MetaISP – Exploiting Global Scene Structure for Accurate Multi-Device Color Rendition

Matheus Souza and Wolfgang Heidrich
King Abdullah University of Science and Technology, Saudi Arabia

Abstract
Image signal processors (ISPs) are historically grown legacy software systems for reconstructing color images from noisy raw sensor measurements. Each smartphone manufacturer has developed its ISPs with its own characteristic heuristics for improving the color rendition, for example, skin tones and other visually essential colors. The recent interest in replacing the historically grown ISP systems with deep-learned pipelines to match DSLR’s image quality improves structural features in the image. However, these works ignore the superior color processing based on semantic scene analysis that distinguishes mobile phone ISPs from DSLRs. Here we present MetaISP, a single model designed to learn how to translate between the color and local contrast characteristics of different devices. MetaISP takes the RAW image from device A as input and translates it to RGB images that inherit the appearance characteristics of devices A, B, and C. We achieve this result by employing a lightweight deep learning technique that conditions its output appearance based on the device of interest. In this approach, we leverage novel attention mechanisms inspired by cross-covariance learn global scene semantics. Additionally, we make use of metadata that typically accompanies raw images, and we estimate scene illuminants when they are not available.

CCS Concepts
• Computing methodologies → Image Manipulation; Computational Photography;

1. Introduction
Mobile devices have emerged as the dominant platform for digital photography. The increasing demand for immediate photo sharing and consumption without manual post-processing has driven the need for advanced color processing capabilities in these devices. To meet this demand, mobile phones are now equipped with cutting-edge, high-level image processing techniques [Inc21], which greatly enhances color processing capabilities.

In contrast, DSLR cameras still produce images with flat colors, placing the burden of color rendition on the end user. Over the past few decades, as camera modules have become smaller and image quality has suffered due to approaching physical limitations, mobile devices have increasingly relied on computational methods to restore high-quality images. The Image Signal Processor (ISP) in a mobile device plays a crucial role in this process. It translates the RAW sensor measurements into an sRGB image using various image processing stages, including interpolation, demosaicking, denoising, white balance, and color restoration. ISPs have accumulated extensive knowledge about color theory to generate visually pleasing images that not only match the scene accurately but also appeal to human perception, taking into account cultural differences.

Figure 1 highlights the importance of modern ISPs in overall image color reproduction. It presents images captured using three different smartphone models and processed using the (relatively simple) open-source OpenISP and the native ISPs of each device.

© 2023 The Authors. Proceedings published by Eurographics - The European Association for Computer Graphics. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.2312/vmv.20231236
Notably, there are strong color and shading differences in the images processed by the native ISPs, even though the images processed by OpenISP appear very similar (and much more "muted" in comparison). This demonstrates two points: 1) The final color reproduction on modern mobile phones is primarily determined by the ISP rather than the camera hardware; and 2) ISPs from different phones make distinct aesthetic choices regarding color reproduction, thereby establishing a brand-specific visual style.

In this work, we propose MetaISP, an ISP that takes as input a RAW image from one mobile phone and can closely reproduce the appearance (color and local contrast characteristics) of any of a range of other phones. This is achieved by training MetaISP to learn to combine meta data and global scene structure in a way that emulates the existing black-box ISPs of these mobile phones. Since the different device characteristics are encoded into a latent feature space, this also makes it possible for MetaISP to interpolate between different device aesthetics.

2. Related Work

2.1. RAW Images and ISPs

In recent years, there has been a growing interest in exploring the unprocessed RAW image space. RAW data, directly captured by the sensor, contains abundant information that has not been influenced or compressed by any Image Signal Processor (ISP). While sRGB images are widely available and commonly used, there have been several studies proposing the translation from sRGB to RAW, followed by performing enhancement tasks in the RAW domain, and finally converting the result back to sRGB [ZAK∗20, XQC21, PB19, NB16]. However, these approaches face certain limitations, primarily due to the inherent challenges in accurately reconstructing RAW images. ISPs are complex software components that are difficult to model and invert successfully.

Working directly in the RAW domain has led to significant advancements in various areas of image processing, including low-light image reconstruction, intrinsic image decomposition, image denoising, reflection removal, real scene super-resolution, illuminant estimation, and image enhancement [CCXK18, BMX∗19, APB21, GCPD16, LHZ∗20]. These methods often replicate parts of the ISP pipeline while leveraging deep learning techniques to match the output with a domain where the images are better conditioned. The idea of replicating an entire camera ISP was early explored by Karaimer and Brown [KB16], while Heide et al. [HST∗14] pioneered the replacement of numerous image processing blocks with a joint optimization approach. More recently, this approach has been extended to matching RAW images captured on mobile devices with high-quality DSLRs using deep learning. Ignatov et al. introduced the PyNet architecture [IVGT20], which utilized paired images from the Huawei P20 (RAW) and Canon 5D Mark IV DSLR (sRGB) to learn how to enhance the quality of mobile device images to match the DSLR level. Subsequently, several challenges [ITZ∗20, ICK∗21] have explored different architectures and strategies to achieve better translations between these domains.

Among the top-reported architectures, AWNet [DLLC20] and MW-ISPnet [ITZ∗20] introduced Multilevel Wavelet CNNs and Residual Channel Attention Blocks. Subsequently, LiteISP [ZWL∗21] further improved upon these works by incorporating their successful techniques along with an alignment strategy based on optical flow and a color prediction network branch. Building upon these advancements, LiteISP proposed an architecture capable of faithfully reproducing the color of mobile ISPs. It leveraged raw metadata and generated high-resolution images as output. However, their approach was limited to learning one ISP at a time. In contrast, MetaISP not only learns a multi-device ISP but also an illumination estimator, enabling both reproduction and interpolation between different device aesthetics.

2.2. Image Translation

The RAW-to-sRGB conversion can be interpreted as an image-to-image translation problem. In recent years, much of the work in this area has focused on solutions based on GANs [GPAM∗20]. The generative characteristic of those models was explored to generate different solutions inside the same domain, or conditioning the generated image to have different characteristics based on a different domain input, which was called multimodal image translation [ZZP∗17] and conditional image translation [IZZE17], respectively. Later, these works inspired image reconstruction tasks like colorization [LMS16], super-resolution [LTH∗17], inpainting [PKD∗16], etc. Recently, works like All-in-One Weather Removal [LTC20], aims to remove degradation caused by adverse weather conditions like rain, snow, or fog with one single network. More recently, TransWeather [VYP22] considered a similar problem based on a transformer-based network instead of generative models. These works usually receive different corruptions as input and attempt to output a restored image.

Using a similar concept, we propose a multi-ISP single image reconstruction network to translate from one single RAW input to a desired sRGB image, conditioning the translations to a given label representing the target device, like early Generative Models [MO14], but applying Convolutional Neural Networks (CNNs), together with attention mechanisms as global feature extractor.

3. Model

In this work, we proposed a RAW-to-RGB mapping that takes into account the possibility of multiple correct translations within different domains represented by three devices: Google Pixel 6 Pro ("Pixel"), Samsung S22 Ultra ("S22"), and iPhone XR ("iPhone"). The objective of our model is to reproduce the ISP of all these devices when provided with the same RAW image. To achieve this, our model is designed to handle small misalignments commonly found in datasets acquired using different phones. We accomplish faithful color reproduction by employing a CNN architecture that can produce diverse translations conditioned on a device embedding, incorporating a lightweight self-attention mechanism. Furthermore, we enhance our translations with a global semantics branch and a metadata awareness branch. The metadata awareness branch utilizes metadata information during training and can estimate it during inference, contributing to a more comprehensive translation process.

Given the raw RGBG image patch $x \in \mathbb{R}^{H \times W \times 4}$, we aim to learn the target output $y_{d} \in \mathbb{R}^{2H \times 2W \times 3}$. The device indicator $d$ represents the Pixel, S22, and iPhone with values 0, 1, and 2, respectively.
Figure 2: a) is an overview of the MetaISP pipeline, b) exemplifies the residual attention block, and c) describes the attention mechanism inside the residual blocks. In this architecture, the RAW patch image $x$ and $e_d$ go through the illumination branch, outputting the scene white balance. Next, the metadata awareness block projects $W_B$ to match the backbone network’s first two levels and aggregate the ISO and exposure time. Later, a downsampled full-resolution version of the image goes through the transformer-based global semantics block to extract global features, matching them inside the bottleneck. Finally, the $e_d$ condition the decoding part with different device embeddings, serving as a scaling factor in the bottleneck and as the query vector inside the decoder attention mechanism. The number of convolutional layers, residual attention blocks and residual blocks are the same as the Figure depicts. The residual block follows the b) diagram without the attention and the innermost skip connection.

3.1. Architecture

Among the architectures proposed in [ITZ* 20], MW-ISPnet stands out by utilizing a Multilevel Wavelet UNet [RFB15] in combination with residual channel attention, resulting in significant improvements in the reconstruction process. Subsequently, LiteISP [ZWL* 21] demonstrated that it is possible to achieve high-quality results while reducing the number of attention residual blocks. MetaISP builds upon this backbone architecture but incorporates modifications to tackle the challenging task of encapsulating different domains within a single model, while avoiding issues such as mode collapse or averaging of device appearances. Additionally, the network efficiently utilizes metadata and effectively extracts global semantics. For a comprehensive overview of the global architecture, refer to Figure 2.

3.1.1. Meta data awareness

The raw file comprises unprocessed or minimally processed data and image metadata, which provides information about the characteristics of the image and the device settings chosen during the capture, considering factors such as lighting conditions and user preferences. While traditional ISPs make use of this data, most current Deep Learning models tend to overlook it. Within a DNG file, standardized metadata includes parameters like the color matrix, white balance weights, exposure settings, and ISO settings.

MetaISP considers the importance of white balance information when reproducing an ISP and incorporates two pipelines. In the first pipeline, where white balance information is available with the RAW image, the network assumes the white balance provided by the iPhone RAW image. Inside the network, the White Balance branch adapts this information to match the scene illuminants of the Pixel and S22 devices. The second pipeline, on the other hand, includes an Illumination Branch capable of estimating the white balance information given the RAW image and the device embedding.

In the first pipeline, the white balance metadata is represented as a scalar multiplier per color channel, with the green channel typically normalized to a value of 1. This information is expanded to $w_d \in \mathbb{R}^4$ in RGBG format. Convolutional layers then upsample this information to match the size of corresponding layers in the reconstruction backbone. This enables the white balance infor-
mation to be injected into the reconstruction process at distinct locations through a channel-wise dot product. In the second pipeline, a lightweight model is employed to learn the white balance information when it is either unavailable or lost. In this case, the ground truth illuminant information from the collected dataset for the Pixel and S22 devices is applied. Therefore, the white balance can be approximated given the RAW image and the device. The network aims to capture global color information for each channel and follows the same procedure as the first pipeline.

Additionally, the model leverages the ISO and exposure time to learn an affine transformation described by the parameters $\alpha$ and $\beta$. This transformation globally alters the appearance of the image, ensuring even light distribution across the entire scene and effectively addressing vignetting issues (corners darker than the image center). The learning process involves projecting both the ISO and Exposure information independently into a higher dimension using linear layers followed by ReLU activation. These two vectors are then concatenated, and two additional linear layers are used to project them back to a lower dimension, matching the output dimension of the first white balancing branch procedure. The final two vectors ($\alpha$ and $\beta$) are used to scale and shift globally the image. Combining ISO and exposure time arises from the fact that both characteristics are related to global scene light condition.

### 3.1.4. Global semantics

Relying solely on local patches is insufficient as they lack information about the relationships between different objects in the scene. Patches only provide texture or cropped structure observations. To address this limitation, MetaISP incorporates a cross-covariance attention mechanism, utilizing a downsampled version of the entire image $x$ pre-processed through bilinear interpolation. The cross-covariance transformer, introduced by [ATC*21], offers a lightweight attention mechanism by computing self-attention with respect to the dimension feature. We leverage this transformer to aggregate information from the entire image, exploiting the known large receptive field property. The final representation of the transformer is obtained using the CLS layer $[TCS*21]$, which is commonly employed in classification tasks to aggregate information of the whole image. The features are then projected using a convolutional layer with a kernel size of 1 to match the dimension of the bottleneck layer. Finally, the projected features are merged with the bottleneck using element-wise multiplication. This approach aligns with current commercial ISPs often perform rough classifications, and the model implicitly learns which features are more important in the scene.

**Normalization** Transformers typically require large amounts of data for effective training. Although the cross-covariance approach helps reduce computational complexity, they still tend to require significant data resources compared to pure CNNs. In this work, we proposed the use of monitor data augmentation, as described in Section 4. Initially, MetaISP is pre-trained using monitor-captured data to increase its semantic diversity. During the subsequent fine-tuning procedure, we retain the statistics learned from the monitor data in the global semantics stage. This means that during fine-tuning, the batch normalization does not compute new mean and

\[ x \in \mathbb{R}^{HW \times dim}. \]
variance based on the current mini-batch, but rather uses the running mean and variance learned during the pre-training with monitor data. In Section 5, we provide a quantitative evaluation of this strategy.

**Light Transformer** Our global semantics branch divides the image in 16x16 patches, with 4 blocks with attention dimensionality of 128 and 4 heads, making it smaller than the tiny version proposed in [ATC*21]. With its additional 1.39M parameters, the model is still lightweight compared with other methods.

### 4. Dataset and Training Details

In the existing literature, there is a lack of large datasets that provide aligned images from different mobile devices. Although PyNet [IVGT20] offered aligned Huawei RAW patches with DSLR images, it did not provide the opportunity to explore strategies like the ones proposed in this work (refer to Section 3.1). Furthermore, DSLR images, unlike current mobile devices, typically require post-processing to achieve good color rendition.

To overcome these limitations, we conducted a data collection process to obtain real-world images from each device. Our dataset consists of 163 diverse scenes captured in indoor, outdoor, night, and daily environments. In addition to the iPhone XR, Google Pixel 6 Pro, and Samsung S22 Ultra, we also collected images from the Xiaomi P60 to evaluate the model in a zero-shot fashion. This resulted in a total of 652 images, which were further divided into patches as described below. During the data collection process, all phones were set to automatic mode for white balance, focus, ISO, and exposure time. The captures were performed using a tripod, with the phones arranged in a grid formation, and a remote controller was used to trigger simultaneous shots.

To enrich our dataset, we also captured images in front of a calibrated HDR monitor. This data was exclusively used for pre-training our model, allowing it to learn semantics from a broader range of scenes multiple devices in real, dynamic environments. The images were captured in a dark room using the Eizo CG3145 4K HDR monitor, which was carefully color calibrated to ensure accurate representation, and the use of a HDR monitor means that we can produce images that exceed the contrast of the camera, so that we obtain realistic image saturation and glare effects. The reference images in the monitor dataset were sourced from Flickr2K [LSK*17] and cover a wide range of landscapes, vegetation, objects, indoor and outdoor scenarios, etc. The monitor is capable of displaying scenes with a dynamic range that exceeds that of the cameras, enabling us to capture realistic exposure and saturation effects for each device. The monitor dataset consists of a total of 5487 RAW/sRGB image pairs with a resolution of 4032 × 3024. The impact of the pre-training strategy is quantitatively demonstrated in Section 5. We will make the entire dataset, including the pre-training and real-world images, publicly available.

### 5. Results

To assess the performance of our model, we conduct a quantitative comparison with state-of-the-art methods using our collected datasets, performing the pre-training with monitor captured data and training with the proposed real-world dataset. Furthermore, we perform a qualitative comparison of all methods against ground-truth images. Additionally, we demonstrate the capability of our model to interpolate between different device styles, as well as its potential to process images from phone models that were not included in the training data, albeit with reduced color accuracy.

Our model is compared against the state-of-the-art methods LitelISP [ZWL*21], MW-ISPNet [ITZ*20], and SwinIR [LCS*21]. These works primarily focus on image enhancement by translating RAW to sRGB. SwinIR, originally designed for image super-resolution, was adapted to handle four-channel raw images. As no previous method has addressed multi-ISP before, we trained each technique separately to match the quality of the three devices in our dataset.

The raw input image for all methods is the iPhone image multiplied by the white balance. However, in our case, we utilized the white balance adaptation strategy described in Section 3.1.1. Our experiments confirm this strategy’s contribution to improving our model’s performance (see Table 2).

For quantitative evaluation, we adopted commonly used metrics, including PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index), and ΔE∞,76 (perceptual color accuracy error). These metrics were measured using ground-truth images that were warped to align with each network’s output using PWC-Net [SYLK18] for all methods. This alignment step was performed to mitigate the contribution of misalignment in the metrics, allowing us to focus on evaluating color accuracy. Please check the supplementary material for more details regarding the alignment.

Table 1 demonstrates that MetaISP outperforms state-of-the-art methods in all the evaluated metrics and across different devices. In an all-in-one fashion, our method can better reconstruct the target devices without having to train individually for each task (refer to supplementary for ablation about this case). Table 1 also highlights that when the illuminants are learned together with the model (MetaISP), the performance is slightly worse. This is attributed to the challenge of accurately estimating illuminants, as minor deviations can result in different color temperatures, thereby changing the image’s overall appearance. However, we still consider this approach to help compensate for cases where illuminant information is not available. In some cases, it performs better when targeting Pixel and Samsung color appearances. This can be explained by the fact that, in this approach, we train the model to match their original illuminants, which could be more accurate than iPhone (original RAW source) values adapted by the white balance branch in the network.

The metrics also evidence the challenging task of ISP reproduction. State-of-the-art image reconstruction networks cannot perform as well as models designed to treat this problem, LitelISP [ZWL*21] and MW-ISPNet [ITZ*20] explore domain-specific knowledge about the ISPs to provide mechanisms that improve the raw-to-rgb reconstruction, which lacks in SwinIR [LCS*21]. MetaISP combines novel deep-learning architectures and insights from legacy ISPs.

Qualitatively, Figure 3 visually demonstrate the closeness of our MetaISP results to the ground truth. With a scalar combined with the RAW file as input, MetaISP can generate a diverse
Table 1: Quantitative comparison of MetaISP and the illuminants variation against state-of-the-art methods evaluated using patches.

<table>
<thead>
<tr>
<th>Model</th>
<th>iPhone XR</th>
<th>Pixel 6 Pro</th>
<th>Samsung S22</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR ↑</td>
<td>ΔE_{ab}^*↑</td>
<td>SSIM ↑</td>
</tr>
<tr>
<td>MW-ISPnet [ITZ’20]</td>
<td>26.32 5.84 0.892</td>
<td>22.39 9.18 0.659</td>
<td>20.92 9.88 0.639</td>
</tr>
<tr>
<td>LiteISP [ZWL’21]</td>
<td>26.87 5.54 0.903</td>
<td>22.97 9.03 0.759</td>
<td>21.90 9.46 0.726</td>
</tr>
<tr>
<td>SwinIR [LCS’21]</td>
<td>25.76 5.92 0.889</td>
<td>21.75 9.22 0.637</td>
<td>20.49 9.98 0.627</td>
</tr>
<tr>
<td>MetaISP</td>
<td>30.14 3.76 0.915</td>
<td>25.49 6.65 0.784</td>
<td>24.26 7.33 0.753</td>
</tr>
<tr>
<td>I-MetaISP</td>
<td>29.12 5.07 0.897</td>
<td>24.89 7.28 0.780</td>
<td>23.66 7.77 0.749</td>
</tr>
</tbody>
</table>

Figure 3: Patch-wise comparison between the different methods and the ground truth. The images mainly represent challenging scenarios such as indoor and night scenes. Our model can accurately reproduce the color appearance of each target device. The alignment strategy described in the supplementary effectively avoids blurry effects that may be observed in SwinIR, resulting in sharper and more visually pleasing results. On the other hand, LiteISP does not suffer from blurry effects but struggles reproducing the color perception.

Interpolation Another feature of our work is the interpolation between different device embeddings, producing appealing intermediate results. Figure 4 shows that possibility with one intermediate result between each device. This number can be increased, creating very smooth transitions.

Zero-shot qualitative evaluation Figure 5 shows how the reconstruction looks when another smartphone is considered without training. On the top-left corner, we have the Xiaomi C40 sRGB, and the other three images are the reconstructions when its RAW image passes through the model. We can notice an overall improvement, with crisp colors, better contrast, and more fine details compared to the native ISP of this inexpensive phone. Please see the supplement for more results.

5.1. Ablation Studies

In our work, we have provided a detailed quantitative evaluation for each contribution, as outlined in Table 2.

A represents the baseline, a network similar to LiteISP but without channel attention. However, this baseline produces slightly inferior results. Next, we added B, the white balance first pipeline as described in Section 3.1.1. This pipeline learns how to retain the white balance information, injecting it into deeper layers, while also adapting the values for different phones’ color perception, instead of only just pre-multiplying it by the white balance scalars. C represents the global semantics branch, which provides knowledge about the global relationship between image patches. This seman-
Table 2: Ablation over the different contributions. Each one of them has a letter from A to E.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>iPhone XR</th>
<th>Pixel 6 Pro</th>
<th>Samsung S22</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR ↑</td>
<td>∆E&lt;sub&gt;ab&lt;/sub&gt;↑</td>
<td>SSIM ↑</td>
</tr>
<tr>
<td>A Baseline</td>
<td>25.29</td>
<td>6.73</td>
<td>0.870</td>
</tr>
<tr>
<td>B + Adapt. Illuminants</td>
<td>25.30</td>
<td>6.30</td>
<td>0.858</td>
</tr>
<tr>
<td>C + Global Semantics</td>
<td>27.23</td>
<td>4.88</td>
<td>0.904</td>
</tr>
<tr>
<td>D + Attention</td>
<td>29.44</td>
<td>4.02</td>
<td>0.904</td>
</tr>
<tr>
<td>E + Iso and Exp</td>
<td>30.14</td>
<td>3.76</td>
<td>0.915</td>
</tr>
</tbody>
</table>

Figure 4: Here, we exemplify the possibility of interpolating between the embeddings of different devices generating intermediate images. We can notice how different ISPs understand the scene and put more emphasis on different features. We have iPhone XR, Pixel 6 Pro, and S22 Ultra from left to right.

Figure 5: Zero-shot processing of images from new devices. MetaISP generating different interpretations from a Xiaomi C40 input without additional training. Top-left Xiaomi C40 (input, top-right simulated Pixel 6 Pro, bottom-left simulated S22 Ultra and bottom-right simulated iPhone XR. Please zoom for more details.

6. Conclusion

In this work, we have successfully built a single model to reproduce the complex color and appearance decisions performed by the ISPs of three commercial devices. Even though each smartphone has a different sensor, lenses, etc., the ISPs are responsible for the brand-specific visual style, and MetaISP learns jointly how to perform this task without collapsing or averaging to one single appearance. With fewer parameters (see supplementary) and performing three tasks at once, the proposed method outperforms the current state-of-the-art in all metrics (PSNR, SSIM, and ∆E<sub>ab</sub>↑), enhancing the original image or creating new appearance possibilities, interpolating between different phone style spaces. This work also exploits the white balance information that usually comes with RAW files, and it’s also robust to cases where this information is not available, learning how to estimate it using a light model that receives the RAW and device embedding as inputs. We also propose a novel idea to extract global semantics extra patches, as well as an attention to account for global communication intra patches.

Finally, in the future, we want to explore different devices, understanding their individual characteristics, learning how to translate from other devices without any further training requirements, expanding the idea of zero-shot presented here, as well as adding more control over the styles.

References


[ATC+21] Ali A., Touvron H., Caron M., Bojanowski P.,
Matheus Souza & Wolfgang Heidrich / MetaISP – Exploiting Global Scene Structure for Accurate Multi-Device Color Rendition


© 2023 The Authors. Proceedings published by Eurographics - The European Association for Computer Graphics.