LAB ShadowFormer

Harshil Bhojwani
Advanced Perception - CS 7180

Dhanush Adithya Balamurugan
Advanced Perception - CS 7180

***Abstract*—Shadow removal has been a challenging problem in computer vision, with recent advancements leveraging deep learning-based architectures like ShadowFormer achieving notable performance. In this work, we propose enhancing ShadowFormer by converting the input representation to alternative color spaces, such as LAB, and incorporating novel architectural modifications. Inspired by the success of LAB color space in CNN-based models, such as LAB-Net, we aim to evaluate its compatibility with transformer-based architectures. Architectural tweaks, such as adding transformer blocks and revising loss functions, will be investigated to ensure the model adapts effectively to these changes. We will evaluate combinations of color space transformations, architectural enhancements, and advanced loss functions through systematic experimentation to determine their impact on shadow removal performance. Our study pioneers the integration of LAB color space with transformer-based models, bridging the gap between prior CNN-based approaches and the emerging transformer paradigm. The proposed modifications aim to establish a new benchmark for shadow removal by enhancing ShadowFormer's capability to handle complex scenes with higher fidelity.**

I. Introduction

Shadow removal is a fundamental task in computer vision, with applications spanning object detection, image editing, and scene understanding. The goal is to eliminate shadows from photos while maintaining the underlying textures and features, which is difficult because of the intricate relationship between illumination and reflectance in darkened areas. Conventional methods sometimes depend on manually created characteristics or presumptions about illumination, but they perform poorly in a variety of real-world situations.

Recent developments in deep learning have greatly enhanced the ability to remove shadows, with state-of-the-art outcomes being achieved by architectures such as ShadowFormer. In order to better understand shadowed and non-shadowed areas in photos, ShadowFormer uses transformer-based architectures to capture long-range dependencies. Although these transformer-based techniques have demonstrated promise, more research into different input representations and architectural improvements can help them reach their full potential.

Motivated by the LAB colour space's success in models based on convolutional neural networks (CNNs), such as LAB-Net, we suggest improving ShadowFormer by transforming input photos into the LAB colour space. By making it easier to characterise shadows as purely luminance fluctuations, the LAB representation—which distinguishes luminance from chromatic components—has shown promise in shadow removal tasks. By combining LAB space with transformer-based designs, we hope to close the gap between the successful CNN-based methods and the new transformer paradigm.

This study also explores new architectural changes to ShadowFormer, like improving loss functions and including transformer blocks. These improvements aim to overcome difficulties in handling complicated scenes with high fidelity and increase the model's responsiveness to the new input representation. Insights into how colour space transformations, architectural upgrades, and sophisticated loss designs interact to affect shadow removal performance will be obtained through methodical testing.

This work advances the field by being the first to combine transformer-based models for shadow removal with LAB colour space. With these initiatives, we hope to raise the bar for shadow removal and improve ShadowFormer's capacity to handle a variety of difficult situations.

II. Related works

Shadow removal has been extensively studied, with approaches evolving from traditional methods relying on handcrafted features to modern deep learning-based solutions. Recent works have explored CNNs and transformer architectures, leveraging their ability to effectively model complex relationships in shadowed scenes.

A. *ShadowFormer*

ShadowFormer is a Retinex-based shadow removal model that exploits information from non-shadow regions to restore shadowed areas. It utilizes a multi-scale channel attention framework to hierarchically capture global information across the image, enabling more effective reconstruction of shadow-affected regions. A key innovation in ShadowFormer is the Shadow-Interaction Module (SIM), which incorporates Shadow-Interaction Attention (SIA) in the bottleneck region to model the contextual correlation between shadow and non-shadow areas. This design allows ShadowFormer to achieve state-of-the-art performance by effectively integrating spatial and contextual information for enhanced shadow removal.

B. *LAB-Net*

LAB-Net is a lightweight deep neural network designed to process shadow images in the LAB color space, leveraging its ability to separate luminance information from color properties. The network employs a two-branch structure, with the L branch handling luminance and the AB branch retaining color properties, allowing for targeted processing of shadow-related features, enhancing the model's capability to cleanse shadow-affected areas effectively.

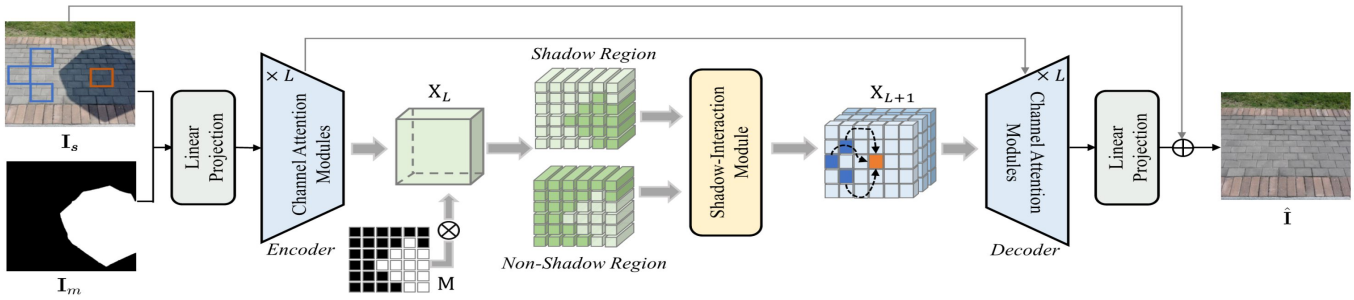

Fig. 1. Overview of the ShadowFormer network. The channel attention transformer-based encoder and decoder are to extract hierarchical information from the input shadow image and reconstruct the shadow-free image, respectively, using a series of channel attention (CA) modules. In the bottleneck stage, we adopt a Shadow-Interaction Module (SIM) to exploit the context information across both spatial and channel dimensions from non-shadow regions to help shadow region restoration.

C. *DeS3*

The DeS3 model presents a shadow removal method without requiring a binary mask, leveraging adaptive attention mechanisms and Vision Transformer (ViT) similarity loss during the sampling phase, rather than the recovering phase, this loss helps guide the reverse sampling toward recovering scene structure, It addresses hard, soft, and self shadows by effectively distinguishing between shadowed regions, objects casting shadows, and underlying structures. The model employs a novel diffusion-based approach for clearer boundary recovery. Each step in the diffusion process brings about changes in the adaptive attention to eliminate self and soft shadows that lack clear boundaries.

III. Methodology

A. *Overall Architecture*

Given a shadow input $\mathbf{I}_s \in \mathbb{R}^{3\times H\times W}$ with the corresponding shadow mask $\mathbf{I}_m \in \mathbb{R}^{H\times W}$, we first apply a linear projection LinearProj$(\cdot)$ to obtain the low-level feature embedding of the input, denoted by $\mathbf{X}_0 \in \mathbb{R}^{C\times H\times W}$, where C is the embedding dimension. Then, we feed the embedding $\mathbf{X}_0$ into the CA transformer-based encoder and decoder, each consisting of L CA modules to stack multi-scale global features. Each CA module consists of two CA blocks, as well as a down-sampling layer in the encoder or an up-sampling layer in the decoder. The CA block sequentially squeezes the spatial information via CA and captures the long-range correlation via a feed-forward MLP [**?**] as follows:

$$\tilde{\mathbf{X}} = \mathrm{CA}(\mathrm{LN}(\mathbf{X})) + \mathbf{X}, \tag{1}$$

$$\hat{\mathbf{X}} = \mathrm{GELU}(\mathrm{MLP}(\mathrm{LN}(\tilde{\mathbf{X}}))) + \tilde{\mathbf{X}}, \tag{2}$$

where LN$(\cdot)$ denotes layer normalization, GELU$(\cdot)$ denotes the GELU activation layer, and MLP$(\cdot)$ denotes a multi-layer perceptron. After passing through L modules within the encoder, we receive the hierarchical features $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_L\}$, where $\mathbf{X}_L \in \mathbb{R}^{2^L C\times \frac{H}{2^L}\times \frac{W}{2^L}}$. We calculate the global contextual correlation via the Shadow-Interaction Module (SIM) based on the pooled feature $\mathbf{X}_L$ in the bottleneck stage. Next, the features input to each CA module of the decoder are is the concatenation of the up-sampled features and the corresponding features from the encoder through skip connection

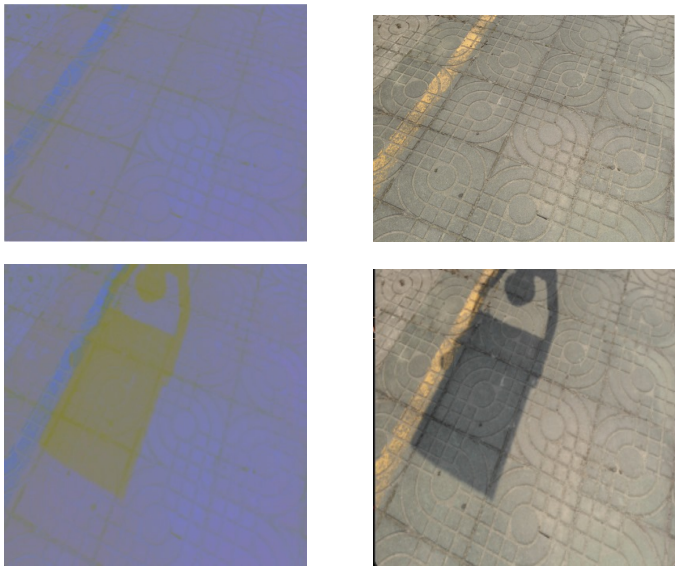

Fig. 2. L*A*B* Transformed images

B. *LAB Input Space*

Most image processing tasks consider images comprising of 3 channel components including Red, Green, and Blue(RGB). The value of each channel ranges from 0-255. Despite the wide adoption of current implementation on shadow removal, the shadow regions often severely deteriorate if represented in the RGB style. Consequently, another color spaced named LAB has aroused the interest of many researchers, where L represents perceptual luminance, A stands for the two unique colors of human vision: red and green while B corresponds to blue and yellow. Note that the value of the L channel changes from 0 to 100 while these of the A and B channels vary from -127 to 128. The LAB color space is a device-independent color system based on physiological characteristics and models the visual perception of human eyes.

Such a representation method stores the shadow-related luminance information in the L channel. Changes in the L

channel do not affect the color information since it is stored in the A and B channels, and vice versa. Therefore LAB color space has been demonstrated to be more suitable for shadow removal

C. *Blurred Binary Mask*

Shadow degradations in the real world exhibit considerable variation and can be classified into two main categories: soft shadows and hard shadows. This classification depends on the light source and the distance between the object and the surface, Soft shadows are characterized by blurred edges and gradual transitions from light to dark, creating penumbra areas, while hard shadows have sharp edges without penumbra. Models that use a binary shadow mask as a guidance for shadow removal, end up representing shadow locations with sharp edges end up to creating a border-like artifact visible on the edges where the shadows are removed from, this is especially true for Soft Shadows as they generally don't have sharp edges.

To tackle this problem, we pass the binary mask through a Gaussian filter to blur the edges of the white shadow mask, for the model to remove the shadow as naturally as possible. without creating any artifacts.

D. *Cosine-Distance*

In shadow removal, getting the color of the parts where the model has successfully removed the shadow as true to the ground truth is crucial. Choosing the right Lost Function can help in tackling this problem, assisting the model in covering all the aspects while retraining.

Originally ShadowFormer on which our model is primarily based uses Charbonnier Loss which is an evolved version of L2 loss, but if the loss is extremely large to be able to maintain the variability better loss switches to an L1 loss.

Considering we are changing the input space of the image to LAB from RGB before inserting it into the model, it is important to ensure that the relations between the colors are captured well between the input image and the ground truth especially when the color channels A and B are separate.

This is why we try to change the Loss Function to Cosine Distance, as it measures the cosine of the angle between two vectors. In a color context, it compares how similar two color vectors are in orientation, regardless of their magnitude, ensuring that the colors are compared based on hue and chromatic relation rather than absolute intensity.

Mathematically it can be defined as:-

$$\text{Cosine Distance Loss} = 1 - \cos(\theta) = 1 - \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$$

Here:

- $\mathbf{u}$ and $\mathbf{v}$ are color vectors (e.g., RGB or Lab representations).
- $\|\mathbf{u}\|$ and $\|\mathbf{v}\|$ are the magnitudes of the vectors.
- $\mathbf{u} \cdot \mathbf{v}$ is the dot product of the two vectors.

Benefits of using this loss function

- **Color Transfer**: Compare palettes in target and predicted images.
- **Image Colorization**: Ensure that the predicted color aligns with expected hues and tones.
- **Palette Matching**:Minimize differences between predicted and target color relations.

Earlier we tried to obtain the cosine distance within the image domain by flattening the image and applying the loss function to every channel of every pixel but the model failed to give anu results therefore as an alternative, we though of extracting mid and low level features of the image by passing it through a pretrained **VGG-16** available in pytorch which which is trained on Image-Net and the calculating the loss based on those feature vectors giving us far superior qualitative results.

E. *Deeper model*

The ShadowFomer model is known to be one of the efficient vision transformer models, being able to perform solely on 1 Nvidia A100 chip as well, though the transition from sRGB to LAB color space needs a deeper model, as the perceptually uniform and luminance-separated nature of LAB enables more sophisticated feature extraction that can be progressively leveraged through deeper architectural layers. By introducing additional encoder and decoder blocks, the model gains an enhanced ability to disentangle and reconstruct color information, particularly exploiting the LAB space's unique channel separation of luminance from chrominance, thereby improving color-aware feature learning and transformation capabilities.

IV. Experiments and Results

A. *Dataset*

The Image Shadow Triplets dataset (ISTD) is a dataset for shadow understanding that contains 1870 image triplets of shadow image, shadow mask, and shadow-free image. ISTD data set suffers from illumination inconsistencies where The shadow/no-shadow pairs in ISTD have different values outside the shadow area, and a lot of images suffer from having more than 2% of the pixels being saturated. This can distort color perception and affect the model that usually requires a balanced intensity distribution.

The first task was to fix the Dataset by using the pixels outside the shadow mask and computing the difference between Shadow and Shadow Free image in that area. Fitting a plane in between the shadow and shadow-free image either as a combination of 3 channels or doing it per channel. Projecting the Shadow Free Image onto that to best match the Shadow Image, using it as Ground Truth once Fixed removing all the images having more than 2% saturated pixels.

Following this all the images were converted to LAB color space we tried to use the **OpenCV** and **SciKit** functions convert the images but these functions do not represent the exact values mathematically therefore we created our own functions and hardcoded the exact values to perform the transformations.

Fig. 3. (a) Shadow Image, (b) Original Shadow Mask, and (c) Smoothened Shadow Mask using Gaussian blur

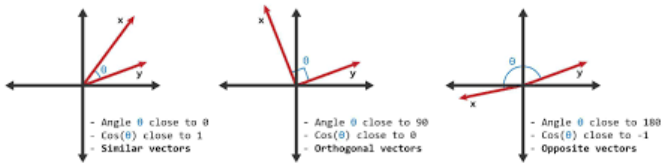

Fig. 4. Cosine Similarity

Below is an example:-
First, the sRGB images are converted to XYZ

$$x = \frac{(r \cdot 0.4124) + (g \cdot 0.3576) + (b \cdot 0.1805)}{0.95047}$$

$$y = \frac{(r \cdot 0.2126) + (g \cdot 0.7152) + (b \cdot 0.0722)}{1.00000}$$

$$z = \frac{(r \cdot 0.0193) + (g \cdot 0.1192) + (b \cdot 0.9505)}{1.08883}$$

Then the XYZ images are converted to LAB

$$x = \begin{cases} x^{\frac{1}{3}}, & \text{if } x > 0.008856, \\ 7.787 \cdot x + \frac{16}{116}, & \text{otherwise.} \end{cases}$$

$$y = \begin{cases} y^{\frac{1}{3}}, & \text{if } y > 0.008856, \\ 7.787 \cdot y + \frac{16}{116}, & \text{otherwise.} \end{cases}$$

$$z = \begin{cases} z^{\frac{1}{3}}, & \text{if } z > 0.008856, \\ 7.787 \cdot z + \frac{16}{116}, & \text{otherwise.} \end{cases}$$

$$L = (116 \cdot y) - 16$$

$$A = 500 \cdot (x - y)$$

$$B = 200 \cdot (y - z)$$

B. Results

We trained the model for 100 epochs for each modification, balancing GPU restrictions with the need to achieve reasonably good performance while maintaining a consistent basis for comparison. This approach allowed us to evaluate the impact of various architectural changes and input transformations systematically, ensuring that the results were both meaningful and comparable across different configurations.

TABLE I
THE QUANTITATIVE RESULTS OF SHADOW REMOVAL USING VARIOUS ITERATIONS OF THE MODEL

Model	*PSNR*	*SSIM*
Original ShadowFormer	30.421639	0.965128
Input as L*A*B*	30.555857	0.981886
L*A*B* with cosine sim	25.431281	0.949069
- with gradient clipping	28.973819	0.953715
- Charbonnier + cosine sim	30.358224	0.983740
Deeper Model	**31.777911**	**0.985228**

C. Modification 1- LAB Color Space

The experiments in our study build upon the baseline ShadowFormer model by applying various modifications to enhance its performance. Starting with the original model, which achieved a performance score of 30.42 and a cosine similarity of 0.965, the first adjustment involved using the LAB* color space for input data, resulting in a slight improvement in both metrics (30.56 and 0.982).

D. Modification 2 - Cosine Distance

Next, the introduction of cosine similarity as an optimization criterion led to a drop in performance (25.43) and cosine similarity (0.949), as we encountered with gradient explosion, and our loss reached infinity after 40 epochs. However, applying gradient clipping improved the model's stability and raised the performance to 28.97 with a slight increase in cosine similarity (0.954). Combining cosine similarity with Charbonnier loss further improved the model, bringing the performance score back up to 30.36 and increasing cosine similarity to 0.984. Finally, the introduction of a deeper model resulted in the highest performance (31.78) and cosine similarity (0.985), highlighting that deeper architectures are more effective in learning complex features of the L*A*B* color space and producing higher-quality outputs.

E. Modification 3 - Combining the Losses

Considering the output obtained through cosine distance loss was either significantly inferior during both cases of it being used, our next experiment was to try and combine the loss of the orginal ShadowFormer model being the Charbonnier with the Cosine Distance loss and see if the model performs

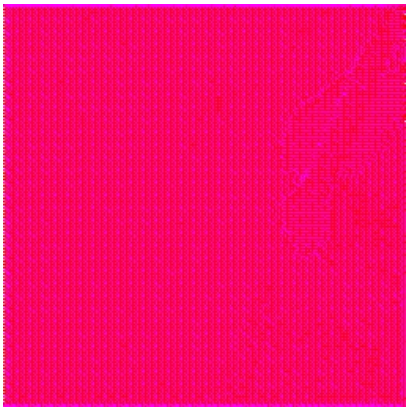

Fig. 5. Model output - cosine distance failure while using the function directly on images(output and ground truth comparison)

any better, though not sigificantly, assuming that combining the loss with take care of the color accuracy as well as visual shape similarity and we can see the output being qualitatively and quantatively far better than the ones in which only Cosine Distance is used as a loss, though the output is close to the orginal output os ShadowFormer.

F. Modification 4 - Deepening the Architecture

Now that we have tried experimenting the original model by changing the input space as well as the loss function, we further went ahead to add additional layers of the encoder and decoder blocks to see how significant would the improvements in results be.

Though ShadowFormer as a model is known to be an Vision Transformer but adding additonal layers did not significantly impact the performance where earlier it took close **170 seconds** to run an epoch. deeping the model increase that time to on average 220 seconds per epoch to train. The results did see a substantial improvement where for the first time a PSNR score crossed the **31** threshold reaching a highest score of **31.77** within 100 epochs, qualitatively also the results speaks for itself.

V. Failures

- **Color Accuracy**: In Images involving a lot of colors there were a few inconsistencies in matching the color profile especially as we concurrently do RGB-LAB-RGB withing training and testing the model, this can be seen in **Figure 6**
- **Cosine Distance**: While initially trying to calculate the loss in image space the results were undesirable.

VI. Future Work

- **Projection**: Instead of Linear Projecting the input images and passing them obtained embeddings to the ViT encoder and decoder blocks, the images can be passed through the extraction layers of **ResNet50** or **ConVNext** in order to get better embeddings for the transformers to work with.
- **Batch Normalization**: Adding Batch Normalization between encoder and decoder blocks inorder to obtain better color accuracy especially for images having different colors .
- **Parallel Processing**: Seperating the L channel from the AB channel and parallel processing them together, finally concatenating them before passing them to the ViT. Seperating the Luminosity ad color channel might help us get better features individually.

VII. Conclusion

Shadow Removal is a persistent challenge in computer vision and remains a complex problem to solve effectively. Developing a robust function to address this issue will enable us to carry out a series of experiments designed to improve its performance.

Transforming images to the LAB color space provides a valuable approach, but it is crucial to adapt the model to process these images appropriately. The features extracted in the LAB space differ significantly from those in the standard sRGB space, necessitating careful tweaking of the architecture.

While Cosine Distance is a widely used loss function in shadow removal tasks, it does not consistently perform well across all contexts. However, there is potential in integrating LAB-based input architectures with the Cosine Distance loss function to achieve improved results.

Techniques such as gradient clipping and advanced feature extraction architectures can significantly enhance the performance of deep neural networks in this domain.

It is equally important to evaluate the trade-offs between computational cost and performance when deepening the architecture to ensure an optimal balance.

Acknowledgment

We would want to thank Professor Bruce Maxwell, Assistant Director of Computer Programs, Northeastern Seattle for taking the CS7180 class and giving us an opportunity to work on this.

References

[1] L. Guo, S. Huang, D. Liu, H. Cheng, and B. Wen, "ShadowFormer: Global Context Helps Image Shadow Removal," *arXiv preprint arXiv:2302.12345*, 2023. [Online]. Available: https://arxiv.org/abs/2302.12345.

[2] X. Wang, L. Guo, X. Wang, S. Huang, and B. Wen, "SoftShadow: Leveraging Penumbra-Aware Soft Masks for Shadow Removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[3] J. Xiao, X. Fu, Y. Zhu, D. Li, J. Huang, K. Zhu, and Z. Zha, "HomoFormer: Homogenized Transformer for Image Shadow Removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 25617–25626.

[4] H. Yang, G. Nan, M. Lin, F. Chao, Y. Shen, K. Li, and R. Ji, "LAB-Net: LAB Color-Space Oriented Lightweight Network for Shadow Removal," *arXiv preprint arXiv:2208.13182*, Aug. 2022, last revised: 4 Sep. 2022 (v2). [Online]. Available: https://arxiv.org/abs/2208.13182.

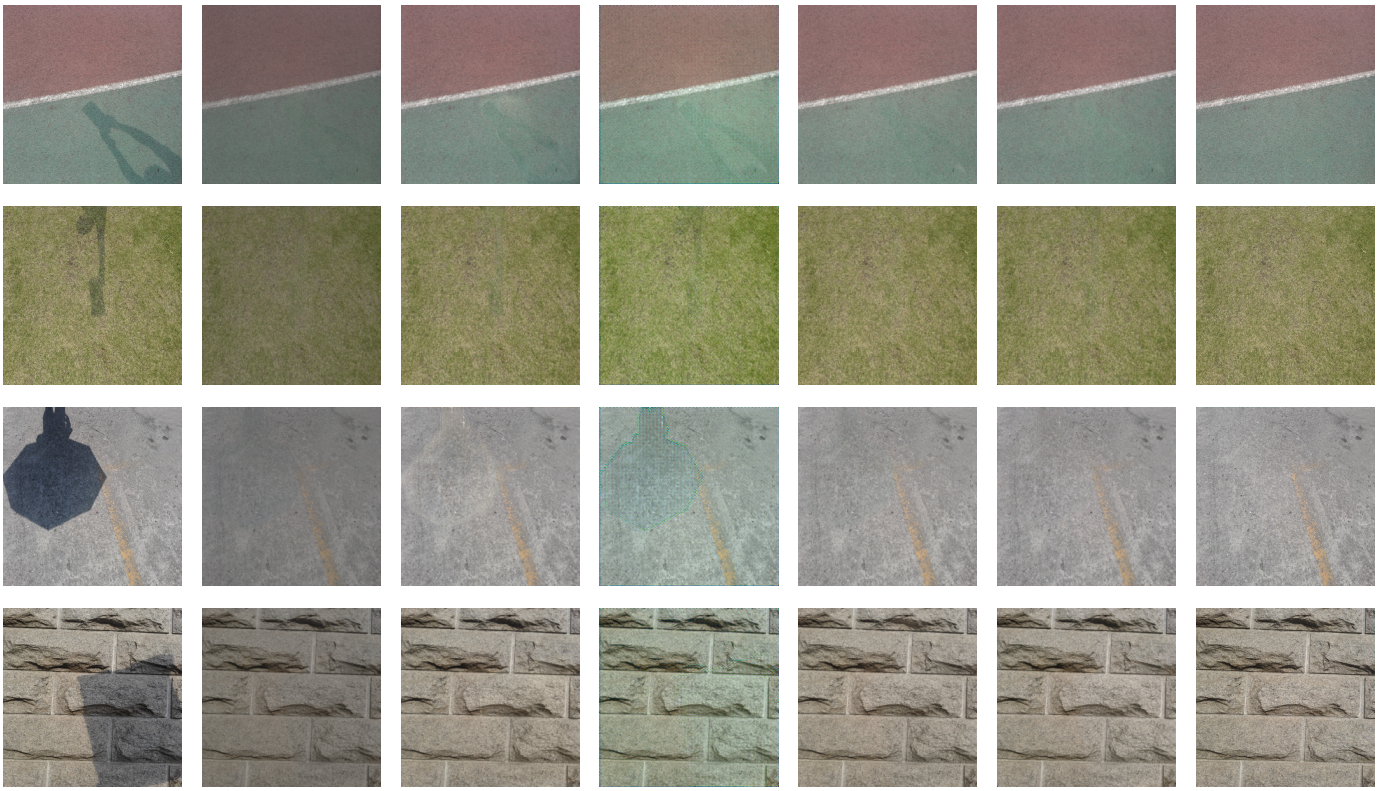

Fig. 6. (a) Shadow Image (b) Original ShadowFormer, (c) LAB, (d) Cosine, (e) Combined, (f) Final (g) Ground Truth

Fig. 7. Failure case - (a) Shadow Image, (b) ground truth, and (c) model output