Supplementary Material
SwitchLight: Co-design of Physics-driven Architecture and Pre-training Framework for Human Portrait Relighting

1. Implementation Details

Training Our method incorporates both pre-training, spanning for 100 epochs, and fine-tuning phases lasting 50 epochs. In the pre-training stage, we utilized the ImageNet dataset, using a batch size of 1024 images, each with a resolution of 256 × 256 pixels. We employed the Adam optimizer; 10k linear warm-up schedule followed a fixed learning rate of $1e^{-4}$. In the fine-tuning stage, we switched to the OLAT datset, with a batch size reduced to 8 images, each at a resolution of 512×512 pixels. The Adam optimizer is used with a fixed learning rate of $1e^{-4}$. The entire training process takes one week to converge using 32 NVIDIA A6000 GPUs.

We pre-train a single U-Net architecture during this process. In the subsequent fine-tuning stage, the weights from this pre-trained model are transferred to multiple U-Nets - NormalNet, DiffuseNet, SpecularNet, and RenderNet. In contrast, IllumNet, which does not follow the U-Net architecture, is initialized with random weights. To ensure compatibility with the varying input channels of each network, we modify the weights as necessary. For example, weights pre-trained for RGB channels are copied and adapted to fit networks with 6 or 9 channels.

Data To generate the relighting training pairs, we randomly select each image from the OLAT dataset. Two randomly chosen HDRI lighting environment maps are then projected onto these images to form a training pair. The images undergo processing in linear space. For managing the dynamic range effectively, we apply logarithmic normalization using the $\log(1 + x)$ function.

Architecture SwitchLight employs a UNet-based architecture, consistently applied across its **Normal** Net, **Diffuse** Net, **Specular** Net, and **Render** Net. This approach is inspired by recent advancements in diffusion-based models [2]. Unlike standard diffusion methods, we omit the temporal embedding layer. The architecture is characterized by several hyperparameters: the number of input channels, a base channel, and channel multipliers that determine the channel count at each stage. Each downsampling stage features two residual blocks, with attention mechanisms integrated at certain resolutions. The key hyperparameters and their corresponding values are summarized in Table. 1.

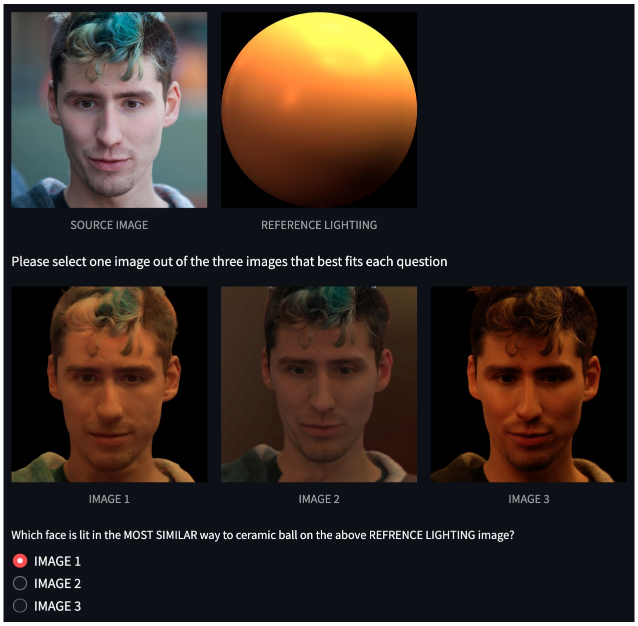
	Normal Net	Diffuse Net	Specular Net	Render Net
In ch	3	6	9	9
Base ch	64	64	64	64
Ch mults	[1,1,2,2,4,4]	[1,1,2,2,4,4]	[1,1,2,2,4,4]	[1,1,2,2,4,4]
Num res	2	2	2	2
Head ch	64	64	64	64
Att res	[8,16,32]	[8,16,32]	[8,16,32]	[8,16,32]
Out ch	3	3	2	3

Table 1. **Network Architecture Parameters.** This table outlines the key hyperparameters and their corresponding values; initial input channels (In ch), base channels (Base ch), and channel multipliers (Ch mults) that set the stage-specific channel counts. It also indicates the number of residual blocks per stage (Num res), the number of channels per head (Head ch), the stages where attention mechanisms are applied based on feature resolution (Att res), and the final output channels (Out ch).

IllumNet is composed of two projection layers, one for transforming the Phong lobe features and another for image features, with the latter using normal bottleneck features as a compact form of image representation. Following this, a cross-attention layer is employed, wherein the Phong lobe serves as the query and the image features function as both key and value. Finally, an output layer generates the final convolved source HDRI.

The **Discriminator** network is utilized during both pre-training and fine-tuning stages, maintaining the same architectural design, although the weights are not shared between these stages. This network is composed of a series of residual blocks, each containing two 3×3 convolution layers, interspersed with Leaky ReLU activations. The number of filters progressively increases across these layers: 64, 128, 256, and 512. Correspondingly, as the channel filter count increases, the resolution of the features decreases, and finally, the network compresses its output with a 3x3 convolution into a single channel, yielding a probability value.

Regarding the activation functions across different networks: NormalNet processes its outputs through ℓ_2 normalization, ensuring they are unit normal vectors. IllumNet, DiffuseNet, and RenderNet utilize a softplus activation (with



Please select one image out of the three images that best fits each question

Which face is lit in the MOST SIMILAR way to ceramic ball on the above REFRENCE LIGHTING image?

● IMAGE 1
○ IMAGE 2
○ IMAGE 3

Figure 1. **User Study Interface** comparing relighting results with prior approaches, focusing on consistency in lighting, preservation of facial details, and retention of original identity.

$\beta = 20$) to generate non-negative pixel values. SpecularNet employs a sigmoid activation fuction, ensuring that both the roughness parameter and Fresnel reflectance values fall within a range of 0 to 1.

2. User Study Interface

Our user study interface is outlined as follows: Participants are shown an input image next to a diffused ball under the target environment map lighting. The primary objective is to compare our relighting results with baseline methods, as depicted in Fig. 1. Evaluation focuses on three criteria: 1) Consistency of lighting, 2) Preservation of facial details, and 3) Retention of the original identity. This comparison aims to determine which image best matches the lighting of the diffused ball while also maintaining facial details and original identity. To ensure unbiased evaluations, we randomized the order of presentation. Participants evaluated 30 random samples from a set of 256. This dataset included 32 portraits from the FFHQ dataset [3], each illuminated under eight distinct lighting conditions.

3. Video Demonstration

We present a detailed video demonstration of our SwitchLight framework. Initially, we use real-world videos from Pexels [1] to showcase its robust generalizability and practicality. Then, for state-of-the-art comparisons, we utilize the FFHQ dataset [3] to demonstrate its advanced relighting capabilities over previous methods. The presentation includes several key components:

1. **De-rendering:** This stage demonstrates the extraction of normal, albedo, roughness, and reflectivity attributes from any given image, a process known as inverse rendering.

2. **Neural Relighting:** Leveraging these intrinsic properties, our system adeptly relights images to align with a new, specified target lighting environment.

3. **Real-time Physically Based Rendering (PBR):** Utilizing the Three.js framework and integrating derived intrinsic properties with the Cook-Torrance reflectance model, we facilitate real-time rendering. This enables achieving 30 fps on a MacBook Pro with an Apple M1 chip (8-core CPU and 8-core GPU) and 16 GB of RAM.

4. **Copy Light:** Leveraging SwitchLight's ability to predict lighting conditions of a given input image, we explore an intriguing application. This process involves two images, a source and a reference. We first extract their intrinsic surface attributes and lighting conditions. Then, by combining the source intrinsic attributes with the reference lighting condition, we generate a new, relit image. In this image, the source foreground remains unchanged, but its lighting is altered to match that of the reference image.

5. **State-of-the-Art Comparisons:** We benchmark our framework against leading methods, specifically Total Relight [4] and Lumos [6], to highlight substantial performance improvements over these approaches.

4. Additional Quantitative Results

Quantitative comparison on decomposed intrinsics

Normal	MAE ↓	MSE ↓	SSIM ↑	LPIPS ↓
TR	0.1315	0.3517	**0.854**	**3.195**
SwitchLight	**0.1306**	**0.3455**	0.8514	3.238

Table 2. Quantitative Evaluation of *Normals* on the OLAT test set.

Albedo (log)	MAE ↓	MSE ↓	SSIM ↑	LPIPS ↓
TR	0.8449	0.0173	0.9282	0.1706
SwitchLight	**0.8097**	**0.0156**	**0.9342**	**0.1618**

Table 3. Quantitative Evaluation of *Albedos* on the OLAT test set.

As shown in Table 2, our method outperforms TR framework in MAE/MSE for normal prediction but lags in SSIM/LPIPS, likely due to the emphasis of SSIM and LPIPS on perceptual similarity rather than geometric accuracy. In contrast, as indicated in Table 3, our method surpasses the TR Framework in all metrics for albedo prediction. Overall, our approach offers a more precise and consistent estimation of intrinsics.

Influence of pre-training data

Relit	MAE ↓	MSE ↓	SSIM ↑	LPIPS ↓
Ours	0.1023	0.0275	0.9002	2.137
w. MMAE H	0.0936	0.0237	0.9060	2.068
w. MMAE IM	**0.0933**	**0.0235**	**0.9051**	**2.059**

Table 4. The impact of pretraining and its dataset characteristics.

We investigate the influence of pre-training data characteristics. We specifically pre-train on either human-specific dataset (‘H’) or ImageNet (‘IM’), ensuring both datasets approximately equal in size. The results in Table 4 reveal that pre-training with ImageNet yields better results. This suggests that using a broad dataset is more preferable to task-specific data for learning transferable features for the religthing task.

5. Visualization of estimated illumination

In Fig.2, we present a visualization of the light estimates as inferred by Illum net. Specifically, we focus on one of the four convolved HDRIs predicted by Illum net to represent the light estimate, selecting the convolved HDRI characterized by a shininess exponent of $p = 16$. The light estimates clearly reflect the directionality, color, and intensity of the lighting conditions, illustrating the effective estimation capabilities of Illum net.

6. Additional Qualitative Results

Further qualitative results are provided in Fig.3, 4, 5, 6, and 7. Each figure illustrates the relighting of a source image in eight distinct target lighting environments. In these figures, our approach is benchmarked against prior state-of-the-art methods, namely SIPR [5], Lumos [6], and TR [4], utilizing images from Pexels [1]. This comparison is enabled by the original authors who applied their models to identical inputs and provided their respective outputs.

We can clearly observe that our method demonstrates notable efficacy in achieving consistent lighting, maintaining softness and high-frequency detail. Additionally, it effectively manages specular highlights and hard shadows, while meticulously preserving facial details, identity, skin tones, and hair texture.

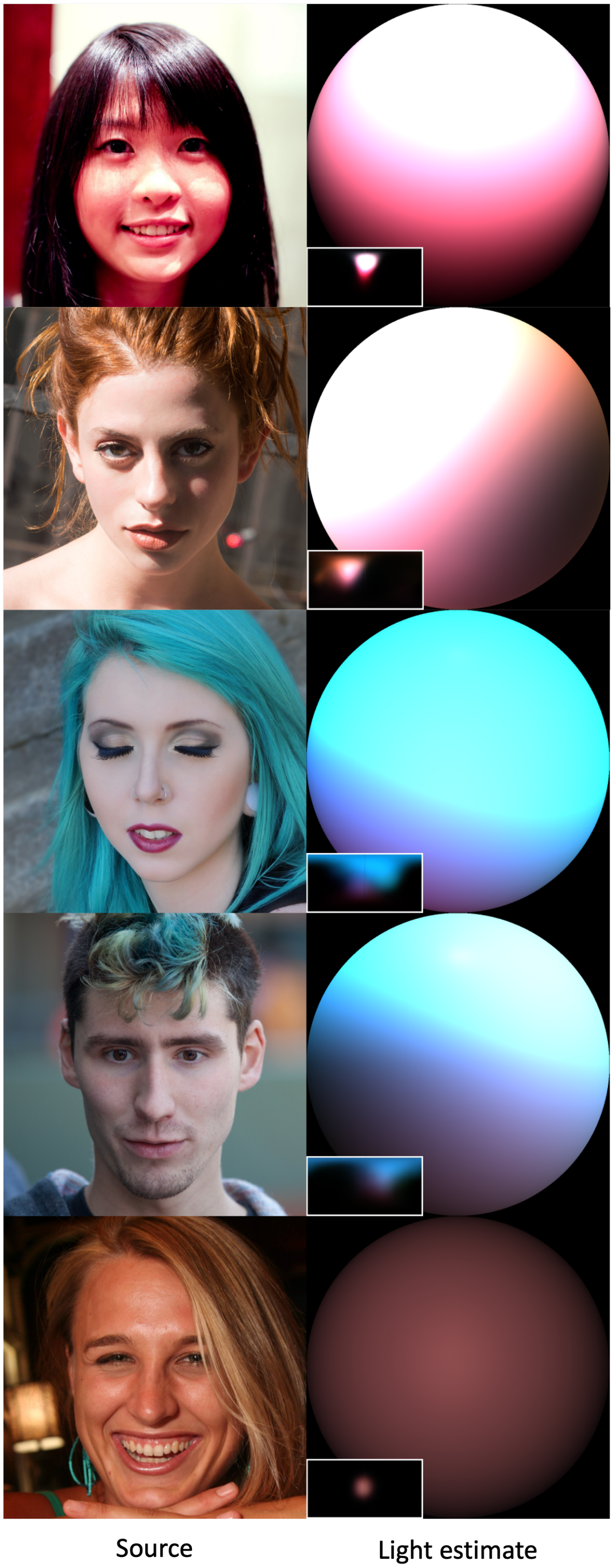


Figure 2. **Visualization of Light Estimates** : Illum net accurately captures the color, intensity, and directionality of lighting conditions across various portraits

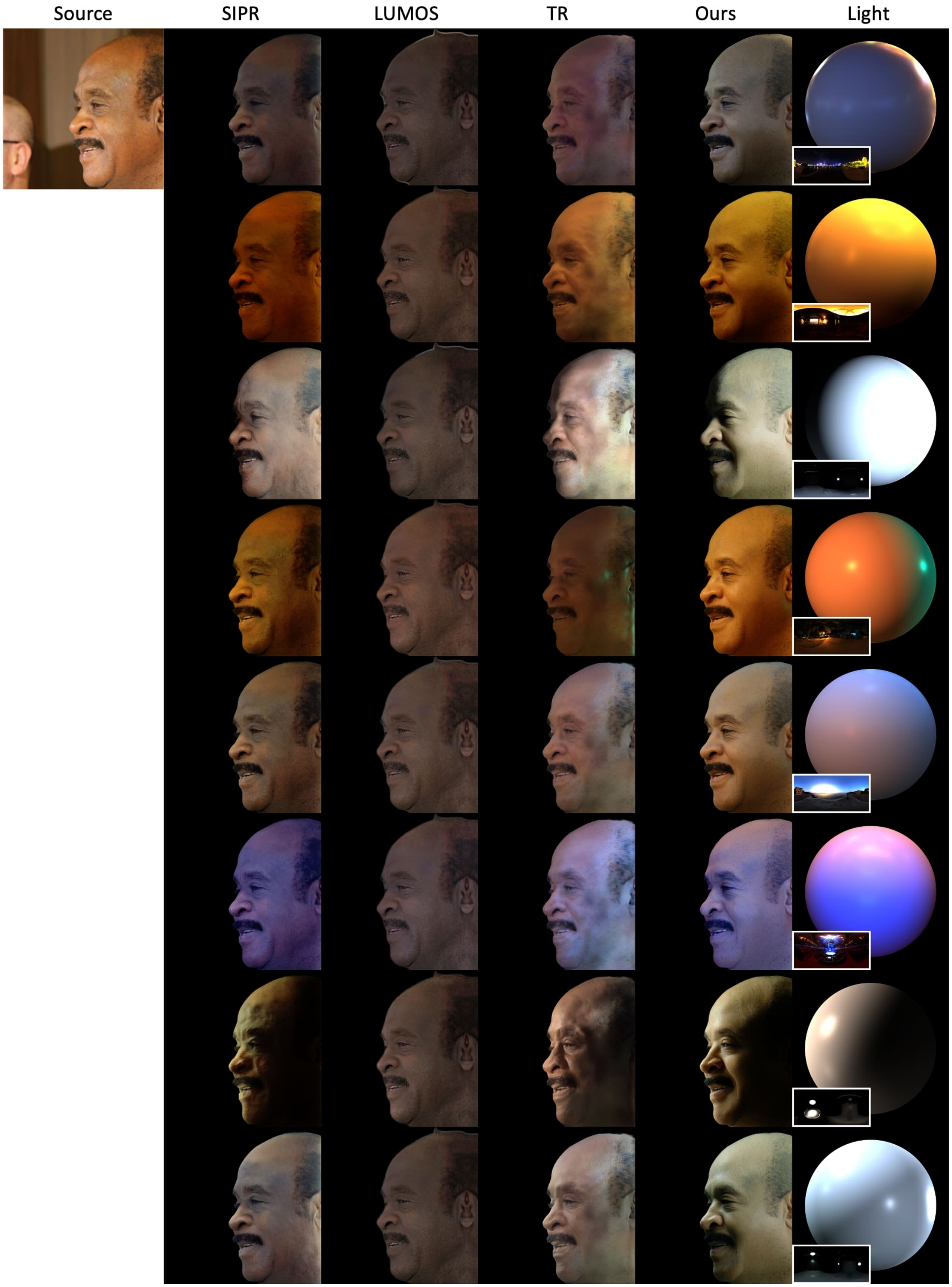


Figure 3. **Qualitative Comparisons** with state-of-the-art approaches.

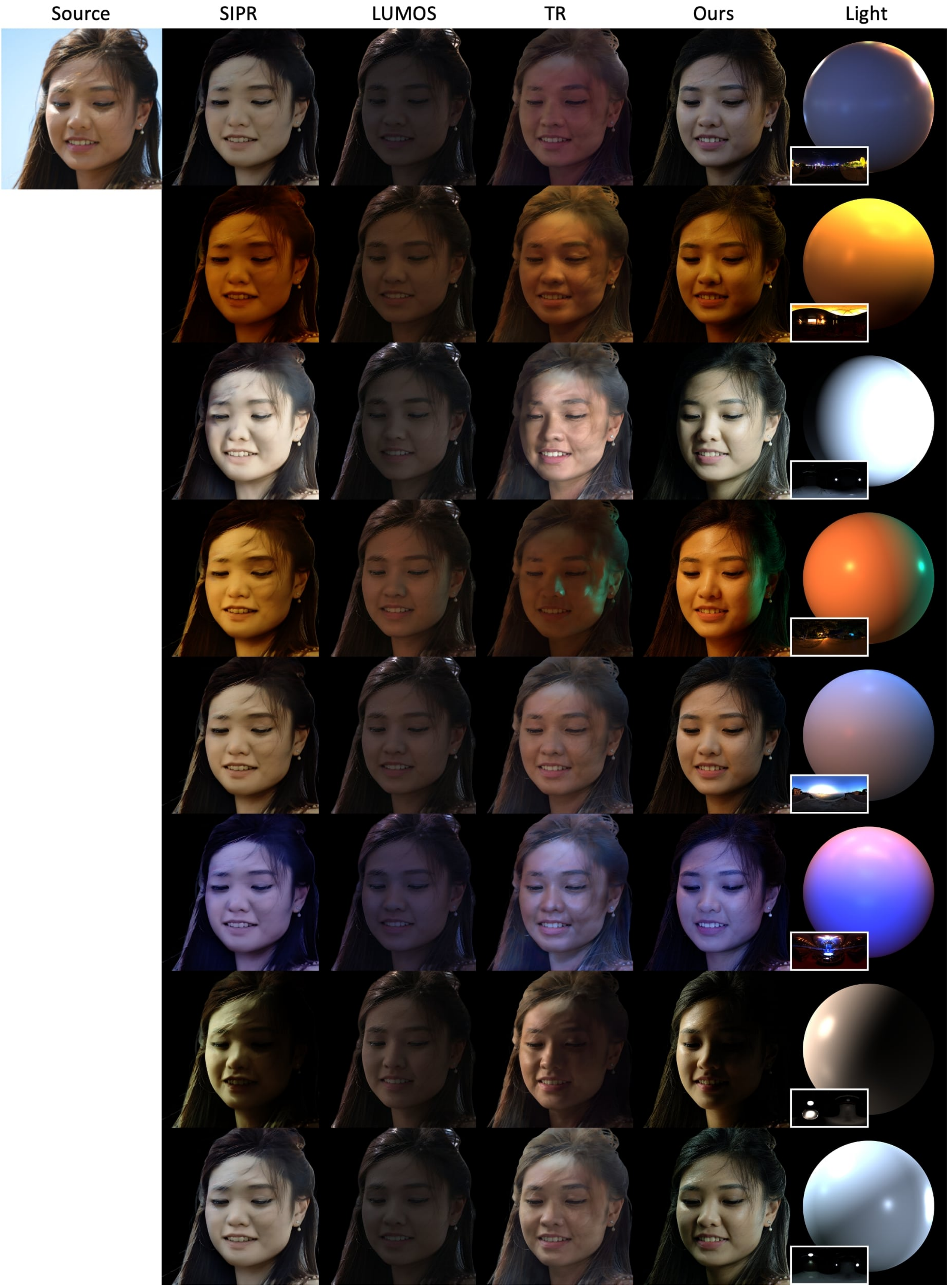


Figure 4. **Qualitative Comparisons** with state-of-the-art approaches.

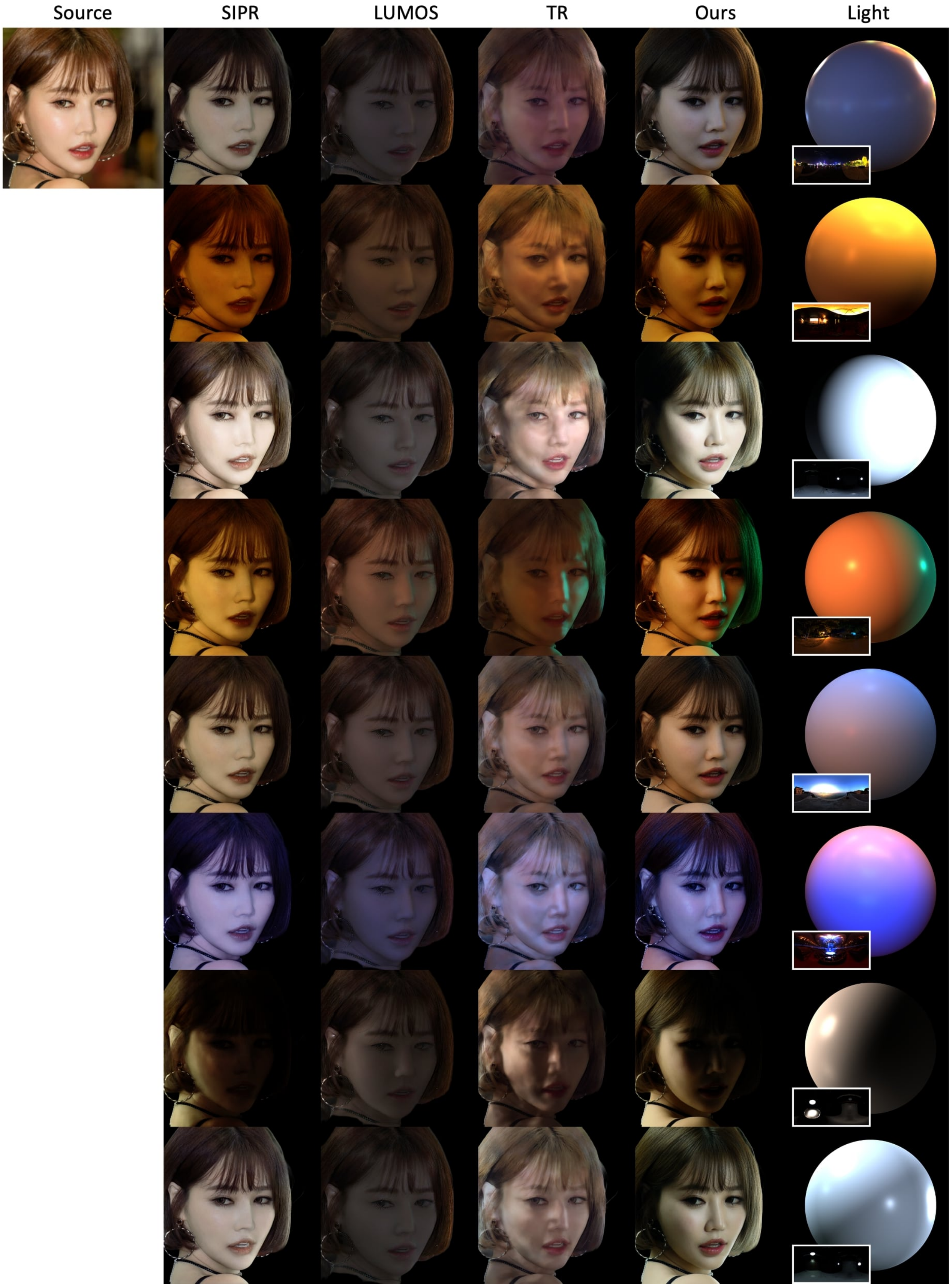


Figure 5. **Qualitative Comparisons** with state-of-the-art approaches.

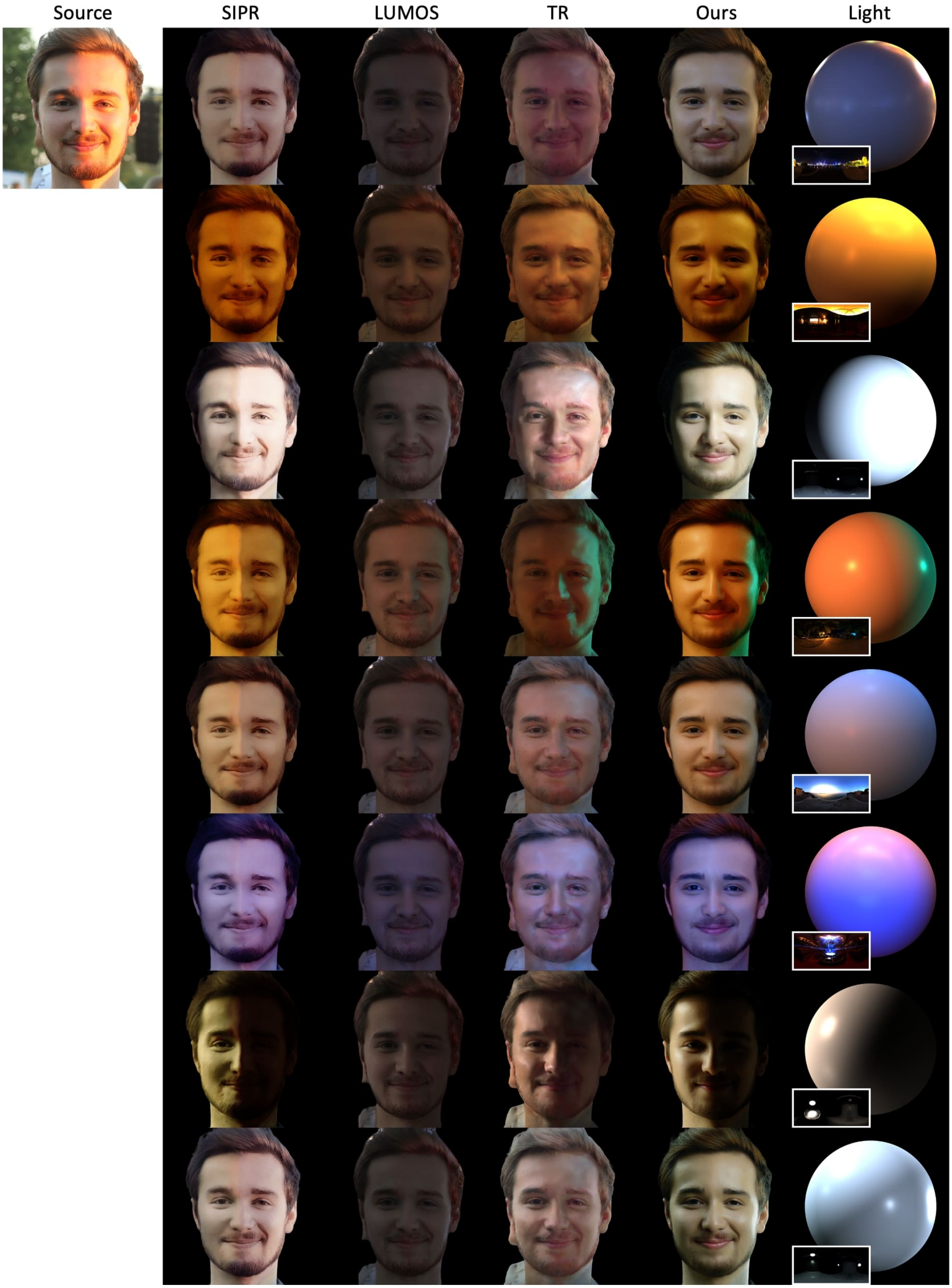


Figure 6. **Qualitative Comparisons** with state-of-the-art approaches.

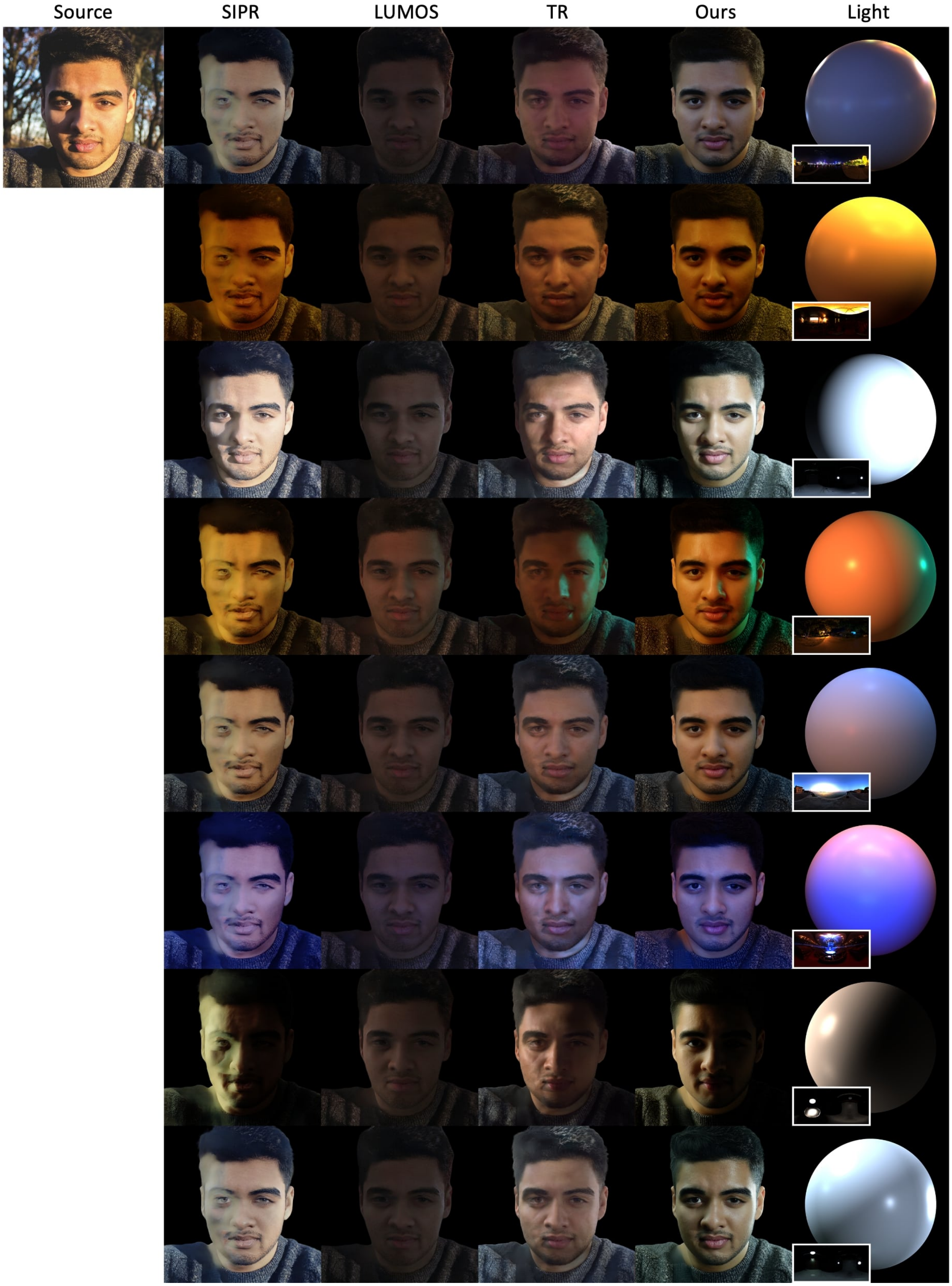


Figure 7. **Qualitative Comparisons** with state-of-the-art approaches.

References

[1] Pexels. https://www.pexels.com. 2, 3

[2] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1

[3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[4] Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG)*, 40(4):1–21, 2021. 2, 3

[5] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 3

[6] Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. *ACM Transactions on Graphics (TOG)*, 41(6):1–21, 2022. 2, 3