that the proposed G-SemTMO prediction gets closest to the images retouched by expert E with a median perceptual colour distance score of 5.53. It also receives better PSNR evaluation than the other two ablations.

V. LEARNING GLOBAL IMAGE ENHANCEMENTS

Comparing tone mapping operators is a difficult task due to subjectivity of the tone mapped results. Although having the 'best' tone mapping outcome is an elusive objective, certain HDR datasets [13] provide reference tone mapped LDR images to reconstruct. To evaluate G-SemTMO's performance against references, we compare the quality of reconstruction of different operators using objective metrics. Conversely, it is also common for HDR images to not have any tone mapped references [38]. Consequently, several operators are developed or trained without a target reference. In such cases comparison becomes challenging, choice of objective metrics are limited and we rely on visual evaluation of objective image parameters such as contrast, saturation etc. In the following subsections, we compare G-SemTMO to several tone mapping and image enhancement operators based on results with and without reference LDR images. The use of different operators, datasets, metrics as well as visual observation just outlines how challenging the task of assessing TMO performance can be.

A. COMPARING ON MIT-FIVEK DATASET REFERENCES

The MIT-FiveK [13] dataset presents ground truth expert tone mapped images. We compare the results of G-SemTMO against the prediction of other supervised learning-based methods, HDRNET [16] (retrained on the same images as our method) and EnlightenGAN [18] (using the pre-trained weights provided by the authors). We also include 4 traditional TMOs: Photoreceptor TM [39], Photographic TM [8], Display Adaptive TM [10] and Bilateral TM [40]. The traditional TMOs do not allow for training and they are included in our comparison to show the difference between trained and non-trained tone mapping. We present our observations based on our subjective assessment and validate them using objective metrics.

Since we are unable to train the official HDRNET Tensorflow implementation due to rather old version of the dependencies, we rely on the PyTorch re-implementation by Jinchen Ge [41]. Gharbi et al. [13] use FiveK dataset to learn style transfer and their network was trained using image pairs comprising of 8-bit input images without corrections and 8-bit images retouched by experts. However, as per author suggestions, we use their network architecture to train for end-to-end tone mapping using 16-bit linear images as input and 8-bit retouched images as output. EnlightenGAN [18] provides weights to relight 8-bit input images. We use EnlightenGAN both on input images in linear RGB colour space and in gamma-encoded RGB space to generate separate results. To generate the results for the 4 traditional TMOs, we used pfstools[1] software.

To assess the tone mapping results, we use 6 objective metrics: PSNR, Multi-scale Structural Similarity Index (MS-SSIM) [42], Visual Difference Predictor for HDR images HDR-VDP-3 [43], hybrid perceptual colour distance metrics HyAB [37] and CIEDE2000 (ΔE_{00}) [44] and Colourfulness-based contrast quality metric (C-PCQI) [45]. More precisely, HyAB and ΔE_{00} are colour-sensitive and assess the closeness of color reproduction to the ground truth. They use computations in the CIELAB colour space to measure perceptual distance from the reference image. It has shown good agreement to subjective preference for small colour deviations. A smaller HyAB and ΔE_{00} score suggests better quality. To evaluate the reproduction of structural details and local contrast preservation, we use MS-SSIM and a colour and patch-based contrast quality metric C-PCQI. A higher score for either suggests a higher measure of structural and contrast reconstruction resulting in better perceptual quality.

Furthermore, for overall quality, we choose traditional PSNR and the HDR-VDP-3 (v3.0.6)[2] Quality correlate (Q) score. It is a measure of the magnitude of distortion corresponding to visibility rather than the mathematical distance between the pixels. The HDR-VDP-3 score attains a maximum of 10 for best perceptual quality and gets lower for poorer reconstruction.

[1] http://pfstools.sourceforge.net/pfstmo.html.
[2] https://sourceforge.net/projects/hdrvdp/files/hdrvdp/.

FIGURE 6. ***Left to Right*: 4 selected images from FiveK dataset. *Row 1*: Target images modified by expert E [13]. *Row 2-6*: Selected TMOs with HyAB and PSNR metric scores. *Row 7*: Tone curve applied per semantic segment by G-SemTMO.**

B. OBSERVATIONS

Fig. 6 presents the results for 4 images from the MIT Adobe FiveK dataset [13] (from the testing set). In the top row, we present the target images manually retouched by *expert E* (used for training HDRNET and G-SemTMO) and the following rows contain the results of each operator. Objective metric scores, HyAB and PSNR with respect to expert E, are also indicated. The last row contains the plots of the per segment gray-scale tonecurves, produced by G-SemTMO for each semantic region. The tone curves are generated by mapping input grayscale values (where $R = G = B$) to output color and then computing the luma value. Our first observation is that, for the selected images, G-SemTMO produces results closer to the expert retouched images than HDRNET and EnlightenGAN trained on the same data.

Another interesting observation can be made when analyzing the per-segment tone curves of G-SemTMO (the bottom row in Fig. 6). Each plot presents the tone curves predicted by G-SemTMO using the semantic hints per segment in a $\log_{10}$ space. We hypothesized that the neighborhood of semantic segments play a part in deciding the tonal adjustment inside the segment. Consequently, different neighborhood result in different tone curves for the same semantic label.

Fig. 7 compares the graph representations of the semantic segments in two images A and B from Fig. 6 (*a4986* and *a5000* respectively). Both images contain a large semantic segment annotated as *vegetation* but the neighbors to *vegetation* in A are different from B. Consequently, from Fig. 6, we observe that the tone curve for *vegetation* is different in the two plots. Hence, we validate that the GCN learns the neighborhood information and predicts different hints for the same semantic label resulting in different tone curves.

Fig. 8 shows the distribution of scores for aforementioned 6 objective metrics: PSNR, MS-SSIM, HDR-VDP-3 Quality, HyAB, C-PCQI and CIEDE2000 over 99 test images. For completeness, the plot also includes the results of the 4 other traditional tone mapping operators apart from HDRNET, EnlightenGAN (in linear and gamma encoded RGB colour space) and G-SemTMO, but as mentioned previously the traditional operators were not trained to reproduce the results of the experts. As reference, we use the test images manually retouched by expert E from the FiveK dataset. Along with

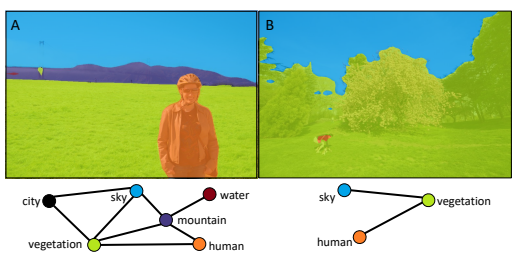


FIGURE 7. **Neighborhood based tonal adjustment. Image A (a4986), Image B (a5000) and their corresponding graph representation of semantic labels. The neighborhood of *vegetation* is different in A from B, the predicted tone curve would be different too.**

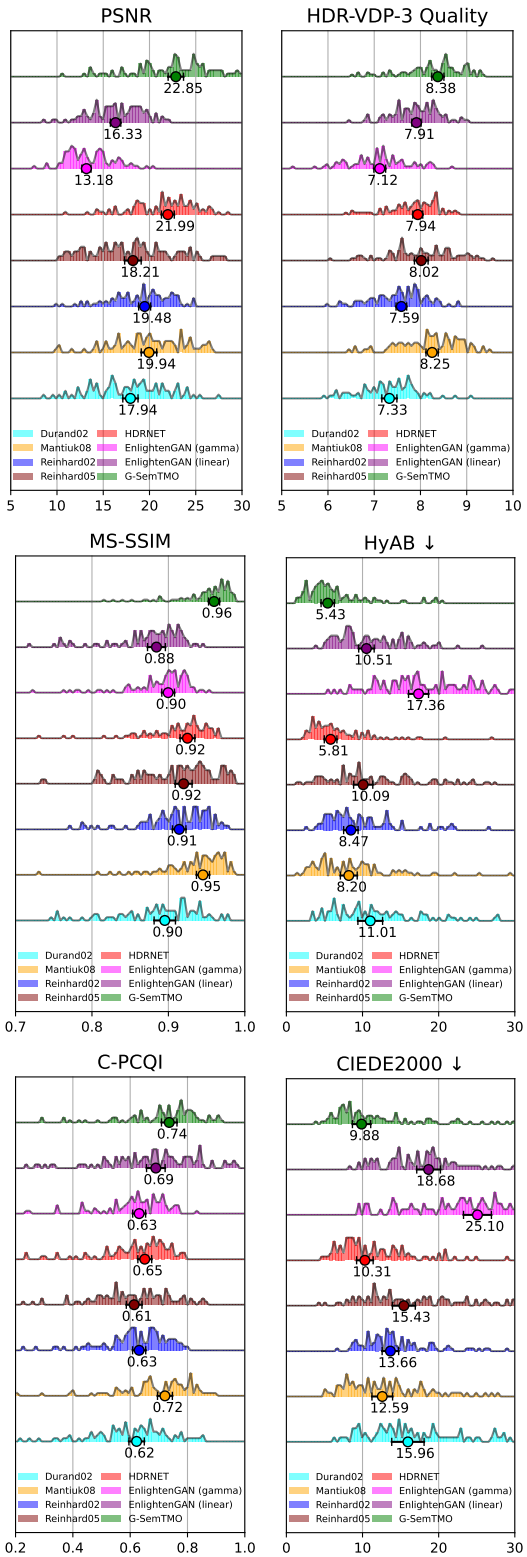


FIGURE 8. ***Objective metric scores*. We present 6 objective metrics: PSNR, HDR-VDP-3 Quality, MS-SSIM, and colour-sensetive metrics HyAB, C-PCQI and CIEDE2000. Each plot presents histograms of scores achieved by 7 TMOs: proposed G-SemTMO, EnlightenGAN [18] (linear and gamma-enc.), HDRNET [16], Photoreceptor TM [39], Photographic TM [8], Display Adaptive TM [10] and Bilateral TM [40]. The median of each histogram is marked with a solid circle and a confidence interval of 95%.**

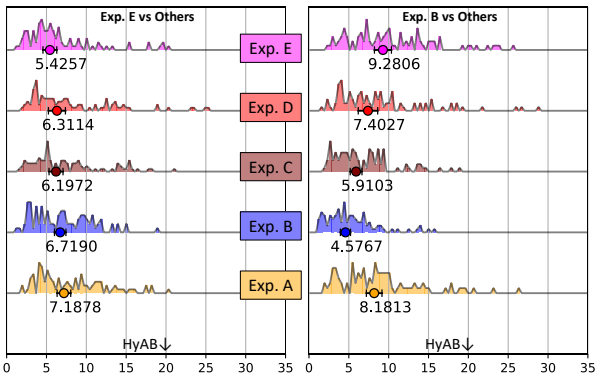


FIGURE 9. ***Learning distinct styles*: HyAB metric scores for expert E (left) and expert B (right) in comparison to the ground truth of other experts.**

the histogram of observed metric scores, we plot the median metric scores for each TMO with an error bar denoting a confidence interval of 95% of the median. The histograms confirm our subjective assessment of Fig. 6. We observe that across all objective metrics, the proposed G-SemTMO has a better median scores and produces results that are closer to the results of expert E compared to the other TMOs. Among of the included objective metrics, HyAB, C-PCQI and CIEDE2000 are colour-sensitive. We notice that HDRNET results rival G-SemTMO closely in terms of colour similarity (HyAB, CIEDE2000) but there is a visible softness which is reflected in worse scores for more spatial metrics (MS-SSIM, HDR-VDP-3, C-PCQI) sensitive to sharpness and local contrast. Fortuitously, the display adaptive tone mapping also produces results that are close to the retouched images of expert E.

C. TRAINING FOR OTHER MIT FIVEK EXPERTS

We trained our network over the same set of training images for the 4 other expert photographers in the FiveK dataset and validated the results over the 99 test images. We use the same hyper-parameters for training as in Sec. III-D6. We observe that there are inconsistencies among the tonal adjustments provided by the experts in the FiveK dataset as a result of which learning tone mapping becomes harder.

Consequently, to validate that our network can differentiate between the styles of each expert and learn tonal adjustment specific to the expert, we compare the prediction of G-SemTMO trained for a particular expert to the other expert ground truth. Fig. 9 shows the performance of networks trained over expert E and B with HyAB metric. We observe that results predicted by network trained over E is closest to the ground truth E than others for 99 images. The same holds true for network trained over expert B. This concludes that the parameters learnt by the network are specific to the expert trained. This is in alignment to Mustafa et al. [12] which advocates learning specific style vectors for better representation of tone mapped images. Our work uses GCN to incorporate semantic information and could learn an approximate style specific mapping for each expert on the FiveK dataset.

Gharbi et al. [16] mention that HDRNET could learn the adjustments made by expert B better. We also notice that our training could learn and infer better for expert B, as validated by the objective metric scores in Fig. 10. Subjectively analysing the enhancements, we find the adjustments made by expert A to be the most inconsistent.

D. COMPARING WITH HDR PHOTOGRAPHIC SURVEY

Our previous results aimed to reconstruct ground truth references produced by experts. In this subsection, we look beyond target ground truth reconstruction and observe the performance of G-SemTMO against traditional methods and data-driven methods focused towards unsupervised or unpaired learning. We choose images from the HDR Photographic Survey dataset [38] for comparison. Fig. 11 presents 2 images tone mapped by traditional TMOs - Display Adaptive TM [10], Photographic TM [8], Photoreceptor TM [39] and proposed G-SemTMO. On cursory observation, we find that G-SemTMO produces results high on contrast, saturation and hence appear more aesthetic than other washed-out tone mapped images. It also preserves details in the shadows and highlights better than others. Fig. 12 presents images from the same dataset tone mapped by data-driven operators - UnCLTMO [21], UnpairedTMO [22], DeepTMO [20] and proposed G-SemTMO. While UnCLTMO and DeepTMO both produce visibly high-contrast images with heightened low frequency details and darker shadows, UnpairedTMO fares low on contrast at the lower-frequencies. G-SemTMO manages to preserve the details on highlights and shadows, and on high and low frequencies while producing results overall high in contrast. It must be reiterated that the traditional TMOs do not allow for training whereas the data-driven methods are trained on the said dataset.

We acknowledge that cursory observation over a handful images is insufficient and highly subjective. However, the objective attributes such as contrast or saturation positively reinforces our hypothesis that G-SemTMO can produce favourable results in a no-reference comparison. To reiterate, the objective to find the best TMO is ill-posed. While other operators provide a unique 'best' representation for a linear image, G-SemTMO is able to produce as many representations as the number of styles it is trained on.

VI. LEARNING LOCAL IMAGE ENHANCEMENTS

It can be argued that the tonal adjustments created by the expert photographers for the FiveK [13] dataset is global in nature. The photographers had access to limited tools and sliders from the Adobe Lightroom photo-retouching application. Although, the sliders can effect non-linear adjustments, they are not as local as using brushes and radial/gradual filters to modify images. It is important to validate the performance of G-SemTMO in learning local tonal adjustments. Consequently, we present a locally enhanced dataset of HDR images, *LocHDR*. We train over our dataset and conduct ablation studies to confirm whether graph convolution manages to learn tone modifications closer to the reference.

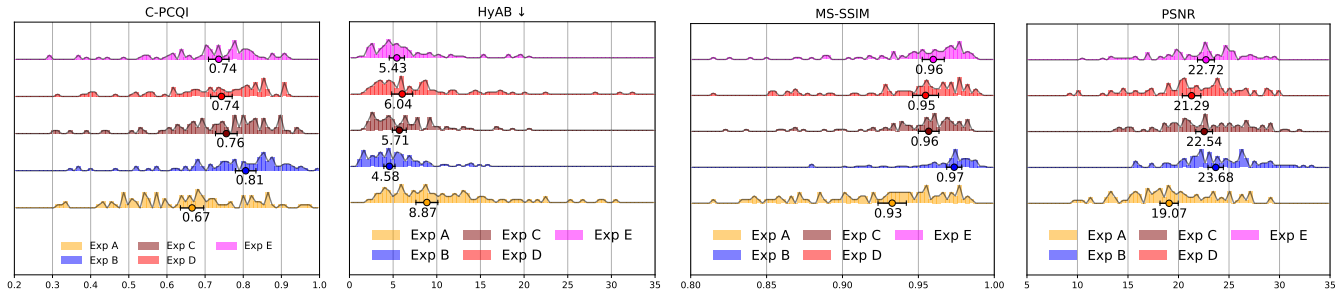


FIGURE 10. **Metric score distributions for C-PCQI, HyAB, MS-SSIM and PSNR for networks trained over 5 experts individually. The plots show the histograms of scores along with medians and its 95% confidence intervals.**

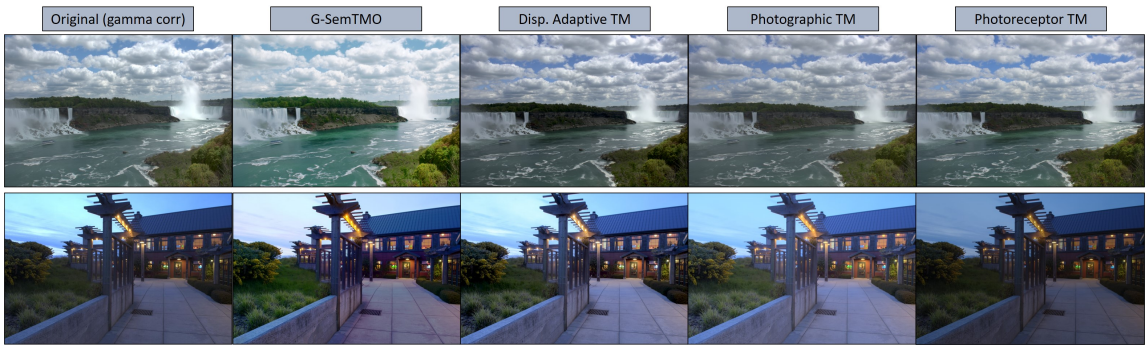


FIGURE 11. **Images from HDR Photographic Survey dataset [38]. Comapring G-SemTMO (trained on expert E) to traditional tone mappers: Display Adaptive TM, Photographic TM and Photoreceptor TM.**

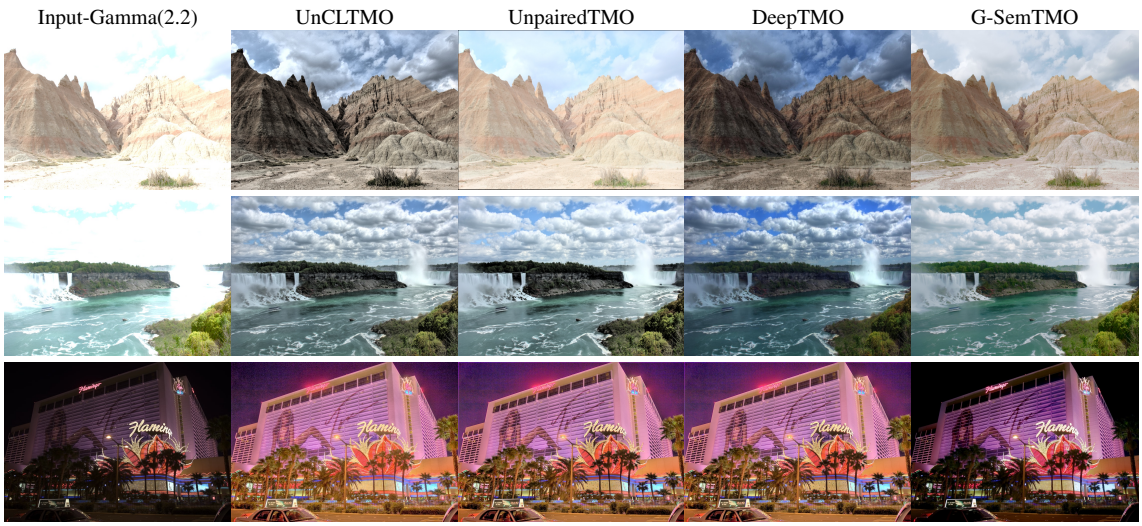


FIGURE 12. **Images from HDR Photographic Survey dataset [38]. Comapring G-SemTMO (trained on expert E) to data-driven tone mappers: UnCLTMO, Unpaired TMO and DeepTMO.**

A. LOCAL ENHANCEMENT HDR DATASET - LOCHDR

We filter the images from FiveK based on their dynamic range. We compute dynamic for images from FiveK range as the logarithm of the ratio of the 99^{th} and 1^{st} percentile of observed luminance and empirically put a threshold of 2.2 to filter out images. Furthermore, to emphasize on local changes, we filter images based on the number of semantic segments and choose images with at least 3 unique semantic labels. Based on our criteria we compile a subset of 781 HDR images. We hire an expert photo-retoucher (henceforth referred to as expert I) who is tasked to apply corrections to the LocHDR dataset using Adobe Lightoom application with emphasis on using brushes and spatial filters. Only the sliders in the Tone section — *Exposure, Contrast, Highlights, Shadows, Whites* and *Blacks* are used and no auto-enhancement or colour, noise, detail adjustments are made.

On closer observation and investigation, we uncover that manual enhancement using local tools leads to tonal enhancement inconsistencies in the LocHDR Dataset. Fig. 13 illustrates how local enhancements across and within images for

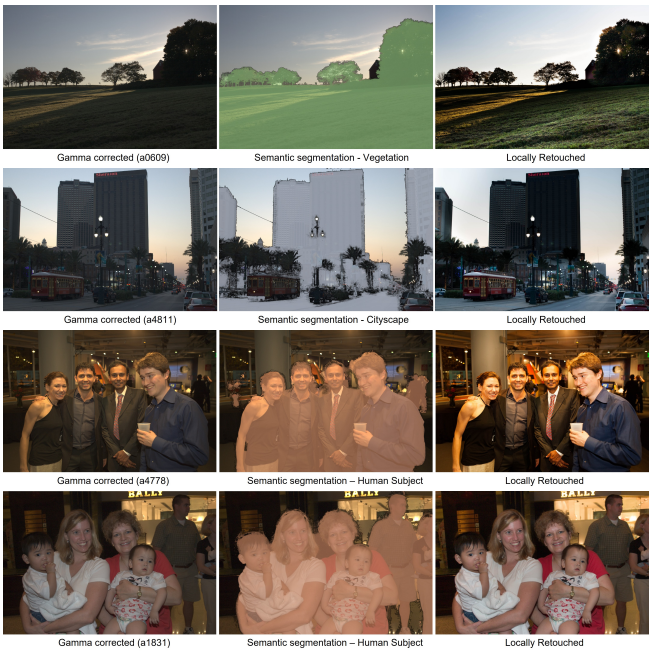


FIGURE 13. Inconsistencies in local enhancements. From Left-Right: the gamma corrected image, chosen semantic class and locally enhanced image. We can observe noticeable variation of local contrast inside masked semantic classes *Vegetation* and *Cityscape* for images a609 and a4811. Furthermore, we notice contrast variations across images a4778 and a1831 for the same semantic class *Human Subject*. Image a4778 has a high local and global contrast making it the enhancement perceptually different that image a1831.

the same semantic label vary. On image 'a0609' we observe that parts of the *Vegetation* segment has received inconsistent exposure gains. Similarly, parts of the *Cityscape* segment in image 'a4811' are well exposed whereas parts of the building are darkened. Comparing the enhanced image 'a4778' to 'a1831' we observe that the former appears 'punchy' with heightened contrast on the *Human Subject*.

Learning tone mapping from reference pairs is similar to learning individual retouching styles. Our expert does not use the same semantic masks which the G-SemTMO uses to learn local enhancements. Hence, it becomes challenging for the network to train over an image dataset if there are local enhancement inconsistencies. The network fails to generalize and converge on the style of adjustment. Previous work on FiveK dataset [13] mentions the inconsistencies in retouches [15], [16] and the use of data splits such as 'Random 250' and 'High Variance 50' [14]–[16] for analysis. We also decide to split our LocHDR based on consistency in style.

B. STYLE-SPECIFIC HIGH CONTRAST ENHANCEMENT - HC200

To evaluate whether the presence of semantic graphs helps G-SemTMO learn local enhancements better, we present a final dataset of *High Contrast 200* images. We filter the LocHDR dataset based on the perceivable contrast effected by expert I to maintain consistent enhancement in training set.

Measuring the perceptual contrast or how 'contrasty' or 'punchy' an image appears is a challenging task. Inspired by multi-level approaches in entropy computation [46] and structural similarity measures [42], we present our own approximation of a multi-level contrast measure. Multi-level contrast follows a multi-grid approach where at each level n, the full resolution image is divided into a grid of $n \times n$ patches and patch-specific variance of pixel intensity is computed. The contrast for level n is the square root of the mean variance. The final multi-level contrast measure is computed as the mean of level-specific contrast scores thereby capturing the global as well as local contrast variations:

$$C_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}\left(\sqrt{\frac{\sum_{p=1}^{i^2} Var_p}{i^2}}\right), \tag{5}$$

where n is the number of levels, and p is the index of patches in a level from 1 to $n \times n$. We empirically set $n = 5$ for our contrast estimation. We choose the 200 images with highest contrast measure. On visible subjective assessment, we can confirm that the HC200 subset mostly contains the high contrast 'punchy' images from LocHDR.

C. TRAINING & INFERENCE

We train over the 200 high contrast images using K-fold cross validation [47] ($K = 4$) with a training-testing data split of $150/50$. Following observations from previous ablation studies, we use the ADAMW solver [34] for optimization, weight decay of $5e-4$ and a scheduled learning rate of 10^{-3} between epoch $0-150$, 10^{-4} after 150^{th} epoch and finally 10^{-5} after epoch 300.

We trained three networks separately to observe the influence of graph convolutions and style-specific training data:

- a network with local semantic information without graph convolutions, Local LUT trained over LocHDR dataset.
- a network with local semantic information with graph convolutions, G-SemTMO trained over LocHDR dataset.
- a network with local semantic information with graph convolutions, G-SemTMO trained over style-specific HC200 dataset.

For comparison, 40 common images are chosen from the test sets of LocHDR and HC200. Fig. 14 shows the results of selected test images for assessment. As mentioned previously, we consider three networks to compare the inference subjectively. For each image in the figure, we see marked improvement in the inference quality of G-SemTMO when trained over style-specific HC200. Network trained over HC200 manages to predict the local contrast in the images closest to the ground truth. This confirms that it is essential for neural networks to be trained on image data with consistent enhancement styles to learn local enhancements better. Previously in Fig. 9, we have shown that G-SemTMO could learn different retouching styles. The training on HC200 shows that for datasets with local adjustments the network can be finetuned by training for specific style

FIGURE 14. Comparing Inference on HC200 dataset. *From left to right*: Gamma corrected source image, 3D LUT Local trained on LocHDR, G-SemTMO trained on LocHDR, G-SemTMO trained on HC200 and Ground truth. HyAB colour distance and PSNR metric scores show significant improvement when G-SemTMO is trained over the style and contrast-specific HC200 images.

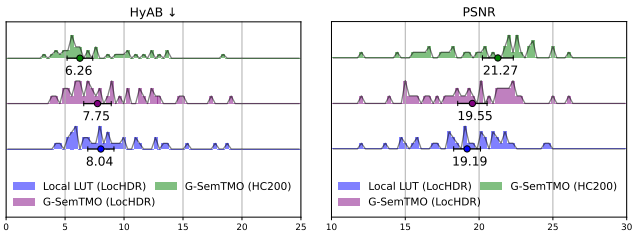


FIGURE 15. Comparing HyAB and PSNR scores for test images inferred by G-SemTMO (trained on HC200 and on LocHDR) and 3D Local LUT (trained on LocHDR). HyAB and PSNR histograms show respective scores over 40 test images common to both testing sets. We observe that G-SemTMO trained on the specific style of HC200 produces significantly better inference and colour closeness than others.

variations inside the data-subset. Fig. 15 compares the three trained networks objectively on the basis of PSNR and HyAB colour closeness. The HyAB and PSNR histograms for the three networks are plotted along with their median score with a confidence interval of 95%. From our experiments, we observe that by training over the entire LocHDR dataset G-SemTMO can only marginally improve over the quality of Local LUT. However, from Fig. 15, we observe significant improvement in G-SemTMO inference quality when trained over a specific style.

VII. CONCLUSIONS AND PERSPECTIVES

In our work, we introduced G-SemTMO, a novel local tone mapping operator, which can learn global and local tonal transformations from semantic-graph representations of images and the spatial arrangement of the semantic regions and LocHDR, a locally tone mapped dataset of HDR images manually retouched by an expert.

We compare the results obtained using G-SemTMO in our experiments and ablation studies to the ones tone mapped by the selected reference TMOs and we can confidently claim that graph-based learning can better incorporate semantic awareness in a TMO. We evaluate G-SemTMO in two ways. First, we show that G-SemTMO can learn global enhancements from MIT Adobe FiveK dataset [13] and reconstruct reference images better than selected traditional and data-driven TMOs. It performs equally well on the HDR Photographic Survey dataset [38] without reference. Second, following our novel dataset of locally tone mapped HDR images we show that G-SemTMO can learn local enhancements by letting the graph convolutional network leverage the spatial arrangement of semantic regions. When comparing over data from MIT FiveK dataset, our results show that our network can produce images closer to the versions manually retouched by expert photographer E than the other methods. When comparing over data from the LocHDR and HC200 dataset, we observe that the presence of graph convolutions help even further in learning local enhancements with consistent tonal modifications in the training image set in comparison to networks without graph convolutions.

However, in the process of developing G-SemTMO, we identify some limitations as well. First, our algorithm is reliant on the semantic segmentation of the images to create a graph of their spatial arrangement of the segments and applying tonal enhancement locally. We observe several cases where the label annotations are improper. Image *a1824* in Fig. 2 contains a segment *city*, which should clearly belong to the segment *water*. Fig. 16 further demonstrates how the segmentation algorithm falsely annotates the foreground as a combination of *vegetation, terrain* and *others*. False classification results in improper tonal enhancement and visible artifacts. Handling improper labels are more challenging with fine grained semantic labelling. Merging labels to coarser segments helps reduce improper annotation to an extent but the requirement for a segmentation algorithm and annotated dataset with labels fit for the use case of photography still remains. This can reduce not just improper labelling but also introduce labels that are closer to an expert photographers' impression of a scene. Second, G-SemTMO is unable to learn colour shifts, white-balance adjustments or tonal rela-

FIGURE 16. Limitation - Improper segmentation. *Bandon Sunset* from HDR Photographic Survey tone mapped by G-SemTMO (left) and its segmentation mask shows inconsistency in the semantic labels.

FIGURE 17. Limitation - Colour shifts. G-SemTMO is unable to learn white balance adjustments and resulting hue shifts. *a4904* from FiveK dataset retouched by expert E (left) shows the magenta hue for underwater image which G-SemTMO (right) finds challenging to reproduce.

tionships which are isolated or sparse in the training dataset. Fig. 17 shows the expert modified image (left) compared to that tone mapped by G-SemTMO (right). Both figures 16 and 17 demonstrate cases which G-SemTMO finds challenging and reflects their poor subjective and objective qualities.

Finally, G-SemTMO in its current state treats all the neighbor semantic segments equally while predicting the latent semantic hints. However, in many cases, semantic segments occupy low percentage of pixels. Image *a5000* in Fig. 7 (bottom) has a very small proportion of pixels annotated as *human* but it impacts the tonal adjustment of its neighbor label *vegetation* equally as the label *sky*. One approach to address this would be to have edge-weighted learning where the GCN considers the edge adjacency along with the edge weight into account based on the size of the semantic segment.

Furthermore, while learning global enhancements from FiveK dataset [13] we observe that our networks trained over the 5 experts learn tone mapping specific to the expert but do not learn any distinct structural modifications. G-SemTMO does not perform local structural modifications (e.g. sharpening) and it focuses instead on color and contrast transformations. Additionally, it must be noted that our network trains on input images with as-shot camera white balance. So, it is unable to reproduce occasional custom white balance modifications made by the expert.

Our LocHDR and HC200 locally tone mapped image datasets provide a valuable contribution for development of data-driven TMOs. G-SemTMO has shown that it can learn local enhancements from the dataset and get closer to ground truth compared to networks without graph convolutions. It can learn better when trained over data with specific style. However, we acknowledge that LocHDR has limitations with inconsistency in enhancements and representation of node permutations owing to relatively low number of images. We look forward to improving the dataset with larger number of tone mapped image pairs with more consistent tonal adjustment and wider representation of semantic graphs.

We acknowledge that conducting psycho-physical experiments [48] or forced-choice preference experiments [49] to evaluate tone mapping quality is a robust methodology. However, we emphasize that the goal of finding the best TMO based on subjective preference is ill-posed. Hence, we reformulate the problem as learning a semantic-aware enhancement specific to an expert's style. Our goal is to produce results that are close to those of an expert photographer rather than to produce the most preferred results. Hence, we do not perform a formal subjective comparison of the results. We aim to learn and generalise enhancement set as reference ground-truth and evaluate the performance based on full reference metrics. We find the existing full-reference objective metrics sufficient for evaluation of that goal. However, we also use our network trained on expert E to tone map images from the HDR Photographic Survey dataset to show that our results can be generalised for unsupervised cases to produce aesthetically pleasing results.

ACKNOWLEDGMENT

This work has been supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 765911 (RealVision ITN). We would also like to extend our gratitude to Ishani Jayawardhana GJKIU for the hours of dedicated image retouching which helped us create our dataset of images.

REFERENCES

[1] A. Adams, *The Print. Vol. 3. The Ansel Adams photography series*. Boston: Little, Brown and Company., 1983.

[2] A. Goswami, M. Petrovich, W. Hauser, and F. Dufaux, "Tone Mapping Operators: Progressing Towards Semantic-awareness," in *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2020, pp. 1–6.

[3] H. Wu, J. Zhang, K. Huang, K. Liang, and Y. Yu, "FastFCN: Rethinking dilated convolution in the backbone for semantic segmentation," *arXiv preprint arXiv:1903.11816*, 2019.

[4] M. Hullin, E. Eisemann, H.-P. Seidel, and S. Lee, "Physically-based Real-Time Lens Flare Rendering," *ACM Transactions on Graphics*, vol. 30, no. 4, p. 1, jul 2011. [Online]. Available: http://portal.acm.org/citation.cfm?doid=2010324.1965003

[5] P. Irawan, J. Ferwerda, and S. Marschner, "Perceptually Based Tone Mapping of High Dynamic Range Image Streams," *EGSR*, 2005.

[6] R. Wanat and R. K. Mantiuk, "Simulating and compensating changes in appearance between day and night vision," *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, vol. 33, no. 4, p. 147, 2014.

[7] SMPTE, "ST 2094-2:2017 Dynamic Metadata for Color Volume Transform — KLV Encoding and MXF Mapping," 2017.

[8] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic Tone Reproduction for Digital Images," in *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, 2002, pp. 267–276.

[9] G. Krawczyk, K. Myszkowski, and H.-P. Seidel, "Lightness Perception in Tone Reproduction for High Dynamic Range Images," in *The European Association for Computer Graphics 26th Annual Conference EUROGRAPHICS 2005*, ser. Computer Graphics Forum, vol. 24, no. 3. Dublin, Ireland: Blackwell, 2005, pp. xx–xx.

[10] R. Mantiuk, S. Daly, and L. Kerofsky, "Display Adaptive Tone Mapping," in *ACM SIGGRAPH 2008 papers*, 2008, pp. 1–10.

[11] Kede Ma, H. Yeganeh, Kai Zeng, and Zhou Wang, "High Dynamic Range Image Compression by Optimizing Tone Mapped Image Quality Index," *IEEE Transactions on Image Processing*, vol. 24, no. 10, pp. 3086–3097, oct 2015. [Online]. Available: http://ieeexplore.ieee.org/document/7111279/

[12] A. Mustafa, P. Hanji, and R. Mantiuk, "Distilling Style from Image Pairs for Global Forward and Inverse Tone Mapping," in *Proceedings of the 19th ACM SIGGRAPH European Conference on Visual Media Production*, 2022, pp. 1–10.

[13] V. Bychkovsky, S. Paris, E. Chan, and F. Durand, "Learning photographic global tonal adjustment with a database of input/output image pairs," in *CVPR 2011*. IEEE, 2011, pp. 97–104.

[14] S. J. Hwang, A. Kapoor, and S. B. Kang, "Context-based Automatic Local Image Enhancement," in *European conference on computer vision*. Springer, 2012, pp. 569–582.

[15] Z. Yan, H. Zhang, B. Wang, S. Paris, and Y. Yu, "Automatic Photo Adjustment using Deep Neural Networks," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 2, pp. 1–15, 2016.

[16] M. Gharbi, J. Chen, J. T. Barron, S. W. Hasinoff, and F. Durand, "Deep Bilateral Learning for Real-Time Image Enhancement," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 118, 2017.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[18] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "EnlightenGAN: Deep Light Enhancement without Paired Supervision," *IEEE transactions on image processing*, vol. 30, pp. 2340–2349, 2021.

[19] R. Montulet, A. Briassouli, and N. Maastricht, "Deep Learning for Robust end-to-end Tone Mapping," in *BMVC*, 2019, p. 194.

[20] A. Rana, P. Singh, G. Valenzise, F. Dufaux, N. Komodakis, and A. Smolic, "Deep Tone Mapping Operator for High Dynamic Range Images," *IEEE Transactions on Image Processing*, vol. 29, pp. 1285–1298, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8822603/

[21] C. Cao, H. Yue, X. Liu, and J. Yang, "Unsupervised HDR Image and Video Tone Mapping via Contrastive Learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[22] Y. Vinker, I. Huberman-Spiegelglas, and R. Fattal, "Unpaired Learning for High Dynamic Range Image Tone Mapping," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 14 657–14 666.

[23] Y. Rong, W. Huang, T. Xu, and J. Huang, "Dropedge: Towards deep graph convolutional networks on node classification," *arXiv preprint arXiv:1907.10903*, 2019.

[24] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," 2017.

[25] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A Review of Methods and Applications," *AI Open*, vol. 1, pp. 57–81, 2020.

[26] X. Wang, Y. Ye, and A. Gupta, "Zero-Shot Recognition via Semantic Embeddings and Knowledge Graphs," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6857–6866.

[27] M. Kampffmeyer, Y. Chen, X. Liang, H. Wang, Y. Zhang, and E. P. Xing, "Rethinking Knowledge Graph Propagation for Zero-Shot Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 487–11 496.

[28] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan, "Semantic Object Parsing with Graph LSTM," in *European Conference on Computer Vision*. Springer, 2016, pp. 125–143.

[29] Z. Wang, T. Chen, J. Ren, W. Yu, H. Cheng, and L. Lin, "Deep Reasoning with Knowledge Graph for Social Relationship Understanding," *arXiv preprint arXiv:1807.00504*, 2018.

[30] Y. Li, X. Fu, and Z.-J. Zha, "Cross-Patch Graph Convolutional Network for Image Denoising," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4651–4660.

[31] ITU-R, "Parameter values for the HDTV standards for production and international programme exchange," ITU-R Recommendation BT.709, 2015.

[32] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic Understanding of Scenes through the ADE20K Dataset," *International Journal of Computer Vision*, vol. 127, no. 3, pp. 302–321, 2019.

[33] E. S. L. Gastal and M. M. Oliveira, "Shared Sampling for Real-Time Alpha Matting," *Computer Graphics Forum*, vol. 29, no. 2, pp. 575–584, May 2010, proceedings of Eurographics.

[34] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," 2019.

[35] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.

[36] M. Fey and J. E. Lenssen, "Fast Graph Representation Learning with PyTorch Geometric," in *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[37] S. Abasi, M. Amani Tehran, and M. D. Fairchild, "Distance metrics for very large color differences," *Color Research & Application*, vol. 45, no. 2, pp. 208–223, 2020.

[38] M. D. Fairchild, "The HDR photographic survey," in *Color and imaging conference*, vol. 2007, no. 1. Society for Imaging Science and Technology, 2007, pp. 233–238.

[39] E. Reinhard and K. Devlin, "Dynamic Range Reduction Inspired by Photoreceptor Physiology," *IEEE transactions on visualization and computer graphics*, vol. 11, no. 1, pp. 13–24, 2005.

[40] F. Durand and J. Dorsey, "Fast Bilateral Filtering for the Display of High-Dynamic-Range Images," in *ACM transactions on graphics (TOG)*, vol. 21, no. 3. ACM, 2002, pp. 257–266.

[41] J. Ge, "HDRnet-PyTorch," accessed: 2022-07-13. [Online]. Available: https://github.com/gejinchen/HDRnet-PyTorch

[42] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale Structural Similarity for Image Quality Assessment," in *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402.

[43] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Transactions on graphics (TOG)*, vol. 30, no. 4, pp. 1–14, 2011.

[44] G. Sharma, W. Wu, and E. N. Dalal, "The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations," *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, vol. 30, no. 1, pp. 21–30, 2005.

[45] K. Gu, D. Tao, J.-F. Qiao, and W. Lin, "Learning a no-reference quality assessment model of enhanced images with big data," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 4, pp. 1301–1313, 2017.

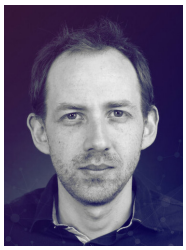
[46] W. Zhang, R. R. Martin, and H. Liu, "A Saliency Dispersion Measure for Improving Saliency-Based Image Quality Metrics," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 6, pp. 1462–1466, 2018.

[47] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.

[48] H.-H. Choi, H.-S. Kang, and B.-J. Yun, "Tone mapping of high dynamic range images combining co-occurrence histogram and visual salience detection," *Applied Sciences*, vol. 9, no. 21, p. 4658, 2019.

[49] A. Ak, A. Goswami, W. Hauser, P. Le Callet, and F. Dufaux, "RV-TMO: Large-Scale Dataset for Subjective Quality Assessment of Tone Mapped Images," *IEEE Transactions on Multimedia*, vol. 25, pp. 6013–6025, 2023.

ABHISHEK GOSWAMI is a Postdoctoral Research Fellow at the Visualisation Lab, WMG, University of Warwick, UK. He was a fellow in the MSCA funded Realvision ITN and has a PhD in Signal and Image Processing from Université Paris-Saclay, France. He received his M.Sc in Computer Science from Saarland University, Germany and B.Tech Engg. degree in Computer Science from IEM, Kolkata, India.



ERWAN BERNARD is a senior researcher of DxO's image processing team. He holds a PhD in Physics and Automatism from ISAE-SUPAERO Toulouse France in partnership with ONERA (Toulouse, France) and Sagem (Massy, France). DxO develops image processing software for amateur and professional photographers.

ARU RANJAN SINGH received an M.Tech. degree in manufacturing science and engineering from the Indian Institute of Technology (IIT), Kharagpur, India, in 2019 and a Ph.D. degree in deep learning for manufacturing from Warwick Manufacturing Group, University of Warwick, U.K. His research interests include machine learning, computer vision, computer graphics, and manufacturing automation.

WOLF HAUSER is the scientific and technical lead of DxO's image processing team. He holds M.Sc. in electrical engineering from University of Stuttgart, Germany, and Télécom ParisTech, France. Over the last 15 years, he has worked on color image processing, image restoration and raw image development. DxO develops image processing software for amateur and professional photographers.

FREDERIC DUFAUX is a CNRS Research Director at Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes (L2S, UMR 8506), where he is head of the Telecom and Networking hub. He received his M.Sc. in physics and Ph.D. in electrical engineering from EPFL in 1990 and 1994 respectively.

Frederic is a Fellow of IEEE. He was Vice General Chair of ICIP 2014, General Chair of MMSP 2018, and Technical Program co-Chair of ICIP 2019 and ICIP 2021. He served as Chair of the IEEE SPS Multimedia Signal Processing (MMSP) Technical Committee in 2018 and 2019. He was a member of the IEEE SPS Technical Directions Board from 2018 to 2021. He is Chair of the Steering Committee of ICME in 2022 and 2023. He was also a founding member and the Chair of the EURASIP Technical Area Committee on Visual Information Processing from 2015 to 2021. He was Editor-in-Chief of Signal Processing: Image Communication from 2010 until 2019. Since 2021, he is Specialty Chief Editor of the section on Image Processing in the journal Frontiers in Signal Processing.

Frederic is also on the Executive Board of Systematic Paris-Region, a European competitiveness cluster which brings together and drives an ecosystem of excellence in digital technologies and DeepTech.

He has been involved in the standardization of digital video and imaging technologies for more than 15 years, participating both in the MPEG and JPEG committees. He was co-chairman of JPEG 2000 over wireless (JPWL) and co-chairman of JPSearch. He is the recipient of two ISO awards for these contributions.

His research interests include image and video coding, 3D video, high dynamic range imaging, visual quality assessment, video surveillance, privacy protection, image and video analysis, multimedia content search and retrieval, and video transmission over wireless network. He is author or co-author of 3 books, more than 200 research publications (h-index=50, 10000+ citations) and 20 patents issued or pending.

RAFAŁ K. MANTIUK is a Professor of Graphics and Displays at the Department of Computer Science and Technology, University of Cambridge (UK). He received Ph.D. from the Max-Planck Institute for Computer Science (Germany). His recent interests focus on computational displays, rendering and imaging algorithms that adapt to human visual performance and deliver the best image quality given limited resources, such as computation time or bandwidth. He contributed to early work on high dynamic range imaging, including quality metrics (HDR-VDP), video compression and tone-mapping. More recently, he led an ERC-funded project on the capture and display system that passed the visual Turing test - 3D objects were reproduced with fidelity, which made them undistinguishable from their real counterparts.