Polarized Color Image Denoising
Supplementary Material

Zhuoxiao Li Haiyang Jiang Mingdeng Cao Yinqiang Zheng
The University of Tokyo

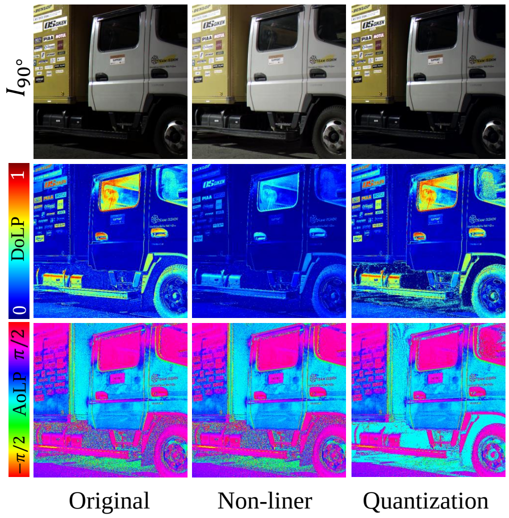


Figure 1. Visual comparison of polarization information, intensity(S_0), DoLP(ρ), AoP(ϕ) extracted from original linear, nonlinearized, or poorly quantized data.

1. Polarization information

Given a raw polarization image, pixels under polarizers of four angles, i.e., 90° , 45°, 0°, and 135° are divided into four channels, termed as I_{90°, I_{45°, I_{0°, and I_{130°. Using the intensities, Stoke parameters can be calculated as follows for describing the linear polarization state of incident light:

$$\begin{aligned} S_0 &= \frac{1}{2}(I_{0^\circ} + I_{45^\circ} + I_{90^\circ} + I_{135^\circ}), \\ S_1 &= I_{0^\circ} - I_{90^\circ}, \\ S_2 &= I_{45^\circ} - I_{135^\circ}. \end{aligned} \tag{1}$$

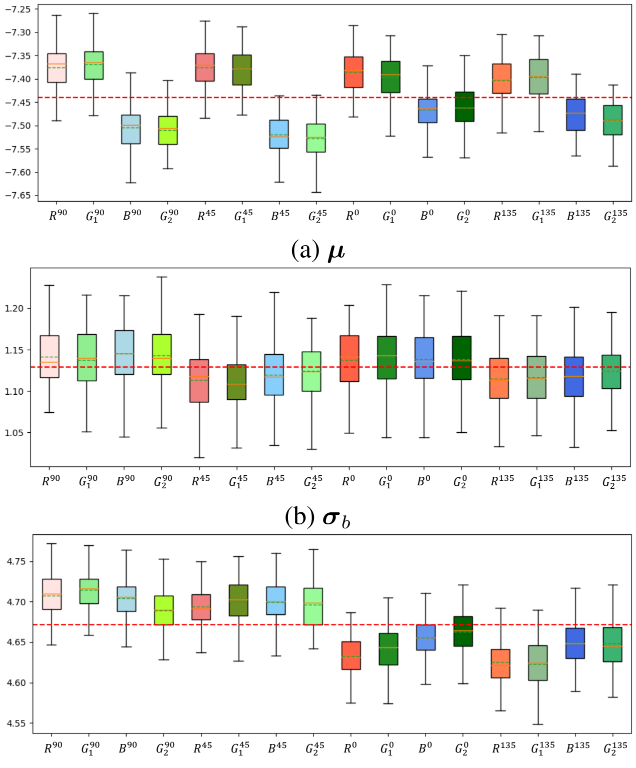
It should be noticed that Stoke parameters build an interconnection among the individual polarization pixel values. Three components are mostly used for polarization information measurements, including the intensity of light (described by S_0), the **Degree of Linear Polarization** (the proportion of fully linearly polarized light in a beam, DoLP, ρ), and the **Angle of Linear Polarization** (the direction of polarization plane, AoLP, ϕ). ρ and ϕ are calculated as:

$$\rho = \frac{\sqrt{S_1^2 + S_2^2}}{S_0}, \phi = \frac{1}{2}\arctan\frac{S_2}{S_1}. \tag{2}$$

These information are essential for material property analysis and surface normal computations. Therefore, many downstream polarization-based computer vision algorithms directly take the measurements as their model inputs [4–6]. It is clear that the equations 1 and 2 only hold for unprocessed raw images or those processed by linear amplifications, e.g., digital gain and white balance. Non-linear computations in ISP, e.g., gamma correction and tone mapping, will destroy the implicit constraints. Moreover, the quantization process of converting 12-bit data to 8-bit values of bright images will introduce quantization noises and reduce accuracy, too [9]. As shown in Figure 1, we exhibit the rendered sRGB images and extracted polarization information of the original raw data and processed data. Among the results, although quantization processing generates almost the same intensities, the diminished precision causes completely wrong ρ and ϕ. We apply gamma correction for data nonlinearization operation. The intensity seems more pleasant to human vision, but the distribution of ρ is noticeably changed. The results show the negative effects of non-linear ISP and quantization on polarization information measurements. This is why we focus on raw-domain processing in this work.

2. Noise model calibration

Here, we provide more noise model calibration results of the IMX250MYR sensor. As shown in Fig. 2, the distribution of parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}_b$ and $\boldsymbol{\sigma}_\tau$ of each channel at $Gain = 24$ varies greatly according to the color and polarization patterns. Fig. 3 illustrates partial results of the linear regression of $\left(\log \hat{K}, \log \boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right)$ and $\left(\log \hat{K}, \log \boldsymbol{\Sigma}_b\right)$. As the covariance of 16 channels is computed, the total regression leads to a 16×16-dimensional linear model. Tendencies of covariances between positively correlated channels satisfy the linear regression model well and covariances between less correlated channels are extremely small that they can be ignored.



(a) $\boldsymbol{\mu}$

(b) $\boldsymbol{\sigma}_b$

(c) $\boldsymbol{\sigma}_\tau$

Figure 2. Box plots of $\boldsymbol{\mu}$, $\boldsymbol{\sigma}_b$ and $\boldsymbol{\sigma}_\tau$ of each channel at $Gain = 24$. The red dashed lines are the samples' averages, and the orange lines and green dashed lines in boxes represent the median and mean values of samples of each channel.

3. Dataset capture

Camera gain and overall system gain. Camera gain is a controllable parameter for the polarization color camera accessed from its interface. It is different from overall system gain, but can be converted through:

$$G(K) = 20 \log \frac{K}{K_0}, \qquad K(G) = K_0 10^{G/20.}, \tag{3}$$

where G and K denote camera gain and corresponding overall system gain, respectively. K_0 is the overall system gain when $G = 0$. From the equations, it is easy to know that, K is doubled when adding 6 to G, e.g., $20 \log 2 \approx 6$. Therefore, Gs selected in our work are $[0, 6, 12, 18, 24]$ and similar to ISO settings of $[100, 200, 400, 800, 1600]$. $G_0 = 0$ is used for reference images and others are applied for noisy images. Then, our detailed data capture protocol is summarized as follows: Two low light factors, 10 and 60 are selected for capturing low light images.

```
Require: $G_0 = 0; K(G)t^* = K_0 t$
for each scene do
    Find a proper shutter speed $t$ that well exposes the scene at camera gain $G_0$;
    Take 100 images at exposure setting $(G_0, t)$ to generate a reference image;
    for each camera gain $G$ do
        $t^* = \frac{K_0}{K(G)} t = t/10^{\frac{G}{20}}$;
        for each low light factor $f$ do
            Capture 5 noisy image at $(G, t^*/f)$;
            Select a noisy image whose intensity is most close to the reference after rescaled by $\times f$;
        end
    end
end
```

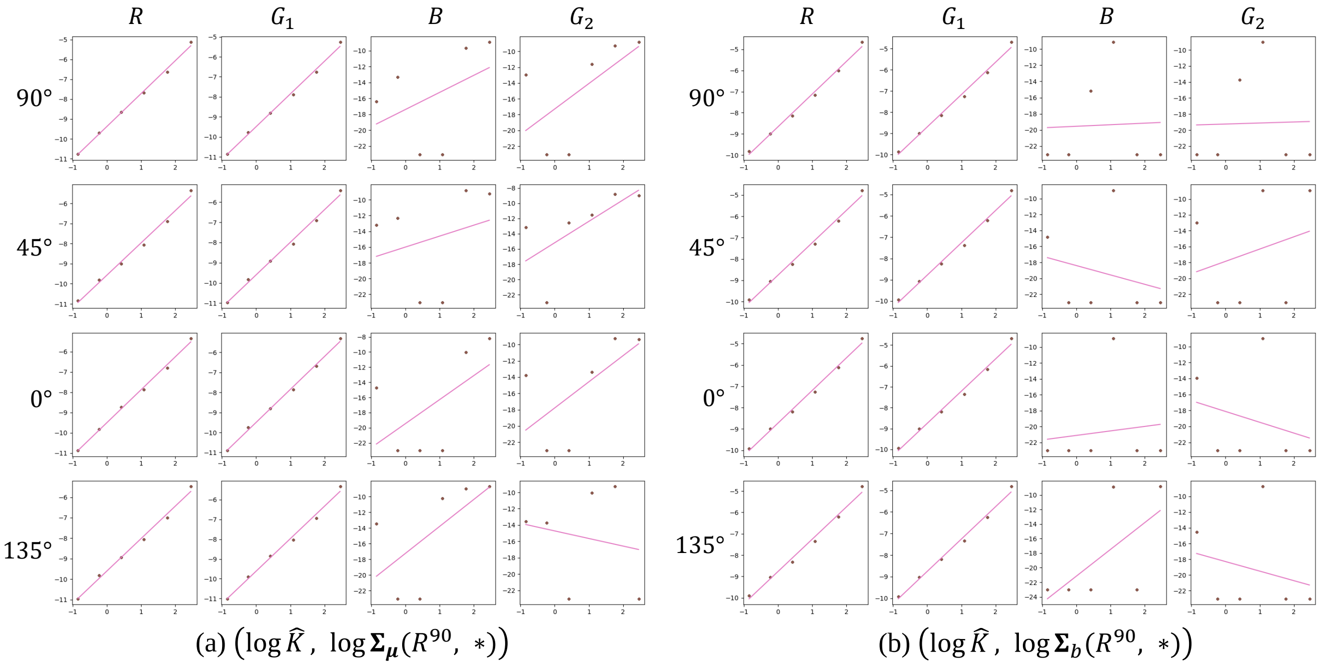
Algorithm 1: Image capture protocol

4. Transformer model

We introduce detailed architecture of proposed Transformer model.

Overall architecture. As shown in Figure 4a, our Transformer model firstly extracts low-level feature embeddings with a 3×3 convolution layer. After that, 3 encoder stages, a bottleneck stage and 3 decoder stages, with successive Transformer blocks in each stage, are applied to handle multi-scale features. After each encoder stage and before each decoder stage, downsample layers and upsample layers, following the implementation of [8], are employed to reduce or recover feature resolutions. Low-level encoder features are delivered to the corresponding decoder stages, where the features of the former decoder are concatenated. Except for the last decoder stage, the channel of concatenated features are halved by a linear projection layer. At last, a 3×3 convolution layer is applied to output a residual to noisy input and generate a denoised image. After the last encoder stage, the feature are converted back into 2D feature maps. Then we establish a skip connection between the shallow and the deep features.

Transformer block. Different from conventional image denoising, polarized color image denoising aims to restore clean signals from noisy inputs while preserving precise polarization information at the same time. To meet the needs, we propose the Transformer block build by stacking (Shifted) Window Multi-heads Attention ((S)W-MA), Window Multi-Shuffled-heads Transposed Attention (W-MT(S)A), and a Locally-Enhanced MLP, as shown in the Figure 4b. While SW-MA is applied to capture spatial attention towards a better denoising performance, SW-MTSA is expected to model interconnections of the channels and



(a) $\left(\log \widehat{K}\,,\ \log \boldsymbol{\Sigma}_{\boldsymbol{\mu}}(R^{90},\ *)\right)$ (b) $\left(\log \widehat{K}\,,\ \log \boldsymbol{\Sigma}_{b}(R^{90},\ *)\right)$

Figure 3. The linear regression results for the joint distribution of (a) $\left(\log \hat{K}, \log \boldsymbol{\Sigma}_{\boldsymbol{\mu}}(R^{90}, *)\right)$ and (b) $\left(\log \hat{K}, \log \boldsymbol{\Sigma}_{b}(R^{90}, *)\right)$ at $Gain = 24$, where $\boldsymbol{\Sigma}(a, b)$ represents the covariance of $\boldsymbol{a}$ and $\boldsymbol{b}$ samples, $*$ denotes arbitrary channels. The labels indicate the channels, e.g., R^{90}. The results indicate lifted parameter covariances $\log \boldsymbol{\Sigma}$ (y-axis) between positively correlated channels satisfy the linear function to $\log \hat{K}$ (x-axis) well.

explore the polarization priors that existed implicity in channle dimension. A LayerNorm (LN) layer is employed before each module (omitted in Figure 4) and a residual connection is applied after each module.

Shifted Window-based Multi-head Attention. Conventional attention modules have shown great capabilities in modeling spatial long-term dependencies for image restoration tasks [2, 12]. However, the time and memory complexity of the key-query dot-product are quadratic to the spatial resolution of input, so the global attention module is unavailable for our high-resolution data. Compared with global attention, SW-MA has shown a better speed-accuracy trade-off [1] with local attention and shifted window strategy. Applying the learnable relative position encoding $\mathbf{B}$ [1], the attention calculation can be formulated as:

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{SoftMax}(\mathbf{Q}\mathbf{K}^{\mathbf{T}}/\alpha + \mathbf{B})\mathbf{V}, \quad (4)$$

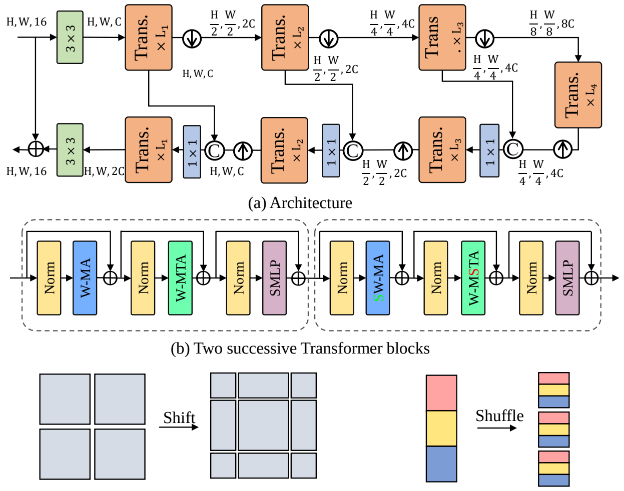
where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{l^2 \times d}$ are the query, key, and value matrices encoded with the same input; α is a learnable temperature parameter for adjus SoftMax activation; l^2 denotes the window size and d represents the number of channels. As the stages get deeper, the SW-MA can capture longer spatial dependencies. Regular and shifted window partitioning is applied alternatively to facilitate interactions across windows.

Window based Multi-Shuffled-heads Transposed Attention. While spatial attention focuses on modeling interactions of features in spatial domain, it neglects to explore physical characteristics existing implicitly in the channel dimension. Therefore, we propose W-MSTA to apply attention across channels. The transposed attention is defined as:

$$\mathrm{TAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{V} \cdot \mathrm{SoftMax}(\mathbf{Q}^{\mathbf{T}}\mathbf{K}/\beta), \quad (5)$$

where β is a learnable parameter, too. Since conventional multi-heads attention divides channels into multiple groups, the interactions through channels are insufficient. To tackle the issue, we introduce the channel shuffle operation [11] to enable intensive connections across channel groups. Following [11], the original features are firstly shuffled along the channel dimension, according to feature dimension and the number of heads. Then the channel-shuffled features are partitioned into non-overlapped windows and fed into the transposed attention module. Lastly, after the attention, the channel-enhanced features are shuffled back to the original arrangement.

Locally-Enhanced MLP. The window participation approach can cause border artifacts even with shift operations, especially for sensitive polarization information. A 3 × 3 depth-wise convolution is employed after the first linear projection layer and followed by GELU activation. The use



(a) Architecture

(b) Two successive Transformer blocks

(c) Shifted window for W-MA

(d) Shuffled heads for W-MTA

Figure 4. (a) The architecture of proposed Transformer model. ↑, ↓, + and C in the circles denote upsample layer, downsample layer, element-wise addition, and feature concatenation, respectively. (b) Two successive Transformer blocks. W-MA and SW-MA are Window-based Multi-head self Attention with regular and window Shift strategies, respectively. W-MTA and W-MSTA are Window-based Multi-head Transposed self Attention with regular and channel Shuffle operations, respectively. (c) and (d) are illustrations of shifted windows and shuffled heads.

of convolution layers in a Transformer block plays crucial roles in enhancing translation invariance in Transformer-based networks [3, 7].

5. Experiment

For proposed Transformer model, from the first encoder stage to the last decoder stage, the number of stacked Transformer blocks are $[2, 6, 6, 8, 6, 6, 2]$, the number of channels are $[48, 96, 192, 384, 192, 96, 96]$, and the number of heads are $[1, 2, 4, 8, 4, 2, 2]$.

5.1. comparison

We show more visual comparisons in Figure 5. It can be observed that Uformer*, Restormer*, and Ours*, trained on synthetic noisy images, restore more vivid details and decrease the oversmoothing issue. Moreover, due to the window-based hybrid attention mechanism, our denoising model is able to remove noises and restore sharp details for both images and polarization information. We further count the number of parameters and FLOPs of Uformer, Restormer, and the proposed model, as (50.89 M, 345.84 Mac), (27.03 M, 452.41 Mac), and (23.42 M, 346.14 Mac). The proposed model outperforms Uformer and Restormer with noticeably decreased parameters and FLOPs.

Table 1. Polarized color image denoising performance in PSNR(dB)/SSIM calculated on images, DoLP and AoLP. **Bold** values present the best best results. N_{read} and N^*_{read} represent read out noises sampled with zero-mean and channel-wise biases respectively.

	image	DoLP	AoLP
	PSNR/ SSIM	PSNR	PSNR
(a) paired w/ bias	33.01/0.922	23.23	14.75
(b) paired	**34.42/0.927**	23.56	14.89
(c) N_{read}	31.56 /0.849	22.79	13.12
(d) $N_{read} + N_p$	33.81/ 0.907	23.64	14.96
(e) $N_{read} + N_p + N_b + N_q$	33.94/0.914	23.69	15.04
(f) $N^*_{read} + N_p + N_b + N_q$	33.93/0.915	23.65	15.07
(g) W/o joint dist. model	33.60/0.927	**23.89**	15.03
(h) W/o $\mathcal{L}_S$	33.93/ 0.914	23.74	14.92
(i) W/o W-MSTA	33.84/0.913	23.64	**15.15**
(j) W/o SW-MA	33.73/0.914	23.74	14.92

5.2. Ablation

Here, more ablation experiment results are exhibited and all ablation experiments are conducted based on our proposed Transformer model with noisy images at ×60 low-light ratio for a significant comparison.

Table 1 (f) and (g) show the results of the model trained with synthetic data generated with our joint noise distribution or that proposed by [10], in which gain-wise and channel-wise noise formulations are not considered. The results indicate our proposed joint distribution can significantly improve the performance of both image denoising and polarization restoration. The comparisons in Fig. 6 (f), (g) indicate our proposed joint noise distribution benefits to image denoising and sharp texture reconstruction.

Table 1 (i), (j), and (f) show ablation experiment results on the proposed Transformer model consisting of (Shifted) Window Multi-heads Attention ((S)W-MA, spatial-domain) and Window Multi-Shuffled-heads Transposed Attention (W-MT(S)A, channel-domain). The comparisons suggest a spatial-domain self-attention contributes more to denoising performance than channel-domain self-attention, but their combination significantly improves the restoration. The results in Fig. 6 show the final proposed model (f) combining spatial and channel domain self-attention restores vivid details than other settings (i), (j).

References

[1] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021. 3

[2] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *CVPR*, pages 11065–11074, 2019. 3

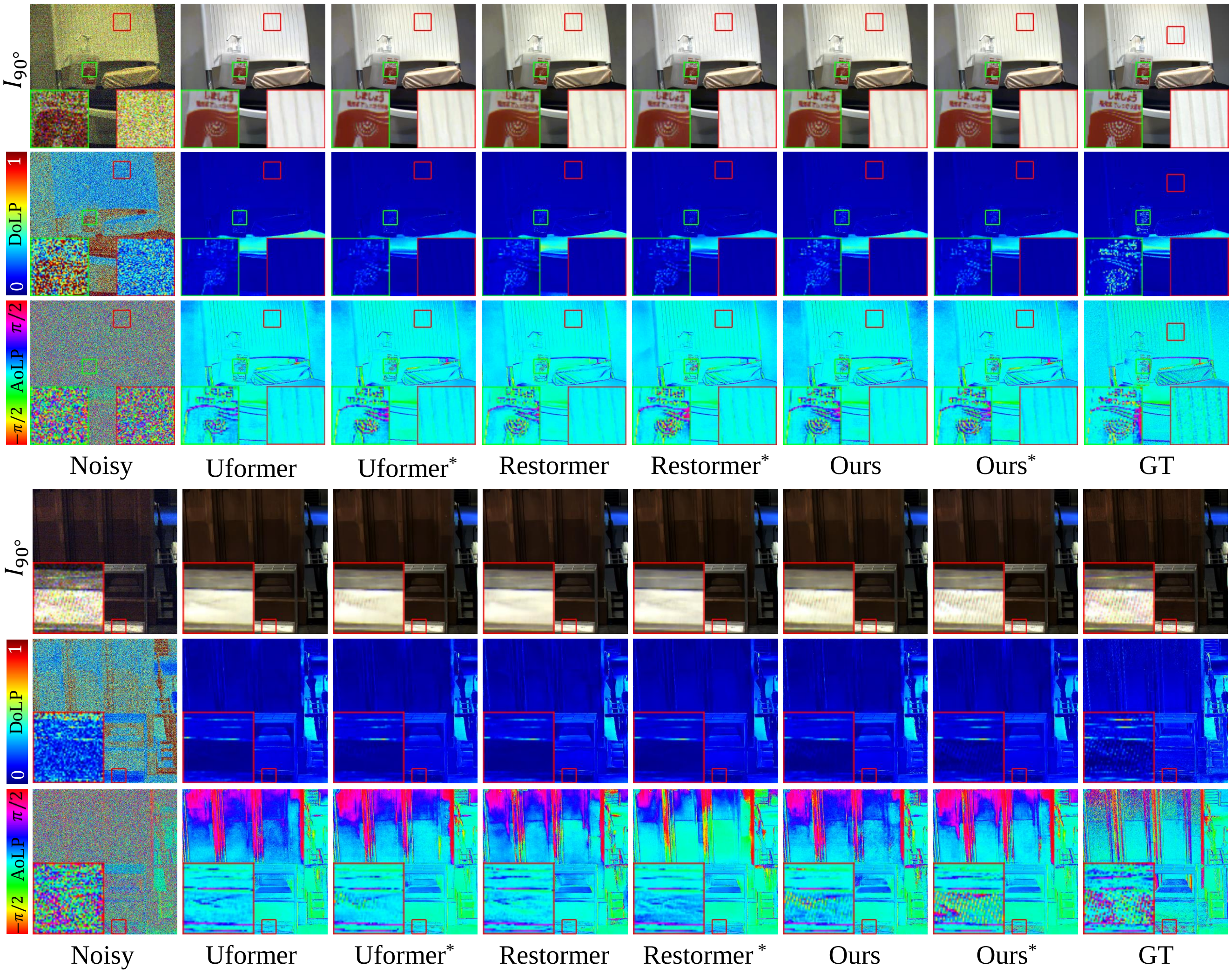


Figure 5. Visual comparison for polarized color image denoising. I_{90°, DoLP and AoLP are exhibited and ”*” represents the model is trained on synthetic noisy images generated via our noise model.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[4] Yoshiki Fukao, Ryo Kawahara, Shohei Nobuhara, and Ko Nishino. Polarimetric normal stereo. In *CVPR*, pages 682–690, 2021. 1

[5] Agastya Kalra, Vage Taamazyan, Supreeth Krishna Rao, Kartik Venkataraman, Ramesh Raskar, and Achuta Kadambi. Deep polarization cues for transparent object segmentation. In *CVPR*, pages 8602–8611, 2020. 1

[6] Chenyang Lei, Xuhua Huang, Mengdi Zhang, Qiong Yan, Wenxiu Sun, and Qifeng Chen. Polarized reflection removal with perfect alignment in the wild. In *CVPR*, pages 1750–1758, 2020. 1

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *ICCV*, 2021. 4

[8] Zhendong Wang, Xiaodong Cun, Jianmin Bao, and Jianzhuang Liu. Uformer: A general u-shaped transformer for image restoration. *arXiv preprint arXiv:2106.03106*, 2021. 2

[9] Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In *CVPR*, pages 2758–2767, 2020. 1

[10] Kaixuan Wei, Ying Fu, Yinqiang Zheng, and Jiaolong Yang. Physics-based noise modeling for extreme low-light photography. *IEEE TPAMI*, 2021. 4

[11] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, pages 6848–6856, 2018. 3

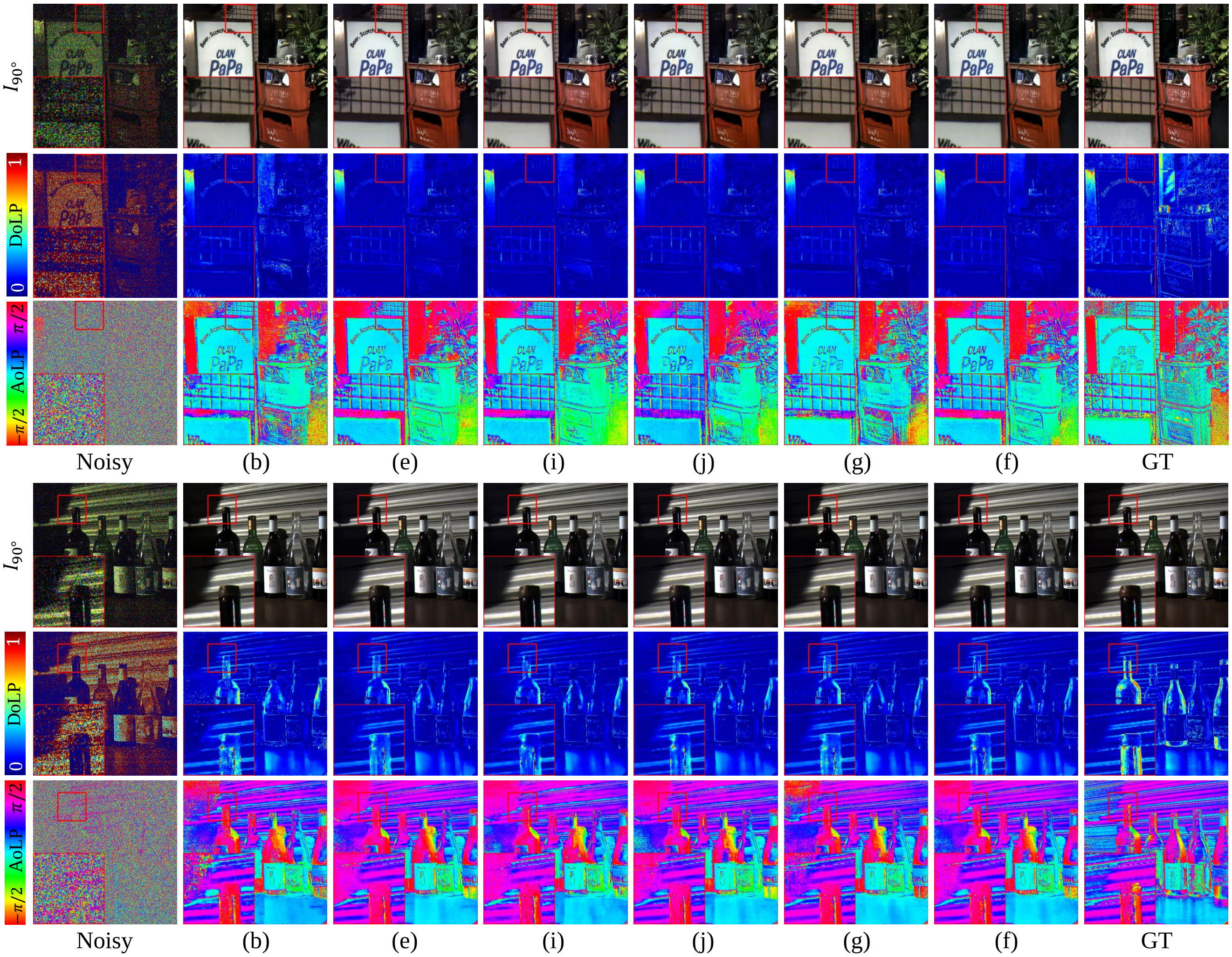


Figure 6. Visual comparison for ablation study. $I_{90°}$, DoLP and AoLP are visualized and the labels represent settings presented in Table 1.

[12] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019. 3