Asymmetric Color Transfer with Consistent Modality Learning

Kaiwen Zheng* Jie Huang* Man Zhou Feng Zhao†
University of Science and Technology of China
{kezh,hj0117,manman}@mail.ustc.edu.cn, fzhao956@ustc.edu.cn

Abstract

The mono-color dual-lens system widely exists in the smartphone that captures asymmetric stereo image pairs, including high-resolution (HR) monochrome images and low-resolution (LR) color images. Asymmetric color transfer aims to reconstruct an HR color image by transferring the color information of the LR color image to the HR monochrome image. However, the inconsistency of spectral resolution and spatial resolution between stereo image pairs poses a challenge for establishing reliable stereo correspondence for precise color transfer. Previous works have not adequately addressed this issue. In this paper, we propose a dual-modality consistency learning framework to assist the establishment of reliable stereo correspondence. According to the complementarity of color and frequency information between stereo images, a dual-branch Stereo Information Complementary Module (SICM) is devised to perform the consistent modality learning in feature domain. Specifically, we meticulously design the stereo frequency and color modulation mechanism equipped in the SICM for capturing the information complementarity between dual-modal features. Furthermore, a parallax attention distillation is proposed to drive consistent modality learning for better stereo matching. Extensive experiments demonstrate that our model outperforms the state-of-the-art methods in the Flickr1024 dataset and has superior generalization ability over the KITTI dataset and real-world scenarios. The code is available at https://github.com/keviner1/SICNet.

1. Introduction

To balance the hardware cost and imaging quality, multi-sensor joint imaging has become the mainstream configuration of smart phones. Mono-color dual-lens system is a classic joint imaging system, which is used in many smart phones, *e.g.*, HONOR Magic3 and HUAWEI P50. It consists of two kinds of sensors: the color one and the monochrome one. Specifically, the former with a Bayer array color filter in front of the image sensor is responsible for separating the incident light into one of three primary colors to capture the color information. However, the Bayer array leads to the blocked incident lights and amplified image noise. Unlike color cameras, the latter directly receives all the incident lights at each pixel without the process of filtering and demosaicing. Though it lacks color properties, it has better light efficiency and provides clearer images than Bayer-filtered color cameras.

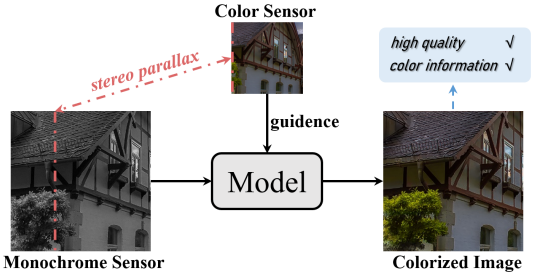


Figure 1. The flowchart of asymmetric color transfer in the mono-color system. It aims to employ one view's low-resolution color image as the reference to guide another view's high-resolution monochrome image for coloring. The big challenge in this task is to establish a reliable correspondence between stereo images with asymmetric spectral and spatial resolutions accompanied by various parallaxes.

Then there are two schemes that can combine the benefits of the mono-color image pairs: stereo image super-resolution and stereo image color transfer. Compared with the super-resolution method, color transfer can make full use of high-quality mono images to ensure visual effect. In the case of different parallax scenarios, transferring the high-frequency information faces more challenges and is prone to copy artifacts, which is more unacceptable than color bias. Therefore, many researchers have focused on the stereo color transfer (the workflow is illustrated in Fig. 1). With regards to the model design, the common pipeline has three steps: first, extracting the feature pairs of stereo images; second, establishing the correspondence between stereo feature pairs; finally, warping the color information of the color image for monochrome image colorization by

*Both authors contributed equally to this research.
†Corresponding author.

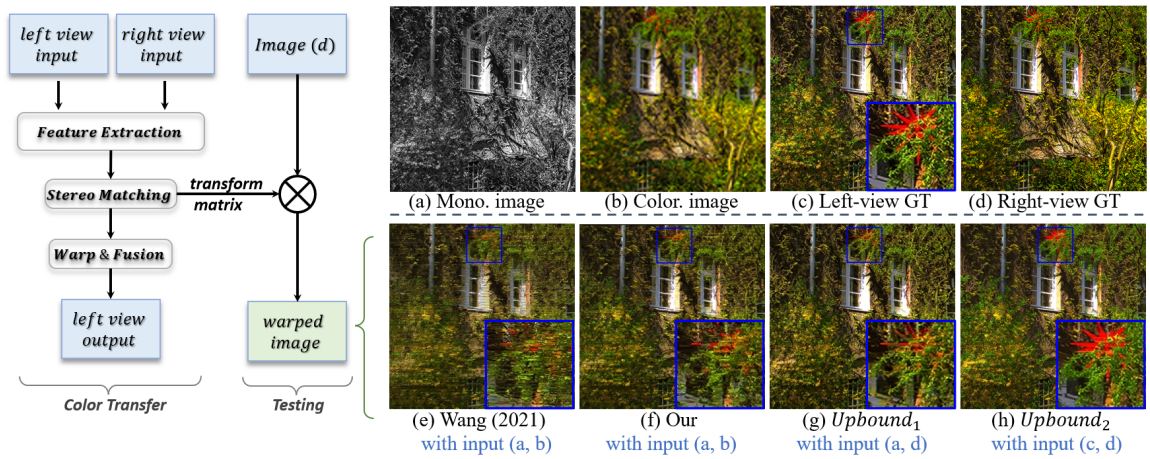


Figure 2. An experimental example for explaining that the higher the modal consistency, the more accurate the stereo matching. We present the $Upbound$ models to exploit the benefit of improved modal consistency in stereo matching. The shared network architecture of the $Upbound$ models is illustrated in Fig. 7. For better visualization, the color image (b) is given in the upsampled format.

the established stereo correspondence. Obviously, establishing an accurate stereo correspondence is vitally essential for elevating the coloring effect. However, the stereo images with asymmetric spectral and spatial resolution result in the high inconsistency of stereo images in modality, which increases the difficulty of accurate stereo matching.

Previous color transfer works [5, 6] under spatial-symmetric attempts to address the asymmetric spectral resolution by extracting the gray map of color images to perform stereo matching with monochrome image. Although these methods achieve consistency in spectral resolution, they ignore that color image-based matching outperforms gray image-based matching [1]. Different from the above, the asymmetric color transfer task suffers from higher inconsistency in both spectral and spatial resolutions, hampering accurate correspondence establishment. However, the SOTA method [14] does not fully consider this issue and treats dual-modal images equally during processing.

The necessity of modal consistency in establishing reliable stereo correspondence is illustrated in Fig. 2. In experiments, we directly warp the right-view image (d) to the left-view one using the stereo transform matrix established by various models. As demonstrated by the results (g, h) of constructed $Upbound_1$ and $Upbound_2$ model, stereo matching becomes more accurate as the image pair's modal consistency grows. In addition, benefiting from consistent modality learning, our warping result (f) is superior to the result (e) generated by the previous SOTA model [14].

In this paper, we propose a novel framework to perform consistent modality learning that assists the establishment of reliable stereo correspondence for better color transfer. Specifically, we fully consider the complementarity of color and frequency information between stereo images with asymmetric spectral and spatial resolutions. Thus, we design a dual-branch Stereo Information Complementary Module (SICM) to carry out consistent modality learning, as depicted in Fig. 3. It contains a meticulously designed Stereo Color Modulation (SCM) block and a Stereo Frequency Modulation (SFM) block for information complementing between dual-modal image feature pairs. In addition, to further increase the accuracy of color transfer by directly guiding the established stereo correspondence, we employ the $Upbound_2$ model learned in the consistent modality scenario for parallax attention distillation. Extensive experiments are conducted on the Flickr1024 dataset and demonstrate the superior performance of our model.

The main contributions of this work are summarized as:

- We propose a consistency learning framework for asymmetric color transfer. Compared with the previous methods, the proposed method achieves the best quantitative and qualitative results, and extensive experiments have proved its excellent generalization performance.
- We design a dual-branch Stereo Information Complementary Module (SICM) to conduct consistent modality learning by information complementation. Inside SICM, the stereo modulation blocks are devised to modulate the color and frequency information to complement the information between the dual branches.
- We introduce a parallax attention distillation strategy to further boost our model to establish more reliable stereo correspondence for elevating the coloring effect.

2. RELATED WORK

The mono-color dual-lens colorization can be regarded as special reference-based colorization under the camera system, but the requirements for accuracy are more stringent. Jeon et al. [8] studied the stereo matching under low

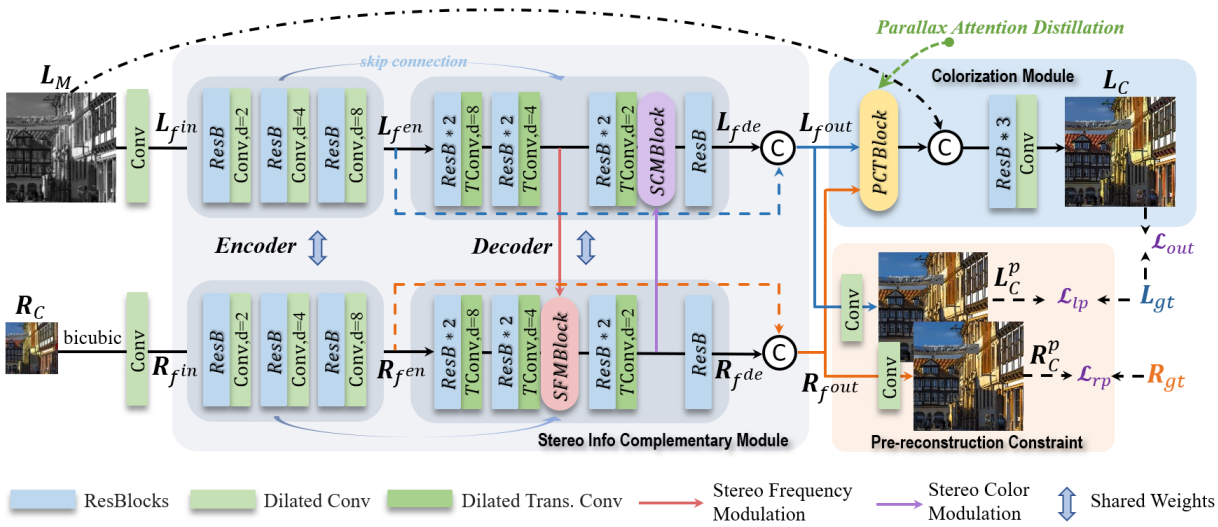


Figure 3. An overview of our proposed framework. Following the complementarity of color and frequency information presented between the input image pair ($\mathbf{L}_M$, $\mathbf{R}_C$), we design a Stereo Information Complementary Module (SICM) to achieve consistent modality learning. Specifically, the SFM block and SCM block are developed to realize the information complementation under the guidance of the Pre-reconstruction Constraint. In the Colorization Module, we present a PCT block to establish reliable stereo correspondence under the supervision of parallax attention distillation for accurate color transfer.

light conditions and realized the symmetric color transfer based on the estimated disparity map. Trinidad et al. [11] designed PixelFusionNet to solve the problem of multi-view image fusion, including color transfer. The two-stage CNN proposed by Dong et al. [5] focuses on more valuable pixels through the attention mechanism and introduces the 3D-Regulation operation to consider more context information. Based on this network, Dong et al. [6] introduced the Cycle-CNN structure with well-designed cycle consistency loss and structure similarity loss to realize the self-supervised colorization. Besides, Dong et al. [4] also designed a coloring model based on the pyramid CNN architecture, which introduces Markov random field (MRF) to update warp information. Recently, Wang et al. [14] raised a more challenging asymmetric color transfer problem and achieved excellent results. However, they do not consider the high modal inconsistency of input stereo images, which hinders reliable correspondence establishment and limits the model's performance.

3. METHOD

In asymmetric color transfer, the input image pair composed of high-resolution left-view monochrome image $\mathbf{L}_M \in \mathbb{R}^{H\times W\times 1}$ and low-resolution right-view color image $\mathbf{R}_C \in \mathbb{R}^{(H/s)\times(W/s)\times 3}$, where H, W, and s denote the image height, image width, and scaling factor, respectively. The asymmetric spectral and spatial resolution between $\mathbf{L}_M$ and $\mathbf{R}_C$ leads to high modality inconsistency, which increases the difficulty of stereo matching. However, introducing pre-trained restoration networks to solve such problem will bring more computational consumption.

Therefore, we propose a framework that achieves consistent modality learning in the feature extraction stage. Specifically, the framework includes a Stereo Information Complementary Module (SICM) and a colorization module. The SICM is cooperation with the pre-reconstruction constraint for consistent modality learning. The colorization module is introduced to utilize the features learned by SICM for accurate color transfer. Firstly, we perform bicubic interpolation on image $\mathbf{R}_C$ to achieve the same size as $\mathbf{L}_M$. Then fed into a convolution layer to obtain original features $\mathbf{L}_{f^{in}} \in \mathbb{R}^{H\times W\times C}$ and $\mathbf{R}_{f^{in}} \in \mathbb{R}^{H\times W\times C}$, where C is the number of channels. Depending on the complementary characteristics between the clear monochrome image and blurred color image, the SICM is presented to perform information complementary between dual branches. It employs two modulation blocks: a Stereo Frequency Modulation (SFM) Block for color branch enhancing frequency information and a Stereo Color Modulation (SCM) Block for monochrome branch supplementing color information. To guide the modal consistency learning in SICM, we pre-reconstruct the modulated features into stereo image pair ($\mathbf{L}_C^p$, $\mathbf{R}_C^p$) and supervise them with ground truth stereo image pair ($\mathbf{L}_{gt}$, $\mathbf{R}_{gt}$).

In the colorization module, we develop a Parallax Color Transfer (PCT) block to implement color transfer in feature domain. The PCT utilizes multi-level features provided by SICM to establish accurate stereo correspondence. Moreover, we design an Upbound model for parallax at-

tention distillation, which establishes precise stereo correspondences by inputting image pairs with entirely consistent modalities. Finally, to further improve the quality of colorization, residual blocks are employed to refine the features for the final result $\mathbf{L}_C$ with the guidance of $\mathbf{L}_M$.

3.1. Stereo Information Complementary Module

As shown in Fig. 3, The SICM is an encoder-decoder architecture consisting of monochrome and color branches. Given the original feature pair ($\mathbf{L}_{f^{in}}$, $\mathbf{R}_{f^{in}}$) as input, the module extracts the encoder feature pair ($\mathbf{L}_{f^{en}} \in \mathbb{R}^{H\times W\times C}$, $\mathbf{R}_{f^{en}} \in \mathbb{R}^{H\times W\times C}$) with a larger receptive field and the decoder feature pair ($\mathbf{L}_{f^{de}} \in \mathbb{R}^{H\times W\times C}$, $\mathbf{R}_{f^{de}} \in \mathbb{R}^{H\times W\times C}$) with pre-aligned modal information. In the end, concatenating them to obtain discriminative feature pair($\mathbf{L}_{f^{out}} \in \mathbb{R}^{H\times W\times 2C}$, $\mathbf{R}_{f^{out}} \in \mathbb{R}^{H\times W\times 2C}$) as output.

Since the rich contextual information is beneficial for establishing accurate stereo correspondence [2], we obtain the encoder's feature pair ($\mathbf{L}_{f^{en}}$, $\mathbf{R}_{f^{en}}$) by stacking ResBlocks and dilated convolutions (with dilation rates of 2, 4, 8) to enlarge the receptive field in a dense sampling way. In decoder, we perform modal consistency learning based on the information complementarity between two branch features. Specifically, the SFM block is developed to modulate frequency information of the color branch with the monochrome branch as the reference. And the SCM block takes the color branch as the reference to supplement the color information for the monochrome branch. To this end, the information of decoder's output feature pair ($\mathbf{L}_{f^{de}}$, $\mathbf{R}_{f^{de}}$) has better modal consistency. Notably, we apply skip connections in the corresponding layer between encoder and decoder in order to minimize the information loss.

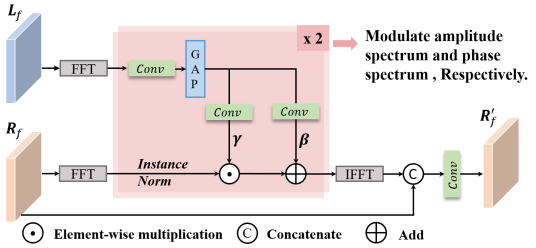


Figure 4. The architecture of SFM Block. It performs alignment-free modulation, i.e., AdaIN, between dual-branch features in the Fourier domain for modal complementary learning.

Stereo Frequency Modulation (SFM) Block. Inspired by previous work [17] on fusing multi-frequency features via normalization methods, the SFM block performs stereo modulation in the frequency domain by means of Adaptive Instance normalization [7]. As shown in Fig. 4, we first transform the current layer input feature pairs ($\mathbf{L}_f \in \mathbb{R}^{H\times W\times C}$, $\mathbf{R}_f \in \mathbb{R}^{H\times W\times C}$) into the frequency domain through Fast Fourier Transform (FFT), and then modulate the amplitude and phase spectrum through learning-based AdaIN, respectively. After modulation, the amplitude spectrum and phase spectrum are transformed back to the spatial domain through Inverse Fast Fourier Transform (IFFT). Finally, the transformed features are concatenated with the original feature and then fused through a 3×3 convolution layer to obtain the output $\mathbf{R}'_f \in \mathbb{R}^{H\times W\times C}$.

When performing the normalization operation to modulate frequency information, we aggregate the global information of the monochrome branch features through 1×1 convolution and Global Average Pooling. Similar to previous learning-based normalization [9, 13], the affine parameters γ and β are learned by 1×1 convolution to scale and shift the normalized color branch features. Our learning-based AdaIN, when modulates x with y as the reference, can be formulated as follows:

$$\begin{aligned} \boldsymbol{AdaIN}(x,y) &= \gamma^y \boldsymbol{IN}(x) + \beta^y \\ &= \gamma^y\left(\frac{x - \mu(x)}{\sigma(x)}\right) + \beta^y \end{aligned} \tag{1}$$

where $\mu(x)$ and $\sigma(x)$ are the channel-wise mean and variance of x, γ^y and β^y are the affine maps learning from y.

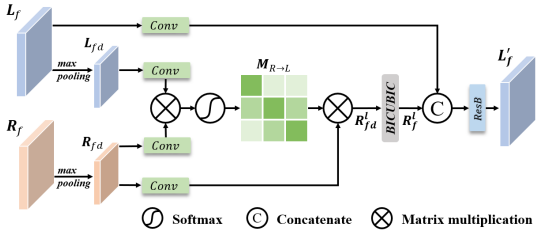


Figure 5. The architecture of SCM Block. To balance performance and complexity, the alignment operation is performed in downsampled resolution.

Stereo Color Modulation (SCM) Block. Given the current layer dual branch feature pair ($\mathbf{L}_f \in \mathbb{R}^{H\times W\times C}$, $\mathbf{R}_f \in \mathbb{R}^{H\times W\times C}$) as input, the SCM block conducts coarse colorization for feature $\mathbf{L}_f$ with feature $\mathbf{R}_f$ as the reference. Since the color information is globally continuous, the degradation of the downsampling operation has less effect on the color information. Therefore, we downsample the feature pair ($\mathbf{L}_f$, $\mathbf{R}_f$) to ($\mathbf{L}_{fd} \in \mathbb{R}^{(H/2)\times(W/2)\times C}$, $\mathbf{R}_{fd} \in \mathbb{R}^{(H/2)\times(W/2)\times C}$) for balancing the coloring effect with computing efficiency. After then, the parallax attention mechanism [12] is employed to establish stereo correspondence between the feature pair ($\mathbf{L}_{fd}$, $\mathbf{R}_{fd}$) and warp the right-view feature $\mathbf{R}_{fd}$ into the left-view feature $\mathbf{R}^l_{fd}$. Finally, we upsample the warped feature $\mathbf{R}^l_{fd}$ into $\mathbf{R}^l_f$ and fuse it with $\mathbf{L}_f$ to achieve color information supplementation. The architecture of the SCM block is shown in Fig. 5. It should be emphasized that the dual-branch feature here is not pre-aligned but only color information modulated.

Pre-reconstruction Constraint. To better drive stereo

modulation blocks for accurate information complementary, we reconstruct the feature pair ($\mathbf{L}_{fde}$, $\mathbf{R}_{fde}$) into image pair ($\mathbf{L}_C^p$, $\mathbf{R}_C^p$), as shown in Fig. 3. Consequently, under the supervision of the ground truth image pair ($\mathbf{L}_{gt}$, $\mathbf{R}_{gt}$), consistent modality learning is more targeted to narrow the gap of modal information between stereo images.

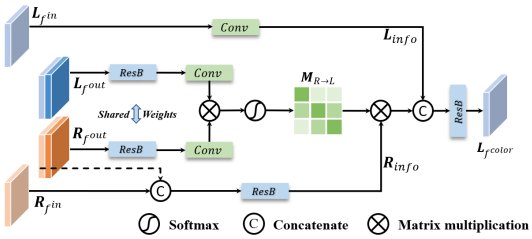


Figure 6. The architecture of PCT Block. It combines multi-level features extracted from SICM for better color transfer.

3.2. Parallax Color Transfer (PCT) Block

In the colorization module, we implement a Parallax Color Transfer (PCT) block to take advantage of multi-level features for accurate stereo matching and color transfer, as shown in Fig. 6. Based on the prior that the stereo image only has the disparity in the horizontal direction, the PCT block adopts the parallax attention mechanism [12] for establishing the stereo correspondence.

For the monochrome branch, feature pair ($\mathbf{L}_{fout}$, $\mathbf{R}_{fout}$) for stereo matching integrates ($\mathbf{L}_{fen}$, $\mathbf{R}_{fen}$) with rich context information and ($\mathbf{L}_{fde}$, $\mathbf{R}_{fde}$) with the modulated modal information. To prevent mutual interference between the features used for reconstruction and the features used for binocular matching, we introduced the original reconstruction features. To prevent the mutual interference between reconstruction and stereo matching during model optimization, we added the original features $\mathbf{L}_{fin}$ for reconstruction. For the color branch, due to the consistent modality learning in SICM, the feature $\mathbf{R}_{fde}$ has much better quality and less noise than the original feature $\mathbf{R}_{fin}$. Therefore, we fuse the $\mathbf{R}_{fde}$ and $\mathbf{R}_{fin}$ before warping the color information.

The complete coloring process of PCT block is as follows: Firstly, the feature pairs ($\mathbf{L}_{fout}$, $\mathbf{R}_{fout}$) are fed into a shared-weight residual block and a convolution layer to obtain ($\hat{\mathbf{L}}_{fout}$, $\hat{\mathbf{R}}_{fout}$) for stereo matching. After matrix multiplication between $\hat{\mathbf{L}}_{fout} \in \mathbb{R}^{H\times W\times 2C}$ and transposed $\hat{\mathbf{R}}^T_{fout} \in \mathbb{R}^{H\times 2C\times W}$, the parallax attention map $\mathbf{M}_{R\to L} \in \mathbb{R}^{H\times W\times W}$ is calculated after the softmax function:

$$\mathbf{M}_{R\to L} = softmax(\hat{\mathbf{L}}_{fout} \otimes \hat{\mathbf{R}}^T_{fout}) \tag{2}$$

Secondly, we fuse the feature $\mathbf{R}_{fin}$ and $\mathbf{R}_{fde}$ by a residual block to obtain $\mathbf{R}_{info} \in \mathbb{R}^{H\times W\times C}$. And the feature $\mathbf{L}_{fin}$ after convolution is recorded as $\mathbf{L}_{info} \in \mathbb{R}^{H\times W\times C}$. Thirdly, warping $\mathbf{R}_{info}$ to the left-view $\mathbf{R}^l_{info}$ by multiplying the matrix of $\mathbf{M}_{R\to L}$ and $\mathbf{R}_{info}$. Finally, residual block is adopted to fuse concatenated $\mathbf{L}_{info}$ and $\mathbf{R}^l_{info}$ to get colorization result $\mathbf{L}_{fcolor} \in \mathbb{R}^{H\times W\times C}$:

$$\begin{aligned}\mathbf{L}_{fcolor} &= f\left(\left[\mathbf{L}_{info}, \mathbf{R}^l_{info}\right]\right) \\ &= f\left(\left[\mathbf{L}_{info}, \mathbf{R}_{info} \otimes \mathbf{M}_{R\to L}\right]\right)\end{aligned} \tag{3}$$

where $f(\cdot)$ stands for the ResBlock, $[\cdot]$ means concatenation operation.

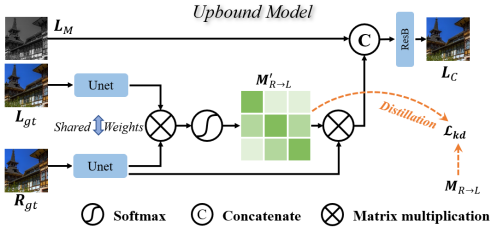


Figure 7. The architecture of $Upbound_2$. The $Upbound_2$ model takes a high-quality image pair as input and establishes accurate stereo correspondence for knowledge distillation.

3.3. Parallax Attention Distillation

As illustrated in Fig. 2, since the modalities of the input image pair (c, d) are completely consistent, the stereo correspondence established by the $Upbound_2$ model is the most precise, which is reflected in the fact that the result (h) has the clearest details and colors. To this end, we employ the $Upbound_2$ model to further drive our model for establishing more reliable stereo correspondence through a proposed parallax attention distillation strategy. As shown in Fig. 7, the $Upbound_2$ network also adopts the parallax attention mechanism as same as the PCT block to establish stereo correspondence $\mathbf{M}'_{R\to L}$. Therefore, we can guide the established stereo correspondence $\mathbf{M}_{R\to L}$ of the PCT block by means of knowledge distillation loss:

$$\mathcal{L}_{kd} = \|\mathbf{M}_{R\to L} - \mathbf{M}'_{R\to L}\|_1, \tag{4}$$

3.4. Loss Function

Our loss function consists of two parts: the output loss $\mathcal{L}_{out}$ for more accurate coloring and the modal consistency loss $\mathcal{L}_{mc}$ for narrowing the modal information gap of stereo images:

$$\mathcal{L} = \mathcal{L}_{out} + \lambda_1 \mathcal{L}_{mc}, \tag{5}$$

where λ is a weighting factor.

The output loss $\mathcal{L}_{out}$ is obtained by calculating the mean absolute error between the left-view ground truth image $\mathbf{L}_{gt}$ and final reconstructed left-view color image $\mathbf{L}_C$:

$$\mathcal{L}_{out} = \|\mathbf{L}_C - \mathbf{L}_{gt}\|_1 \tag{6}$$

In order to restrict the consistent modality learning in SICM and make it oriented to accurate stereo matching, we

Table 1. Quantitative comparison results on Flickr1024 testset. As can be seen, our methods attained the best performance over all noise setups while requiring fewer parameters than previous state-of-the-art techniques. We also present the optimal generalization ability in the KITTI dataset; see details in Supplementary Material.

Model	Setup1		Setup2		Setup3		#Param (M)
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	
Welsh (2002) [16]	24.53	0.806	23.88	0.758	23.86	0.751	-
Zhang (2016) [18]	22.26	0.885	21.99	0.852	21.99	0.852	32.60
Jeon (2016) [8]	28.47	0.922	28.09	0.901	27.32	0.887	-
Dong (2019) [5]	29.48	0.925	28.71	0.893	28.59	0.890	2.30
Dong (2020) [6]	19.86	0.805	17.85	0.723	17.72	0.722	0.12
Su (2020) [10]	20.25	0.725	20.13	0.713	20.13	0.713	34.18
Chen (2021) [3]	27.58	0.919	27.17	0.899	27.11	0.894	2.99
Wang (2021) [14]	30.50	0.941	29.94	0.927	29.72	0.925	1.35
Ours	**31.10**	**0.948**	**30.40**	**0.932**	**30.31**	**0.932**	1.09

elaborately design the pre-reconstruction loss and knowledge distillation loss. The modal consistency loss can be formulated as:

$$\begin{aligned}\mathcal{L}_{mc} &= \mathcal{L}_{lp} + \mathcal{L}_{rp} + \lambda_2 \mathcal{L}_{kd} \\ &= \|\mathbf{L}_C^p - \mathbf{L}_{gt}\|_1 + \|\mathbf{R}_C^p - \mathbf{R}_{gt}\|_1 + \lambda_2 \mathcal{L}_{kd},\end{aligned} \tag{7}$$

where $\mathcal{L}_{lp}$ and $\mathcal{L}_{rp}$ correspond to the constraints of color modulation for left-view pre-reconstructed image $\mathbf{L}_C^p$ and frequency modulation for right-view pre-reconstructed image $\mathbf{R}_C^p$, respectively. $\mathbf{L}_{gt}$ and $\mathbf{R}_{gt}$ are the ground truth stereo images without decolor and degradation.

4. EXPERIMENTS

4.1. Pre-reconstruction Results

In this section, we visualize the pre-reconstructed images within the pre-reconstruction constraint. As depicted in Fig. 8, compared with the input image pair (a, b), the reconstructed image pair (c, d) possess higher modal consistency, which is conducive to the establishment of more accurate stereo correspondence. It illustrates the effectiveness of the consistent modality learning performed by SICM through the stereo information modulation.


(a) Mono. image (b) Color image (c) Reconstructed (a) (d) Reconstructed (b)

Figure 8. An example of modal consistency enhancing. The pre-reconstructed (c, d) exhibits better modal consistency.

4.2. Datasets

We use the stereo image dataset Flikr1024 [15] for training and testing. The dataset contains 1024 pairs of high-quality color images with different parallax. In order to simulate the real mono-color system, we regard the decolored left-view image as collected by the monochrome sensor and the right-view image scaled by 1/4 factor as collected by the color sensor. On this basis, we add signal-dependent Gaussian noises with different given standard deviations (STDs) to simulate the light-efficiency difference between the color and mono sensors, as shown in Table 2.

Table 2. Noise setups (k denotes noise-free signal intensity).

Noise STD.	Color Image	Mono. Image
Setup1	0	0
Setup2	$0.03\sqrt{k}$	$0.01\sqrt{k}$
Setup3	$0.07\sqrt{k}$	$0.01\sqrt{k}$

4.3. Implementation Details

We divided the Flickr1024 dataset into a training set containing 912 image pairs and a test set containing 112 image pairs. For both the training set and test set, we crop the image pair and resize the color image with a scaling factor of 4; then, we will obtain 160 $\times$ 480 left-view monochrome images and 40 $\times$ 120 right-view color image pairs. Finally, we have 5662 training image pairs and 636 test image pairs in the experiment. The proposed network is implemented with the PyTorch framework. We trained the model on an NVIDIA GTX 2080ti GPU. In training, we set the batch size to 1, the number of channels to 48, and the initial learning rate to 1.0×10^{-3}. We use AdamW as the optimizer and dynamically update the learning rate with the stepped strategy, which is set to decay at the rate of 0.5 every 15 epochs. The initial weighting factors λ_1 and λ_2 of loss function are set into 0.2 and 0.1. The λ_1 with a stepped decay of 0.1 for every 25 epochs. In testing, we choose peak signal-to-noise ratio and structural similarity as the evaluation indexes.

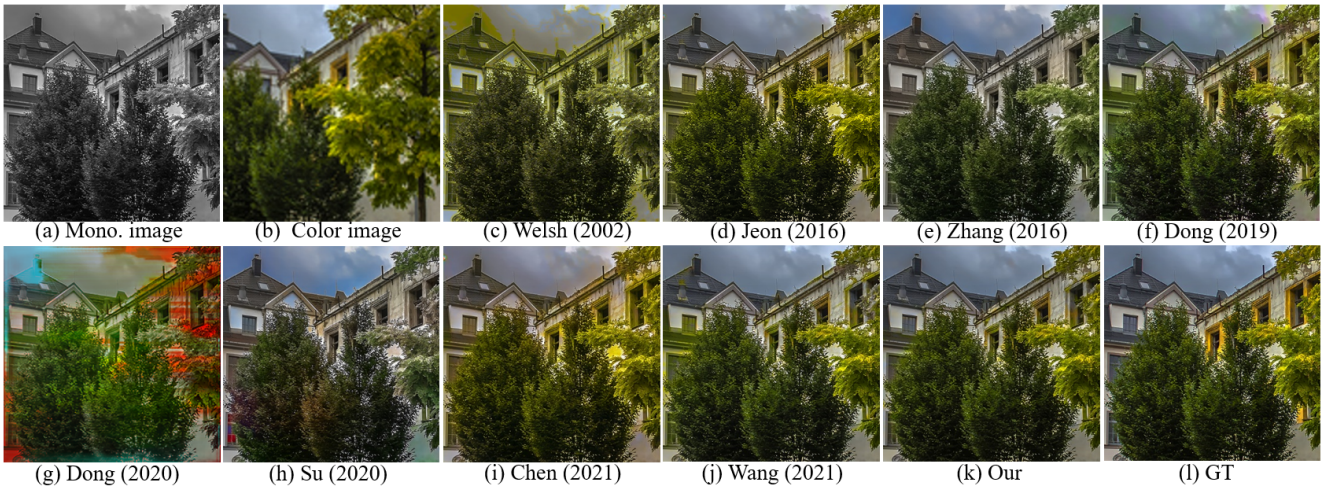


Figure 9. Visualization results on the Flickr1024 dataset. As can be seen, our method has better coloring ability and shows excellent nonlocal color transfer in the occluded area. More visualization results are presented in the Supplementary Material.

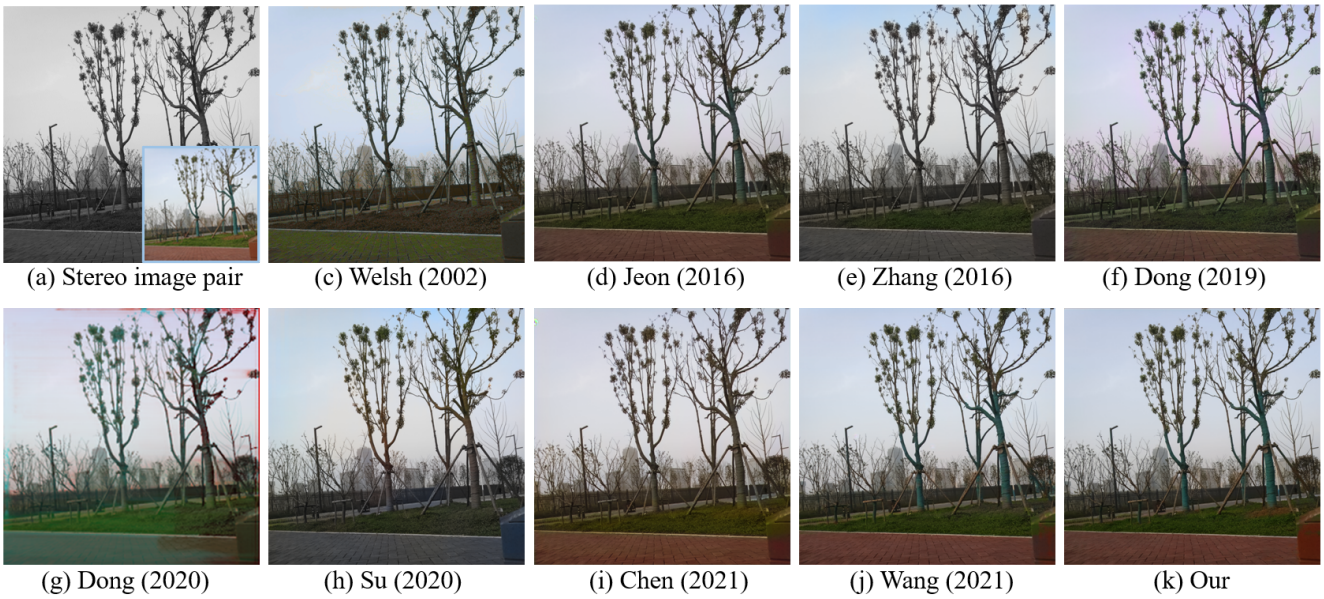


Figure 10. Visualization results on the real-world scenes. The stereo image pairs are collected by the smartphone HUAWEI P9 equipped with the mono-color dual-lens system. More visualization results are presented in the Supplementary Material.

4.4. Quantitative and Qualitative Results

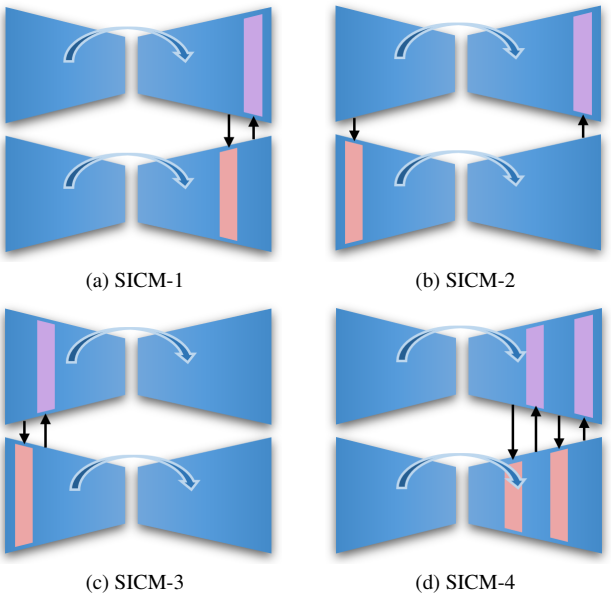
Baseline Methods. For performance comparison, we not only compare with the color transfer works [5, 6, 8, 14] based on the mono-color dual-lens system but also compare the exemplar-based colorization method [16] and the automatic colorization method [10, 18]. In addition, we introduce the representative stereo super-resolution method CPASSR [3] and make it suitable for the asymmetric color transfer.

Flickr1024. We compare all these methods on the Flickr1024 testset with three noise settings, as shown in Table 1. It can be seen that our model outperforms the baseline methods on all noise setups with less parameters. Benefiting from the enhancement of modal information consistency, our model exhibited excellent robustness on noise-added setups. The visualization results exhibited in Fig. 9 indicate that our method achieves the best colorization effects and excellent nonlocal modeling ability in occlusion areas.

Real-world Scenarios. To evaluate each model's generalization performance, we collect real-world image pairs with the HUAWEI P9 smartphone, which has a mono-color dual-lens configuration. The model used for generaliza-

Table 3. Ablation study on the consistent modality learning and the modulation position inside SICM.

SCM	$\mathcal{L}_{lp}$	SFM	$\mathcal{L}_{rp}$	$\mathcal{L}_{kd}$	PSNR (dB)	SSIM
				✓	30.57	0.943
✓	✓			✓	30.78	0.946
		✓	✓	✓	30.84	0.947
✓	✓	✓	✓		30.96	0.947
✓	✓	✓	✓	✓	**31.10**	**0.948**



(a) SICM-1 (b) SICM-2

(c) SICM-3 (d) SICM-4

Figure 11. Four different modulation architectures of SICM.

tion is trained based on the noise-free data (Setup1) of the Flickr1024 dataset. The comparison of the coloring results of the real scene is shown in Fig. 10. Our method gives the best real-world asymmetric color transfer results.

4.5. Ablation Studies

In this section, we conduct several ablation experiments to prove the effectiveness of all the proposed consistent modality learning-oriented techniques. First, we ablate the stereo modulation blocks (i.e., SFM and SCM block) with corresponding loss functions $\mathcal{L}_{lp}$, $\mathcal{L}_{rp}$, as shown in Table 3. The results indicate that both color modulation and frequency modulation are beneficial to the network. In addition, the effectiveness of the parallax attention distillation strategy is also proved by the ablation results.

Table 4. Ablation experiments on different SICM architectures.

Architecture	PSNR (dB)	SSIM
SICM-1	**31.10**	**0.948**
SICM-2	28.16	0.920
SICM-3	27.42	0.913
SICM-4	31.03	0.947

Furthermore, we also explored several variants of SICM to find the best modulation position. Specifically, we set up four different architectures for comparative experiments, as shown in Fig. 11. First, the SICM-1 performs frequency and color modulation in the decoder. Second, we design the SICM-2 to modulate the frequency in the encoder and the color in the decoder. Third, the modulation blocks of SICM-3 both in the encoder. Finally, to explore whether multiple modulations can further improve the performance, we designed SICM-4 equipped with four modulation blocks. As shown in Table 4, compared with SICM-2 and SICM-3 that modulate information in the encoder, SICM-1 and SICM-4 modulate information in the decoder show better results. Besides, the additional modulation blocks inside the SICM-4 do not improve the performance further.

5. CONCLUSION

In this paper, we develop a novel asymmetric color transfer framework, which performs consistent modality learning for establishing reliable stereo correspondence. Specifically, we present a Stereo Information Complementary Module (SICM) with stereo information modulation blocks (i.e., SFM and SCM block) to achieve consistent modality learning by information complementation. The visualization results exemplify that our SICM successfully narrow the gap between stereo modal information. In addition, we introduce a PCT block and the parallax attention distillation strategy to assist the establishment of reliable stereo correspondence for more precise color transfer. Extensive experiments prove that the consistent modality learning in our method is beneficial for establishing accurate stereo correspondence to promote the color transfer effect. Compared with other state-of-the-arts, we achieve superior performance and generalization capability. Furthermore, our model gives the best visualization results when applied to real-world scenarios.

Acknowledgments. This work was supported by the JKW Research Funds under Grant 20-163-14-LZ-001-004-01, and the Anhui Provincial Natural Science Foundation under Grant 2108085UD12. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

References

[1] Michael Bleyer and Sylvie Chambon. Does color really help in dense stereo matching. In *Proceedings of the International Symposium 3D Data Processing, Visualization and Transmission*. Citeseer, 2010.

[2] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018.

[3] Canqiang Chen, Chunmei Qing, Xiangmin Xu, and Patrick Dickinson. Cross parallax attention network for stereo image super-resolution. *IEEE Transactions on Multimedia*, 2021.

[4] Xuan Dong, Weixin Li, and Xiaojie Wang. Pyramid convolutional network for colorization in monochrome-color multi-lens camera system. *Neurocomputing*, 450:129–142, 2021.

[5] Xuan Dong, Weixin Li, Xiaojie Wang, and Yunhong Wang. Learning a deep convolutional network for colorization in monochrome-color dual-lens system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8255–8262, 2019.

[6] Xuan Dong, Weixin Li, Xiaojie Wang, and Yunhong Wang. Cycle-cnn for colorization towards real monochrome-color camera systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10721–10728, 2020.

[7] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.

[8] Hae-Gon Jeon, Joon-Young Lee, Sunghoon Im, Hyowon Ha, and In So Kweon. Stereo matching with color and monochrome cameras in low-light conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4086–4094, 2016.

[9] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

[10] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-aware image colorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7968–7977, 2020.

[11] Marc Comino Trinidad, Ricardo Martin Brualla, Florian Kainz, and Janne Kontkanen. Multi-view image fusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4101–4110, 2019.

[12] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning parallax attention for stereo image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12250–12259, 2019.

[13] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 606–615, 2018.

[14] Yicheng Wang, Jiayong Peng, Yueyi Zhang, Shan Liu, Xiaoyan Sun, and Zhiwei Xiong. Asymmetric stereo color transfer. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1–6, 2021.

[15] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Flickr1024: A large-scale dataset for stereo image super-resolution. In *Proceedings of the International Conference on Computer Vision Workshops*, pages 3852–3857, Oct 2019.

[16] Tomihisa Welsh, Michael Ashikhmin, and Klaus Mueller. Transferring color to greyscale images. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, pages 277–280, 2002.

[17] Yingchen Yu, Fangneng Zhan, Shijian Lu, Jianxiong Pan, Feiying Ma, Xuansong Xie, and Chunyan Miao. Wavefill: A wavelet-based generation network for image inpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 14114–14123, 2021.

[18] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision*, pages 649–666. Springer, 2016.