RATIO-RFE: FROM APPEARANCE FITTING TO SOURCE-RELATIVE RELIGHTING FIELD ESTIMATION FOR IMAGE HARMONIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Image harmonization aims to adjust a composite foreground to match the background while preserving its content and structure. Existing methods typically learn the desired target appearance directly, requiring appearance adjustment and content preservation to be handled within the same prediction process. This work instead formulates image harmonization as *source-relative relighting field estimation*. Motivated by the approximately shared foreground reflectance between the composite and harmonized images, the required source-to-target change is represented as a source-relative log-ratio relighting field, by canceling the shared reflectance. The harmonized foreground can then be obtained from the observed source through analytical reconstruction based on relighting field, without explicitly recovering reflectance or absolute lighting. Based on this formulation, a source-relative Relighting Field Estimation (in the log-Ratio form) based image harmonization framework, termed as Ratio-RFE, is proposed. It is developed with Phong-inspired Ambient-Diffuse-Specular (ADS) representations, where the Ambient, Diffuse and Specular components are estimated separately based on their different characteristics. An ADS-guided dynamic aggregation is further developed, where these representations guide spatially adaptive aggregation for unified relighting field estimation. Moreover, transformation-aligned learning is used to directly constrain the predicted log-ratio relighting field together with the reconstructed appearance. Experiments demonstrate state-of-the-art performance on iHarmony4 and strong cross-dataset generalization to ccHarmony without fine-tuning. Moreover, although trained only at 256×256, Ratio-RFE generalizes directly to 1024×1024 and 2048×2048 inputs without retraining, demonstrating strong cross-resolution generalization.

1 INTRODUCTION

Image compositing enables flexible visual content creation by inserting foreground objects into new backgrounds. However, differences in lighting conditions, capture settings, or post-processing between the source images can make the inserted foreground appear inconsistent with its surroundings. Image harmonization addresses this problem by adjusting the foreground appearance to match the background while preserving its content and structure (Tsai et al., 2017). A defining property of this task is that the foreground is already present: its content is given, while the appearance change required to fit the new environment need be estimated.

Existing methods commonly formulate image harmonization as appearance fitting through learned image mappings that generate harmonized images directly (Cong et al., 2020; Ling et al., 2021; Zhou et al., 2024; Chen et al., 2023) or through various appearance fitting functions including linear fitting, look-up tables (LUTs), curves, and filters (Cong et al., 2022; Ke et al., 2022; Xue et al., 2022; Guerreiro et al., 2023). In these formulations, the lighting effects to be changed and the structural content to be preserved remain intertwined in the image appearance. The model therefore needs to coordinate lighting adjustment and structure preservation within the same appearance-fitting process. There are also studies exploring intrinsic-image decomposition (Guo et al., 2021b; Careaga et al., 2023) or disentangled representations (Guo et al., 2021a; 2023) to separate the lighting and structure. However, explicit reflectance–lighting decomposition can be ambiguous, and errors in separating the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

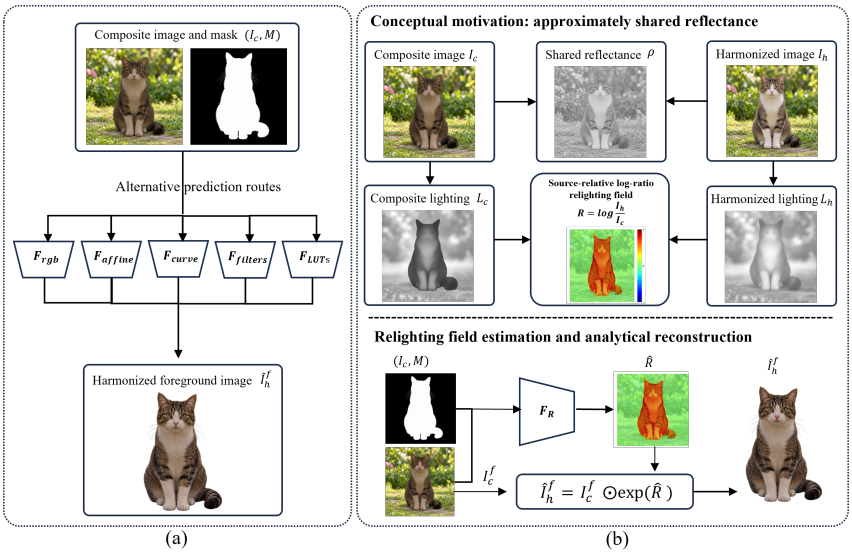

Figure 1: Motivation for source-relative relighting field estimation. (a) Representative image harmonization methods based on learned image mapping or various appearance-fitting functions, shown as alternative routes. These methods jointly handle lighting adjustment and structure preservation within the same appearance-fitting process. (b) Our source-relative relighting field estimation based harmonization. It exploits approximately shared foreground reflectance to represent the required appearance change as a source-relative relighting field in log-ratio form. The model then only needs to estimate the relighting field decoupled from the reflectance, and the final output can be generated via analytical reconstruction with the estimated relighting field.

two can cause lighting edits to alter foreground texture and local structure. This motivates a fundamental design question: how can the intrinsic properties shared between the source and target be exploited to isolate the required lighting-related change without accurate structure/appearance decomposition or estimation?

We address the above problem from an image-formation perspective based on lighting and reflectance. Under an idealized multiplicative model of lighting and reflectance, the composite/source and harmonized/target image share approximately the same foreground reflectance, which approximately cancels in a target-to-source intensity ratio. This motivates estimating lighting-related change of the source foreground without explicitly recovering either reflectance or absolute lighting maps. We term this source-relative lighting change of the foreground as a relighting field. A source-relative log-relighting formulation can be constructed in which a relighting field is represented in a log-ratio form and applied to the observed foreground through analytical reconstruction:

$$I_h^f = I_c^f \odot \exp(R_{\text{ratio}}(L_c^f, L_h^f)), \qquad R_{\text{ratio}}(L_c^f, L_h^f) = \log \frac{I_h^f}{I_c^f} \approx f(\hat{L}_c^f,\ \hat{L}_h^f | I_c^f, I_c^b), \tag{1}$$

where L_c^f and L_h^f denote the effective lighting contributions of the source foreground I_c^f and target foreground I_h^f, respectively. R_{ratio} represents the source-relative relighting field defined in the form of the log-ratio of the foreground. It cancels the reflectance and is determined by the lighting, while f describes its estimation function conditioned on the observed source foreground I_c^f and background I_c^b. Here $\hat{L}_c^f$ and $\hat{L}_h^f$ are conceptual lighting-state estimates; the network directly predicts the relighting field without explicitly recovering these lighting maps. The logarithm changes multiplicative correction from various factors into additive coordinates, with $R_{\text{ratio}} = 0$ corresponding to the identity transformation. This shifts the learning target from what target appearance to fit to how to relight the source foreground, explicitly linking relighting field estimation with analytical reconstruction. Figure 1 contrasts representative appearance-fitting based methods with our source-relative relighting field based formulation, illustrating the shared-reflectance motivation, and the corresponding field estimation and analytical reconstruction.

Upon the above formulation, an Ambient-Diffuse-Specular guided relighting field estimation (ADS-RFE) model is further developed based on the Phong shading, where relative lighting change depends jointly on ambient, diffuse, and specular contributions. Three branches are designed: the Ambient pathway emphasizes global and low-frequency appearance variation, the Diffuse pathway models spatially varying shading patterns, and the Specular pathway focuses on localized high-frequency features. A mask-aware multi-kernel dynamic aggregation module (ADS-MK) is further introduced, where complementary ADS representations guide the spatial aggregation of dynamically filtered foreground responses according to local foreground context, yielding a unified source-relative relighting field. Moreover, our ADS-RFE framework aligns supervision in the domain of the relighting field: each paired source and target foreground provides a log-ratio-form relighting field supervision target, which is used alongside image reconstruction. The image-space objective constrains the resulting appearance, while the auxiliary ratio-space objective directly constrains the relighting field applied to the source.

Our contributions are summarized as follows:

- **Source-relative log-ratio relighting formulation and analytical reconstruction for image harmonization.** Motivated by the approximate cancellation of shared foreground reflectance, we formulate harmonization through source-relative relighting field estimation in a log-ratio form and embed the corresponding analytical reconstruction into the network. To our knowledge, we are the first to reformulate image harmonization as source-relative relighting field estimation.
- **Phong-inspired ADS representation and dynamic aggregation for relighting field estimation.** We develop complementary feature pathways for global appearance variation, spatially varying shading, and localized high-frequency variation, and use these features to guide the spatial aggregation of dynamically filtered decoder features for relighting field estimation.
- **Transformation-aligned learning in the domain of relighting field.** Building on the explicit relighting formulation, we complement image reconstruction with direct log-ratio-form relighting field supervision to make the network explicitly learn the transformation in the domain of relighting field for image harmonization.

Experiments on iHarmony4 (Cong et al., 2020) demonstrate the best overall performance among the compared methods, while direct evaluation on ccHarmony (Niu et al., 2023) without fine-tuning shows strong cross-dataset generalization. Cross-resolution evaluation and ablation studies further validate the proposed formulation and its architectural components.

2 Related Work

2.1 Image Harmonization by Appearance Fitting

Most image harmonization methods formulate the task as learning the target foreground appearance from a composite image. Early deep approaches directly predict harmonized images (Tsai et al., 2017; Cong et al., 2020), while later methods improve foreground–background interaction through region-aware normalization, dynamic feature adaptation, or stronger generative modeling (Ling et al., 2021; Chen et al., 2023; Zhou et al., 2024). Another line of work adopts structured appearance transformations, such as RGB mappings, curves, filters, and look-up tables, for more efficient or interpretable harmonization (Cong et al., 2022; Ke et al., 2022; Xue et al., 2022; Guerreiro et al., 2023). Despite different implementations, these methods mainly optimize the desired target appearance, where lighting adjustment and foreground-content preservation are handled within the same appearance-fitting process. In contrast, we explicitly model the required change relative to the observed source foreground as a source-relative log-ratio relighting field and recover the harmonized image through analytical reconstruction.

2.2 Intrinsic and Illumination-Aware Harmonization

Several studies incorporate intrinsic decomposition or explicit illumination modeling into image harmonization. Intrinsic Image Harmonization decomposes composite images into reflectance and

illumination components (Guo et al., 2021b), while subsequent works exploit intrinsic, reshading, or illumination-aware representations to better model lighting inconsistency between foreground and background (Bao et al., 2022; Careaga et al., 2023; Hu et al., 2024; Ren et al., 2024). Transformer-based methods further explore disentangled representations for separating appearance-related variation from foreground content (Guo et al., 2021a; 2023). Although these methods provide a more structured treatment of illumination inconsistency, explicitly recovering reflectance, illumination, or other intermediate lighting representations introduces additional estimation ambiguity. In contrast, we exploit approximately shared foreground reflectance only as an image-formation prior, allowing it to approximately cancel in the target-to-source intensity ratio and directly defining a source-relative log-ratio relighting field without explicitly recovering reflectance or absolute lighting.

3 METHOD

Given a composite image $I_c \in [0,1]^{H\times W\times 3}$ and a binary foreground mask $M \in \{0,1\}^{H\times W}$, our goal is to harmonize the foreground while retaining its content and the original background. Ratio-RFE estimates a source-relative relighting field and applies it through analytical reconstruction. We first formulate this relighting field and examine its lighting structure, then describe the network used to estimate it and the corresponding learning objective.

3.1 SOURCE-RELATIVE RELIGHTING FORMULATION VIA REFLECTANCE CANCELLATION

Under an idealized multiplicative image-formation model, foreground appearance is expressed as $I_t^f = \rho_t^f \odot L_t^f$, where $t \in \{c, h\}$ denotes the composite or harmonized state and the superscript f denotes the foreground. Here, ρ_t^f denotes reflectance, and L_t^f represents the effective lighting contribution. Since harmonization aims to retain the foreground object and its material properties, approximately shared reflectance is assumed, $\rho_h^f \approx \rho_c^f$. Taking the target-to-source intensity ratio and exploiting this approximate reflectance consistency, the source-relative relighting field can be expressed as

$$R_{\text{ratio}} = \log \frac{I_h^f}{I_c^f} = \log \frac{\rho_h^f \odot L_h^f}{\rho_c^f \odot L_c^f} \approx \log \frac{L_h^f}{L_c^f}. \tag{2}$$

The ratio approximately cancels the shared foreground reflectance, while the logarithm converts the multiplicative source-to-target lighting change into an additive representation, with $R_{\text{ratio}} = 0$ corresponding to the identity transformation. The source foreground provides an explicit reference for the required appearance change. The image-formation model supplies a physical motivation for this representation rather than an exact radiometric description of nonlinear RGB images.

To examine the lighting structure underlying this relighting field R_{ratio}, a simplified Phong-inspired model (Phong, 1975) is adopted, where effective lighting is decomposed into ambient, diffuse, and specular contributions, $L_t^f = A_t^f + D_t^f + S_t^f$. Omitting the foreground superscript and considering positive lighting and ambient terms, $d_t = D_t/A_t$ and $s_t = S_t/A_t$ are defined, yielding $L_t = A_t(1 + d_t + s_t)$. The corresponding source-relative log-ratio becomes

$$\begin{aligned} R_{\text{ratio}} &\approx \log \frac{A_h + D_h + S_h}{A_c + D_c + S_c} \\ &= \log \frac{A_h}{A_c} + \log \frac{1 + d_h + s_h}{1 + d_c + s_c} = \log \frac{A_h}{A_c} + \log \left(1 + \frac{\Delta d + \Delta s}{q_c}\right). \end{aligned} \tag{3}$$

where $q_c = 1 + d_c + s_c$, $\Delta d = d_h - d_c$, and $\Delta s = s_h - s_c$. Since the potentially large base-lighting change is isolated in $\log(A_h/A_c)$, the remaining normalized diffuse and specular changes are relatively small compared with q_c. Thus $|(\Delta d + \Delta s)/q_c| \ll 1$, and with the first-order approximation $\log(1 + x) \approx x$, Equation 3 can be transformed to

$$R_{\text{ratio}} \approx \underbrace{\log \frac{A_h}{A_c}}_{R_A} + \underbrace{\frac{\Delta d}{q_c}}_{R_D} + \underbrace{\frac{\Delta s}{q_c}}_{R_S}. \tag{4}$$

Equation 3 shows that the relighting field depends jointly on changes in the three lighting contributions, while equation 4 provides a local interpretation of their effects in log space. Together

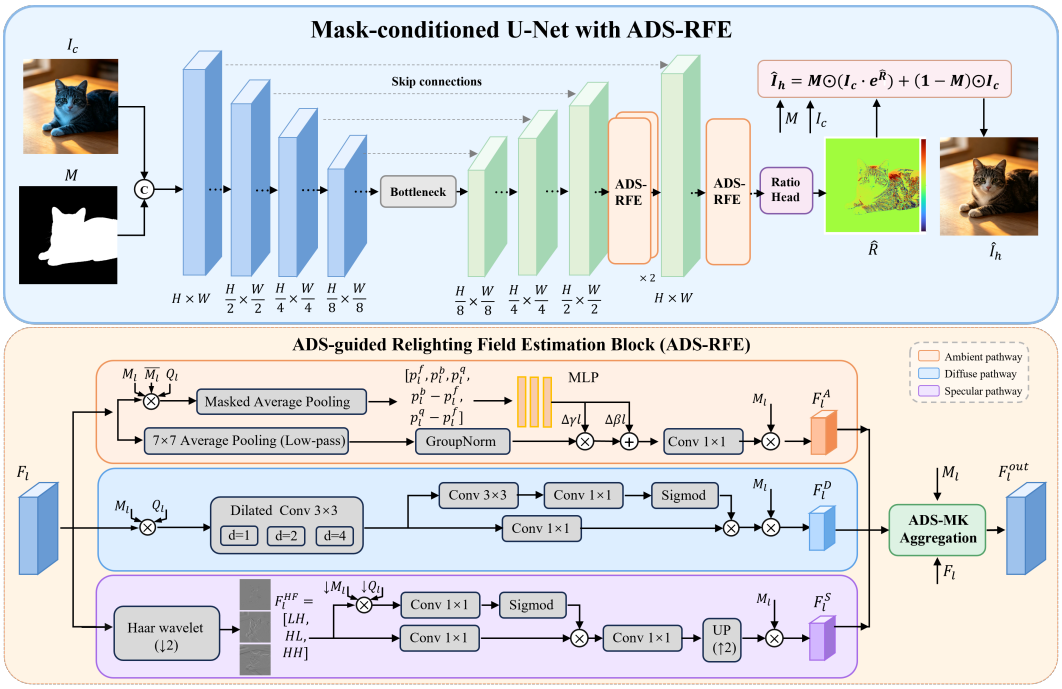

Figure 2: Overview of the proposed Ratio-RFE framework. Given a composite image I_c and its foreground mask M, a mask-conditioned U-Net extracts multi-scale representations, with Phong-inspired ADS-RFE blocks inserted at selected decoder stages. Each ADS-RFE block constructs complementary Ambient, Diffuse and Specular representations, and uses them to guide ADS-MK aggregation. The final decoder representation is fed to a ratio head to predict a unified source-relative log-ratio relighting field $\hat{R}$. The harmonized image $\hat{I}_h$ is then analytically reconstructed from I_c and $\hat{R}$, while the background is directly preserved.

with their characteristic spatial behavior, this analysis motivates complementary features for global, spatially varying, and localized appearance changes. The terms R_A, R_D, R_S guide representation design and are not independently predicted or supervised; the network estimates a unified relighting field without imposing the first-order approximation on its output.

Given the predicted field $\widehat{R}$, the harmonized image follows directly from the inverse of the log-ratio definition:

$$\widehat{I}_h = (1 - M) \odot I_c + M \odot \left[I_c \odot \exp(\widehat{R}) \right]. \tag{5}$$

The source foreground therefore serves as both the reference for relighting field prediction and the input to analytical reconstruction. A zero relighting field yields the identity mapping, while the background remains unchanged.

3.2 ADS-Guided Relighting Field Estimation

Network overview. Building on Section 3.1, complementary ADS features are constructed to estimate the unified source-relative relighting field. A mask-conditioned U-Net with RAIN normalization (Ling et al., 2021) is adopted, taking $[I_c, M]$ as input. At selected decoder stages, ADS-RFE blocks construct complementary Ambient, Diffuse, and Specular representations, which guide spatially adaptive aggregation through ADS-MK, while the final decoder feature is mapped to the source-relative log-ratio relighting field $\widehat{R}$.

Let $F_l \in \mathbb{R}^{H_l \times W_l \times C_l}$ denote the decoder feature at stage l, and M_l its corresponding foreground mask. The three pathways produce F_l^A, F_l^D, F_l^S. The background mask is defined as $\bar{M}_l = 1 - M_l$ and the surrounding-context mask as $Q_l = \mathrm{Dilate}(M_l) - M_l$. Channel concatenation is denoted by $[\cdot]$, while $\mathrm{Conv}_{k\times k}$ denotes a convolution with kernel size $k \times k$.

Ambient pathway. To account for global appearance discrepancies, the Ambient pathway uses regional statistics to modulate low-frequency foreground features. Masked average pooling produces foreground, background, and surrounding-context descriptors p_l^f, p_l^b, and p_l^q. An MLP maps $[p_l^f, p_l^b, p_l^q, p_l^b - p_l^f, p_l^q - p_l^f]$ to channel-wise modulation parameters $\Delta\gamma_l$ and $\Delta\beta_l$, which are applied to the low-frequency feature F_l^{low} obtained by average pooling and group normalization:

$$F_l^A = M_l \odot \text{Conv}_{1\times1}\left(\Delta\gamma_l \odot F_l^{\text{low}} + \Delta\beta_l\right). \tag{6}$$

Diffuse pathway. To capture spatially varying shading transitions, the Diffuse pathway combines foreground features with nearby context. Cascaded dilated convolutions with dilation rates 1, 2, and 4 process $[M_l \odot F_l, Q_l \odot F_l, M_l, Q_l]$ to yield an expanded context representation H_l^D. A learned spatial gate G_l^D modulates the projected features within the foreground:

$$F_l^D = M_l \odot G_l^D \odot \text{Conv}_{1\times1}(H_l^D). \tag{7}$$

Specular pathway. To represent localized appearance variations associated with highlights, the Specular pathway applies Haar wavelet decomposition to F_l. The half-resolution LH, HL, and HH subbands are combined into F_l^{HF}. A spatial gate G_l^S is predicted from foreground and surrounding-context high-frequency features, using masks resized to the subband resolution. The gated feature is projected and upsampled to the decoder resolution:

$$H_l^S = \text{Conv}_{1\times1}\left(F_l^{\text{HF}}\right) \odot G_l^S, \qquad F_l^S = M_l \odot \text{Up}\left(\text{Conv}_{1\times1}(H_l^S)\right). \tag{8}$$

ADS-guided dynamic aggregation. Inspired by the mask-aware global dynamic design of HDNet (Chen et al., 2023), we develop ADS-MK, a mask-aware multi-kernel dynamic aggregation module that uses ADS features to guide the spatial aggregation of dynamically filtered responses. At decoder stage l, the dynamic filtering operator takes F_l as input and produces one background response and three foreground candidates:

$$[Y_l^{\text{bg}}, Y_l^1, Y_l^2, Y_l^3] = \mathcal{K}_l^{\text{MK}}(F_l), \tag{9}$$

Here, $\mathcal{K}_l^{\text{MK}}$ denotes our mask-aware multi-kernel dynamic filtering operator. All three ADS pathways jointly condition the spatial mixing weights of the foreground candidates:

$$[g_l^1, g_l^2, g_l^3] = \text{Softmax}\left(\mathcal{T}_l([F_l, F_l^A, F_l^D, F_l^S, M_l])\right), \tag{10}$$

where $\mathcal{T}_l$ is a learnable weight-prediction network. Softmax normalizes the three candidate weights at each spatial location. The foreground candidates are spatially aggregated and combined with the background feature:

$$Y_l^{\text{fg}} = \sum_{j=1}^{3} g_l^j \odot Y_l^j, \qquad \widetilde{F}_l = \bar{M}_l \odot Y_l^{\text{bg}} + M_l \odot Y_l^{\text{fg}}. \tag{11}$$

Thus, the ADS features guide the local combination of foreground candidates. The fused feature $\widetilde{F}_l$ is passed to subsequent decoding layers. The ratio head maps the final decoder feature to a raw three-channel output Z, from which a unified log-ratio field is obtained as $\widehat{R} = M \odot r_{\max} \tanh(Z)$, where $r_{\max} > 0$ controls the range of the predicted relighting field. The harmonized image is then obtained through equation 5.

3.3 Transformation-aligned learning in the domain of relighting field

To align the learning with the relighting field estimation based formulation, supervision on both the reconstructed appearance and the relighting field are applied to the source foreground. For paired composite and ground-truth images, the target log-ratio is defined as

$$R^{\text{gt}} = M \odot \text{clip}\left(\log \frac{I_h^{\text{gt}}}{I_c}, -r_{\max}, r_{\max}\right), \tag{12}$$

Table 1: Quantitative comparison on iHarmony4. The best and second-best results are marked in **bold** and underlined, respectively. ↓/↑ indicate lower/higher is better. All denotes results on the full test set. Raiha results are reported under the non-reference setting.

Method	Venue	HAdobe5k		HFlickr		HCOCO		Hday2night		All	
		MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑
Composite	–	345.54	28.16	264.35	28.32	69.37	33.94	109.65	34.01	172.47	31.63
iS^{2}AM (Cun & Pun, 2020)	TIP	21.60	38.28	69.43	33.65	16.15	39.40	40.39	37.87	24.13	38.41
RainNet (Ling et al., 2021)	CVPR	43.35	36.22	110.59	31.64	29.52	37.08	57.40	34.83	40.29	36.12
BargainNet (Cong et al., 2021)	ICME	39.94	35.34	97.32	31.34	24.84	37.03	50.98	35.67	37.82	35.88
Intrinsic (Guo et al., 2021b)	CVPR	43.02	35.20	105.13	31.34	24.92	37.16	55.53	35.96	38.71	35.90
D-HT (Guo et al., 2021a)	ICCV	38.53	36.88	74.51	33.13	16.89	38.76	53.01	37.10	30.30	37.55
Harmonizer (Ke et al., 2022)	ECCV	21.89	37.64	64.81	33.63	17.34	38.77	33.14	37.56	24.26	37.84
DCCF (Xue et al., 2022)	ECCV	19.90	38.27	60.41	33.94	14.87	39.52	49.32	37.88	22.05	38.50
CDTNet (Cong et al., 2022)	CVPR	20.62	38.24	68.61	33.55	16.25	39.15	36.72	37.95	23.75	38.23
PCT-Net (Guerreiro et al., 2023)	CVPR	20.86	40.11	44.30	35.13	10.72	40.77	44.74	37.65	18.04	39.89
HDNet (Chen et al., 2023)	ACM MM	13.58	41.17	47.39	35.81	11.60	41.04	31.97	38.85	16.55	40.46
SP-IC (Cai et al., 2024)	TMM	18.17	38.91	68.85	33.88	14.82	39.73	41.32	37.90	22.47	38.81
AICT (Meng et al., 2024)	NeurIPS	25.87	38.12	53.42	34.42	10.50	40.58	43.65	37.50	20.38	39.12
WAVSSNet (Zhou et al., 2026)	IJCV	16.30	41.01	46.09	35.64	10.30	41.16	50.26	37.69	16.77	40.43
Raiha (Wang et al., 2026)	PR	13.32	41.55	45.41	36.44	10.53	41.82	30.38	39.77	15.60	41.10
Ratio-RFE	**Ours**	**10.74**	**42.49**	**40.73**	**36.84**	**9.67**	**42.33**	**26.62**	**40.46**	**13.76**	**41.73**

where I_h^{gt} denotes the ground-truth harmonized image. The target is clipped to the same range as the bounded prediction $\widehat{R}$. We train the framework end-to-end using a joint objective that combines a foreground mean squared reconstruction error with auxiliary foreground ℓ_1 ratio supervision:

$$\mathcal{L} = \frac{1}{N_f}\Big[\big\|M \odot (\widehat{I}_h - I_h^{\mathrm{gt}})\big\|_F^2 + \lambda_R \big\|M \odot (\widehat{R} - R^{\mathrm{gt}})\big\|_1\Big], \tag{13}$$

where $N_f = 3\sum_p M_p$ is the number of foreground color values, p indexes spatial locations, and λ_R balances the two terms. The image loss constrains the reconstructed foreground appearance, while the auxiliary ratio loss directly constrains the estimated relighting field.

4 Experiments

4.1 Experimental Setup

Ratio-RFE is evaluated on iHarmony4 (Cong et al., 2020) and ccHarmony (Niu et al., 2023). The official iHarmony4 splits are used, with all training images resized to 256×256. MSE and PSNR are reported on iHarmony4, while fMSE is additionally reported on ccHarmony. For cross-resolution evaluation, fMSE, MSE, PSNR, and SSIM are reported following AICT (Meng et al., 2024). For ccHarmony, the model trained only on iHarmony4 is directly evaluated on its 1,180 test pairs at 256×256, without using the ccHarmony training set or any fine-tuning.

The model is trained on iHarmony4 using Adam for 120 epochs with a batch size of 12. The initial learning rate is 10^{-3} and is reduced by a factor of 10 at epochs 100 and 110. We use $r_{\max} = 4$ and $\lambda_R = 0.1$ throughout. Unless otherwise specified, all ablation variants share the same training and evaluation settings.

4.2 Comparison with State-of-the-Art Methods

Ratio-RFE is compared with representative image harmonization methods on iHarmony4 and ccHarmony. As shown in Table 1, Ratio-RFE achieves the best overall performance on iHarmony4. Without any ccHarmony fine-tuning, it also achieves the best results in Table 2, demonstrating strong cross-dataset generalization. Figure 3 further shows improved illumination and color consistency while preserving foreground details.

4.3 Cross-Resolution Generalization

Although Ratio-RFE is trained exclusively at 256×256, the same checkpoint is directly applied to 1024×1024 and 2048×2048 inputs without retraining or fine-tuning, following the high-resolution

Table 2: Quantitative comparison on ccHarmony at 256×256. Ratio-RFE is trained only on iHarmony4 and evaluated on ccHarmony without fine-tuning. The best and second-best results are marked in **bold** and underlined, respectively.

Method	Venue	fMSE↓	MSE↓	PSNR↑
DoveNet (Cong et al., 2020)	CVPR	880.94	110.84	31.64
RainNet (Ling et al., 2021)	CVPR	519.32	58.11	34.78
Intrinsic (Guo et al., 2021b)	CVPR	636.28	83.72	33.64
D-HT (Guo et al., 2021a)	ICCV	514.47	55.73	35.07
iS^{2}AM (Cun & Pun, 2020)	TIP	264.84	28.83	36.05
CDTNet (Cong et al., 2022)	CVPR	264.51	27.87	36.62
Harmonizer (Ke et al., 2022)	ECCV	402.09	43.31	34.68
DCCF (Xue et al., 2022)	ECCV	259.83	29.25	36.62
GiftNet (Niu et al., 2023)	ICCV	235.20	24.55	37.59
AICT (Meng et al., 2024)	NeurIPS	232.66	24.14	38.46
WAVSSNet (Zhou et al., 2026)	IJCV	293.98	41.13	38.54
Ratio-RFE	**Ours**	**152.25**	**18.45**	**39.78**

Table 3: Cross-resolution comparison on the HAdobe5k subset. Ratio-RFE is trained only at 256×256. The best and second-best results are marked in **bold** and underlined, respectively.

1024×1024					2048×2048				
Method	**fMSE↓**	**MSE↓**	**PSNR↑**	**SSIM↑**	**Method**	**fMSE↓**	**MSE↓**	**PSNR↑**	**SSIM↑**
iS^{2}AM	168.85	25.03	38.29	0.9846	iS^{2}AM	271.59	46.37	36.57	0.9838
CDTNet-256	152.13	21.24	38.77	0.9868	CDTNet-512	159.13	23.35	38.45	0.9853
Harmonizer	229.17	32.67	35.75	0.9081	Harmonizer	208.93	27.79	36.58	0.8992
DCCF	196.41	22.63	38.39	0.9846	DCCF	197.23	23.00	38.34	0.9851
PCT-Net	171.04	22.57	39.39	0.9860	PCT-Net	156.56	20.38	39.83	0.9870
AICT	156.81	19.50	39.67	0.9864	AICT	147.99	**17.92**	40.07	**0.9871**
Ratio-RFE	**97.40**	**14.91**	**41.35**	**0.9902**	**Ratio-RFE**	**126.23**	19.85	**40.13**	0.9870

evaluation protocol of AICT (Meng et al., 2024). As shown in Table 3, Ratio-RFE achieves the best results across all four metrics at 1024×1024. At 2048×2048, it achieves the best fMSE and PSNR and the second-best MSE, while AICT obtains a marginally higher SSIM. These results demonstrate strong generalization beyond the training resolution. Additional cross-resolution ablations further show that this advantage does not arise from the architecture alone. Compared with direct RGB prediction under the same ratio-space supervision, the complete log-ratio formulation reduces MSE from 18.702 to 14.910 at 1024×1024 and from 23.745 to 19.845 at 2048×2048, showing a consistent advantage across resolutions. Detailed results are provided in the Appendix.



Figure 3: Qualitative comparisons on iHarmony4 and ccHarmony. The first two rows show examples from iHarmony4, while the last row shows an example from ccHarmony. From left to right: composite input, RainNet, Harmonizer, AICT, Raiha, Ratio-RFE, and ground truth.

Table 4: Ablation of source-relative prediction and transformation-aligned learning. The best and second-best results are marked in **bold** and underlined, respectively.

Prediction	Objective	fMSE↓	MSE↓	PSNR↑	SSIM↑
Direct RGB	$\mathcal{L}_{\text{img}}$	224.063	22.07	39.63	0.993
Direct RGB	$\mathcal{L}_{\text{img}} + \lambda_R \mathcal{L}_R$	189.452	20.98	40.41	0.993
Log-ratio	$\mathcal{L}_{\text{img}}$	193.856	19.02	40.22	0.993
Log-ratio	$\mathcal{L}_{\text{img}} + \lambda_R \mathcal{L}_R$	**137.580**	**13.76**	**41.73**	**0.994**

Table 5: Ablations of (a) Phong-inspired ADS pathways and (b) ADS-MK aggregation. The best and second-best results are marked in **bold** and underlined, respectively.

(a) Phong-inspired ADS pathways

A	D	S	fMSE↓	MSE↓	PSNR↑	SSIM↑
×	×	×	321.16	32.20	37.89	0.990
✓	×	×	196.92	19.24	40.07	0.993
×	✓	×	192.86	18.93	40.21	0.993
×	×	✓	186.18	18.37	40.29	0.993
✓	✓	✓	**137.58**	**13.76**	**41.73**	**0.994**

(b) ADS-MK aggregation

Strategy	fMSE↓	MSE↓	PSNR↑	SSIM↑
Sum	259.53	26.15	38.70	0.991
Concat	248.06	24.80	38.89	0.991
Scalar	276.68	27.16	38.51	0.991
ADS-MK	**137.58**	**13.76**	**41.73**	**0.994**

4.4 Ablation Studies

Ablation studies examine the source-relative prediction formulation and supervision, the Phong-inspired ADS representation, and the ADS-MK aggregation strategy.

Source-relative formulation and transformation-aligned learning. The effects of prediction formulation and supervision are isolated while keeping the ADS architecture unchanged. For the direct-RGB variants, the relighting field used in $\mathcal{L}_R$ is derived from the predicted harmonized foreground using the same clipped log-ratio transformation as equation 12. As shown in Table 4, both log-ratio prediction and ratio-space supervision improve the baseline, while their combination yields the strongest overall performance.

Phong-inspired ADS representation and ADS-MK aggregation. As shown in Table 5, each ADS pathway improves over the baseline, while their joint use yields a substantial gain, supporting their complementary roles in modeling different lighting variations. ADS-MK further outperforms direct summation, channel concatenation, and adaptive scalar weighting, demonstrating the effectiveness of ADS-guided dynamic aggregation.

Additional analyses of the relighting range $r_{\max}$, foreground-ratio robustness, cross-resolution formulation variants, and additional qualitative and ADS visualizations are provided in the Appendix.

5 Conclusion

We presented Ratio-RFE, an image harmonization framework that reformulates foreground appearance adjustment as source-relative relighting field estimation with analytical reconstruction. By exploiting approximately shared foreground reflectance as an image-formation prior, the proposed formulation represents the required source-to-target appearance change in a log-ratio form without explicitly recovering reflectance or absolute lighting. To estimate this relighting field, ADS-RFE blocks construct complementary Ambient, Diffuse, and Specular representations for global low-frequency variation, spatially varying shading, and localized high-frequency variation, while ADS-MK uses these representations to guide the adaptive aggregation of dynamic foreground responses. Transformation-aligned learning further constrains the estimated relighting field in addition to the reconstructed appearance. Experiments demonstrate state-of-the-art performance on iHarmony4, while direct evaluation on ccHarmony without fine-tuning shows strong cross-dataset generalization. Ablation studies further validate the proposed relighting formulation, ADS representations, ADS-MK aggregation, and direct ratio-space supervision. Moreover, although trained only at 256×256, Ratio-RFE generalizes directly to 1024×1024 and 2048×2048 inputs without retraining.

AI USE STATEMENT

Generative AI tools were used solely for grammar correction and language polishing to improve the clarity and readability of the manuscript. All research ideas and technical content were developed by the authors, all experiments were conducted by the authors, and all results and conclusions were verified by the authors.

REFERENCES

Zhongyun Bao, Chengjiang Long, Gang Fu, Daquan Liu, Yuanzhen Li, Jiaming Wu, and Chunxia Xiao. Deep image-based illumination harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18542–18551, 2022.

Xun Cai, Qingjie Shi, Yanbo Gao, Shuai Li, Wei Hua, and Tian Xie. A structure-preserving and illumination-consistent cycle framework for image harmonization. *IEEE Transactions on Multimedia*, 26:51–64, 2024. doi: 10.1109/TMM.2023.3260620.

Chris Careaga, S. Mahdi H. Miangoleh, and Yağız Aksoy. Intrinsic harmonization for illumination-aware compositing. In *SIGGRAPH Asia 2023 Conference Papers*, 2023. doi: 10.1145/3610548.3618178.

Haoxing Chen, Zhangxuan Gu, Yaohui Li, Jun Lan, Changhua Meng, Weiqiang Wang, and Huaxiong Li. Hierarchical dynamic image harmonization. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 1422–1430, 2023. doi: 10.1145/3581783.3611747.

Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. DoveNet: Deep image harmonization via domain verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8394–8403, 2020.

Wenyan Cong, Li Niu, Jianfu Zhang, Jing Liang, and Liqing Zhang. BargainNet: Background-guided domain translation for image harmonization. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2021. doi: 10.1109/ICME51207.2021.9428394.

Wenyan Cong, Xinhao Tao, Li Niu, Jing Liang, Xuesong Gao, Qihao Sun, and Liqing Zhang. High-resolution image harmonization via collaborative dual transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18470–18479, 2022.

Xiaodong Cun and Chi-Man Pun. Improving the harmony of the composite image by spatial-separated attention module. *IEEE Transactions on Image Processing*, 29:4759–4771, 2020. doi: 10.1109/TIP.2020.2975979.

Julian Jorge Andrade Guerreiro, Mitsuru Nakazawa, and Björn Stenger. PCT-Net: Full resolution image harmonization using pixel-wise color transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5917–5926, 2023.

Zonghui Guo, Dongsheng Guo, Haiyong Zheng, Zhaorui Gu, Bing Zheng, and Junyu Dong. Image harmonization with transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14870–14879, 2021a.

Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. Intrinsic image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16367–16376, 2021b.

Zonghui Guo, Zhaorui Gu, Bing Zheng, Junyu Dong, and Haiyong Zheng. Transformer for image harmonization and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (11):12960–12977, 2023. doi: 10.1109/TPAMI.2022.3207091.

Zhongyun Hu, Jiahao Li, Xue Wang, and Qing Wang. Spatially-varying illumination-aware indoor harmonization. *International Journal of Computer Vision*, 132(7):2473–2492, 2024.

Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson W. H. Lau. Harmonizer: Learning to perform white-box image and video harmonization. In *Computer Vision – ECCV 2022*, pp. 690–706, 2022. doi: 10.1007/978-3-031-19784-0_40.

Jun Ling, Han Xue, Li Song, Rong Xie, and Xiao Gu. Region-aware adaptive instance normalization for image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9361–9370, 2021.

Quanling Meng, Qinglin Liu, Zonglin Li, Xiangyuan Lan, Shengping Zhang, and Liqiang Nie. High-resolution image harmonization with adaptive-interval color transformation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 13769–13793, 2024.

Li Niu, Linfeng Tan, Xinhao Tao, Junyan Cao, Fengjun Guo, Teng Long, and Liqing Zhang. Deep image harmonization with globally guided feature transformation and relation distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7723–7732, October 2023.

Bui Tuong Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18 (6):311–317, 1975. doi: 10.1145/360825.360839.

Mengwei Ren, Wei Xiong, Jae Shin Yoon, Zhixin Shu, Jianming Zhang, HyunJoon Jung, Guido Gerig, and He Zhang. Relightful harmonization: Lighting-aware portrait background replacement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6452–6462, 2024.

Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3789–3797, 2017.

Haolin Wang, Ming Liu, Zifei Yan, Chao Zhou, Longan Xiao, and Wangmeng Zuo. Retrieval-augmented image harmonization. *Pattern Recognition*, 172:112556, 2026. doi: 10.1016/j.patcog.2025.112556.

Ben Xue, Shenghui Ran, Quan Chen, Rongfei Jia, Binqiang Zhao, and Xing Tang. DCCF: Deep comprehensible color filter learning framework for high-resolution image harmonization. In *Computer Vision – ECCV 2022*, pp. 300–316, 2022. doi: 10.1007/978-3-031-20071-7_18.

Pengfei Zhou, Fangxiang Feng, and Xiaojie Wang. DiffHarmony: Latent diffusion model meets image harmonization. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pp. 1130–1134, 2024.

Zhaorun Zhou, Jinshan Pan, and Jinhui Tang. Wavelet-based adaptive vision state space network for image harmonization. *International Journal of Computer Vision*, 134(9):412, 2026. doi: 10.1007/s11263-026-02981-2.

A Cross-Resolution Formulation Analysis

To further examine whether the cross-resolution generalization of Ratio-RFE comes from the proposed formulation rather than the network architecture alone, we compare direct RGB prediction and log-ratio prediction under the same backbone and supervision settings. The model is trained only at 256×256 and directly evaluated at 1024×1024 and 2048×2048 without retraining or fine-tuning.

As shown in Table 6, combining log-ratio prediction with ratio-space supervision consistently yields the lowest fMSE and MSE and the highest PSNR at both resolutions. In particular, compared with direct RGB prediction under the same ratio-space supervision, the complete formulation reduces MSE from 18.702 to 14.910 at 1024×1024 and from 23.745 to 19.845 at 2048×2048. Although the direct-RGB variants obtain marginally higher SSIM, the complete log-ratio formulation provides substantially better foreground error and pixel fidelity across resolutions.

Table 6: Cross-resolution ablation of the prediction formulation and transformation-aligned ratio-space supervision. All models are trained only at 256×256 and directly evaluated at higher resolutions. The best and second-best results are marked in **bold** and underlined, respectively.

Prediction	Objective	1024×1024				2048×2048			
		fMSE↓	MSE↓	PSNR↑	SSIM↑	fMSE↓	MSE↓	PSNR↑	SSIM↑
Direct RGB	$\mathcal{L}_{\text{img}}$	136.542	21.389	40.024	0.9907	162.790	25.745	39.079	0.9881
Direct RGB	$\mathcal{L}_{\text{img}} + \lambda_R \mathcal{L}_R$	122.883	18.702	40.474	**0.9912**	156.570	23.745	39.192	**0.9884**
Log-ratio	$\mathcal{L}_{\text{img}}$	133.051	20.016	40.343	0.9884	173.263	26.459	38.783	0.9850
Log-ratio	$\mathcal{L}_{\text{img}} + \lambda_R \mathcal{L}_R$	**97.395**	**14.910**	**41.346**	0.9902	**126.234**	**19.845**	**40.126**	0.9870

Table 7: Quantitative comparison across different foreground-ratio ranges on iHarmony4. Results of the compared methods are taken from WAVSSNet (Zhou et al., 2026). The best and second-best results are marked in **bold** and underlined, respectively.

Method	0–5%			5–15%			15–100%		
	PSNR↑	MSE↓	fMSE↓	PSNR↑	MSE↓	fMSE↓	PSNR↑	MSE↓	fMSE↓
DoveNet (Cong et al., 2020)	38.37	14.34	580.88	33.95	42.82	476.10	29.24	147.10	474.73
RainNet (Ling et al., 2021)	39.49	12.17	557.29	34.76	34.61	397.42	29.56	128.11	420.34
Intrinsic (Guo et al., 2021b)	40.35	9.52	400.10	35.67	28.35	317.68	30.78	101.99	324.11
Harmonizer (Ke et al., 2022)	41.33	7.31	307.26	36.84	20.76	230.09	32.49	66.32	217.74
DCCF (Xue et al., 2022)	41.45	6.96	292.09	37.14	19.16	212.78	32.85	60.73	202.36
PCTNet(ViT) (Guerreiro et al., 2023)	42.32	5.56	240.74	38.53	14.71	164.55	34.30	47.59	155.85
WAVSSNet (Zhou et al., 2026)	43.46	5.29	248.99	39.23	13.44	155.95	34.70	48.08	158.41
Ratio-RFE	**45.31**	**3.21**	**146.08**	**40.24**	**10.43**	**121.07**	**35.06**	**42.78**	**137.44**

Table 8: Sensitivity analysis of the relighting range $r_{\max}$ using the basic log-ratio formulation without ADS-RFE pathways.

$r_{\max}$	fMSE↓	MSE↓	PSNR↑	SSIM↑
2	375.891	36.888	37.167	0.989
3	362.544	34.793	37.295	**0.990**
4	**310.570**	**29.499**	**37.982**	**0.990**
5	373.047	35.243	37.259	0.989
6	466.275	44.916	36.290	0.988

B Robustness across Foreground Ratios

Following the foreground-ratio evaluation protocol used in WAVSSNet (Zhou et al., 2026), the iHarmony4 test set is divided into three groups according to the foreground area ratio: 0–5%, 5–15%, and 15–100%. The results of the compared methods are taken from WAVSSNet (Zhou et al., 2026).

As shown in Table 7, Ratio-RFE achieves the best PSNR, MSE, and fMSE in all three foreground-ratio ranges. The consistent gains across small, medium, and large foreground regions indicate that the proposed source-relative formulation remains effective across substantially different foreground sizes.

C Sensitivity to the Relighting Range

The relighting range $r_{\max}$ controls the magnitude of the predicted log-ratio field. To isolate its effect from the ADS architecture, this analysis is conducted using the basic log-ratio formulation without the ADS-RFE pathways. All variants use the same training configuration, and the checkpoint for each setting is selected according to the same validation criterion.

Table 8 shows that the performance improves as $r_{\max}$ increases from 2 to 4, while further enlarging the prediction range degrades the overall results. A small $r_{\max}$ may restrict the required source-to-target correction, whereas an excessively large range makes stable relighting-field estimation more difficult. Based on this analysis, $r_{\max} = 4$ is used throughout the experiments.



Figure 4: Qualitative ablation of the Phong-inspired ADS pathways. From left to right, we show the composite input, the baseline without ADS pathways, variants using only the Ambient, Diffuse, or Specular pathway, the full ADS model, and the ground truth. Individual pathways provide complementary appearance corrections, while their joint use yields more consistent harmonization across diverse lighting and appearance variations.

D Qualitative Analysis of ADS Pathways

Qualitative pathway ablation. Figure 4 complements the quantitative ablation in Table 5. Each individual pathway improves the baseline in different cases, while the full ADS configuration produces more consistent harmonization across diverse lighting, color, and local appearance variations. These observations further support the complementary roles of the three pathways.

Feature visualization. Figure 5 visualizes the learned pathway representations for representative inputs. Distinct spatial response patterns can be observed across the Ambient, Diffuse, and Specular pathways, suggesting that they capture complementary appearance characteristics for relighting-field estimation. These pathway features are learned representations inspired by the Phong model rather than explicit physical decompositions of the Ambient, Diffuse, and Specular components.

E Additional Qualitative Results

Figures 6 and 7 provide additional qualitative comparisons on iHarmony4 and ccHarmony, respectively. For each example, an enlarged region corresponding to the marked box is shown below the full image to facilitate comparison of local color, appearance consistency, and foreground detail preservation.

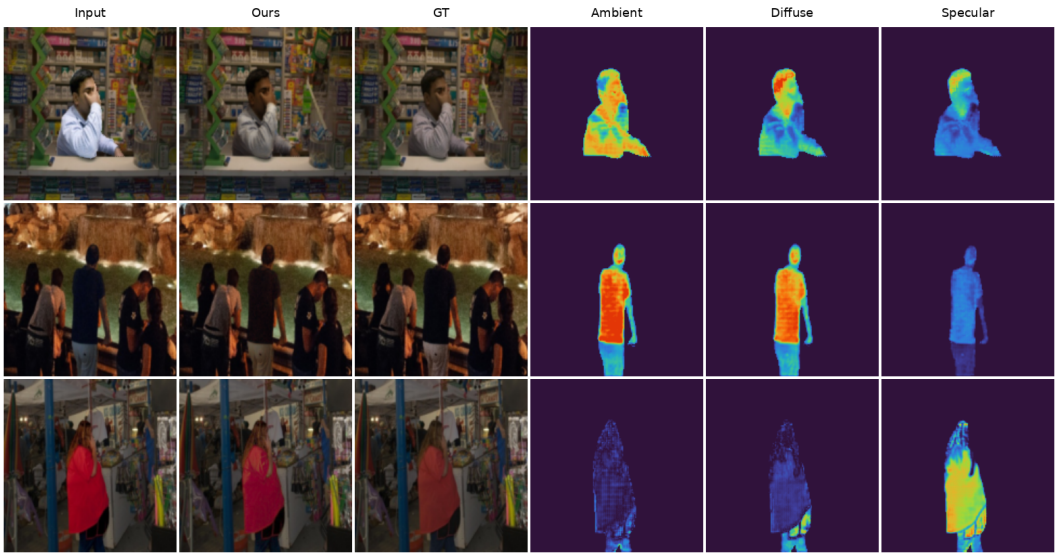

Figure 5: Feature visualizations of the Ambient, Diffuse, and Specular pathways on representative examples. Distinct response patterns are observed across the three pathways, indicating that they capture complementary appearance characteristics for relighting-field estimation.

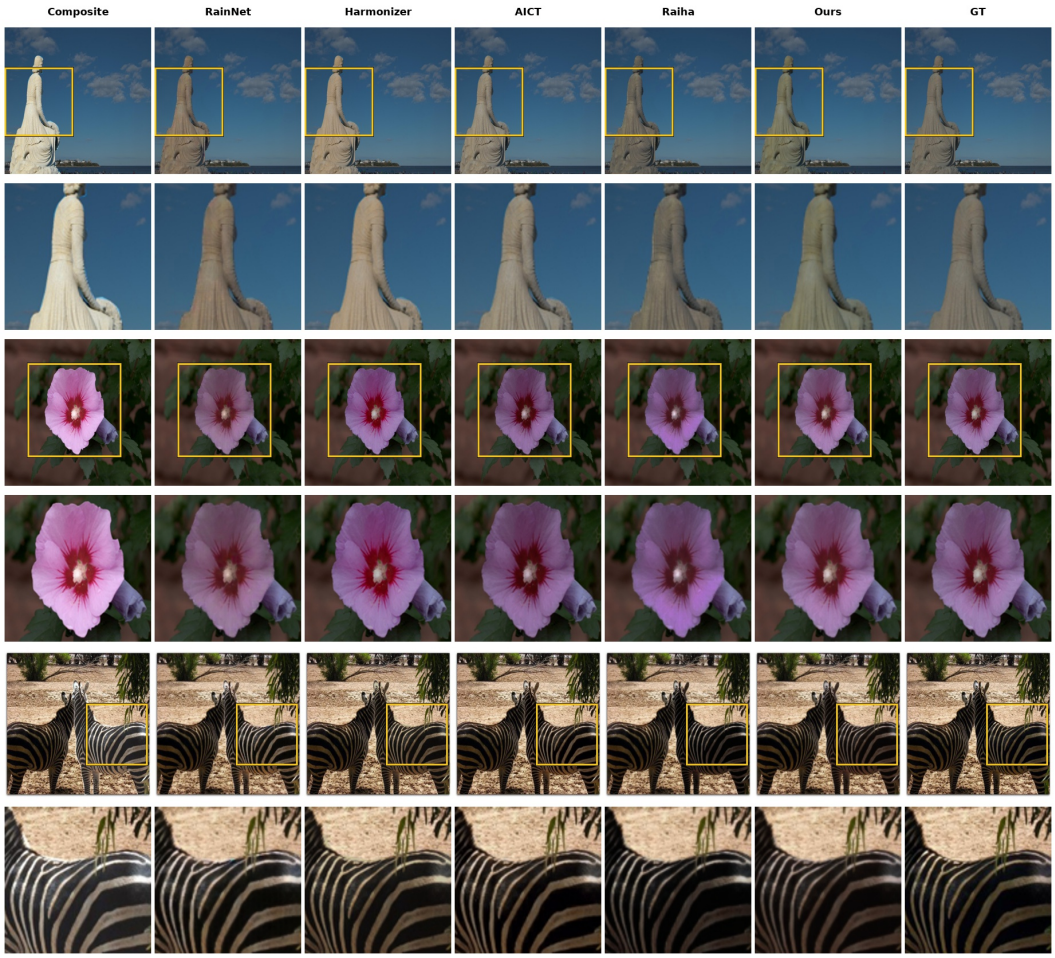

Figure 6: Additional qualitative comparisons on iHarmony4. For each example, the first row shows the full image, while the second row shows an enlarged region corresponding to the yellow box. The enlarged views provide a closer comparison of local color consistency, appearance correction, and foreground detail preservation.



Figure 7: Additional qualitative comparisons on ccHarmony. Ratio-RFE is directly evaluated using the model trained only on iHarmony4, without fine-tuning on ccHarmony. For each example, the first row shows the full image, while the second row shows an enlarged region corresponding to the yellow box. The enlarged views better reveal differences in local color and appearance consistency.