PhyFusion: Physics-Aware Infrared and Visible Image Fusion via Modality-Specific Physical Priors

A. Physical Prior Decomposition: Theory and Validation

In this section, we provide additional theoretical and empirical evidence for the proposed physical prior decomposition, including its infrared formulation, decomposition validity, and visible-domain analysis.

A.1. Theoretical Derivation of Infrared Priors

We first derive the infrared physical priors from the thermal imaging formation process and show how the proposed TeV-style [2] decomposition can be reformulated into an efficient learnable representation.

A typical infrared image captures the superposition of an object's self-emitted thermal radiation and the reflected radiation from its surroundings. Assuming a typical atmospheric transmissivity of $\tau_{\text{atm}} \approx 1$, the thermal radiation intensity S_λ at wavelength λ can be approximated as:

$$S_\lambda \approx e_\lambda B_\lambda(T) + (1 - e_\lambda)\Phi_{\text{env}}. \tag{1}$$

where e_λ is the object's emissivity, $B_\lambda(T)$ is the blackbody spectral irradiance defined by Planck's law [3], and Φ_{env} is the environmental irradiance. Please note that in the main paper, we use an uppercase $\mathbf{E}$ to denote the lowercase e used here.

To avoid the $O(N^2)$ complexity of modeling pixel-to-pixel environmental reflections, Φ_{env} is approximated using a linear combination of a downsampled, locally averaged infrared image $\hat{S}_\lambda$.

As illustrated in Fig. 1, empirical data indicates that the emissivity e of common materials remains approximately constant over the camera's working wavelength range $[\lambda_{\min}, \lambda_{\max}]$. By applying this constant-emissivity assumption and integrating the spectral signals over the entire wavelength band, the complex integrations can be simplified into matrix operations. This leads directly to the core modeling conclusion of the TeV decomposition method:

$$S = e \odot \mathbf{T} + (1 - e) \odot \mathbf{V}\hat{S} \tag{2}$$

where $\odot$ denotes element-wise multiplication. $e = [e_\alpha]_{\alpha=(x,y)} \in \mathbb{R}^{H\times W}$ is the emissivity matrix, $\mathbf{T} = [\mathbf{T}_\alpha]_{\alpha=(x,y)} \in \mathbb{R}^{H\times W}$ is the temperature matrix, and $\mathbf{V} = [\mathbf{V}_\alpha]_{\alpha=(x,y)} \in \mathbb{R}^{H\times W\times m}$ is the thermal vector matrix.

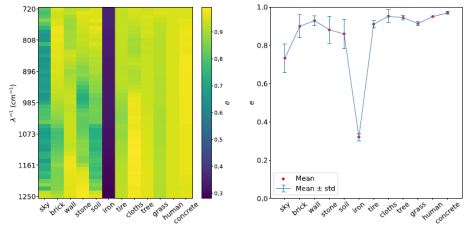

Figure 1. Common materials and their emissivities at different wavelengths. e represents emissivity, λ represents wavelength. HADAR [1] collects the emissivities of some of the common materials in the figure. Left: the emissivities of different materials at different wavelengths. Right: the statistical result of the emissivities of different materials. Both figures indicate that emissivity of a special materials changes little at different wavelengths.

Based on the derived relationship (Fig. 2), a self-supervised decomposition network is trained to adaptively predict (e, $\mathbf{T}$, $\mathbf{V}$) by optimizing $\mathbb{E}_{S\in p_{\text{data}}} \left\|\tilde{S} - S\right\|_2^2$:

$$\tilde{e}, \tilde{\mathbf{T}}, \tilde{\mathbf{V}} = C_{\text{ir}}(S), \tag{3}$$

where S is the ground-truth infrared image, and the reconstructed image $\tilde{S}$ is obtained via the reconstruction function Rec (Eq. (2)):

$$\tilde{S} = \text{Rec}(\tilde{e}, \tilde{\mathbf{T}}, \tilde{\mathbf{V}}). \tag{4}$$

A.2. Empirical Validation of Infrared Priors

We next validate whether the decomposed infrared priors are physically meaningful and beneficial for fusion by reconstruction analysis and controlled ablation experiments. we present the infrared image, the reconstructed infrared image, and (e, $\mathbf{T}$, $\mathbf{V}$) in Fig. 3.

First, we note that the physical priors obtained by decomposing an infrared image can be reconstructed. Second, although the reconstructed result exhibits no noticeable visual difference from the original image, a certain yet acceptable discrepancy is observed when we compute the

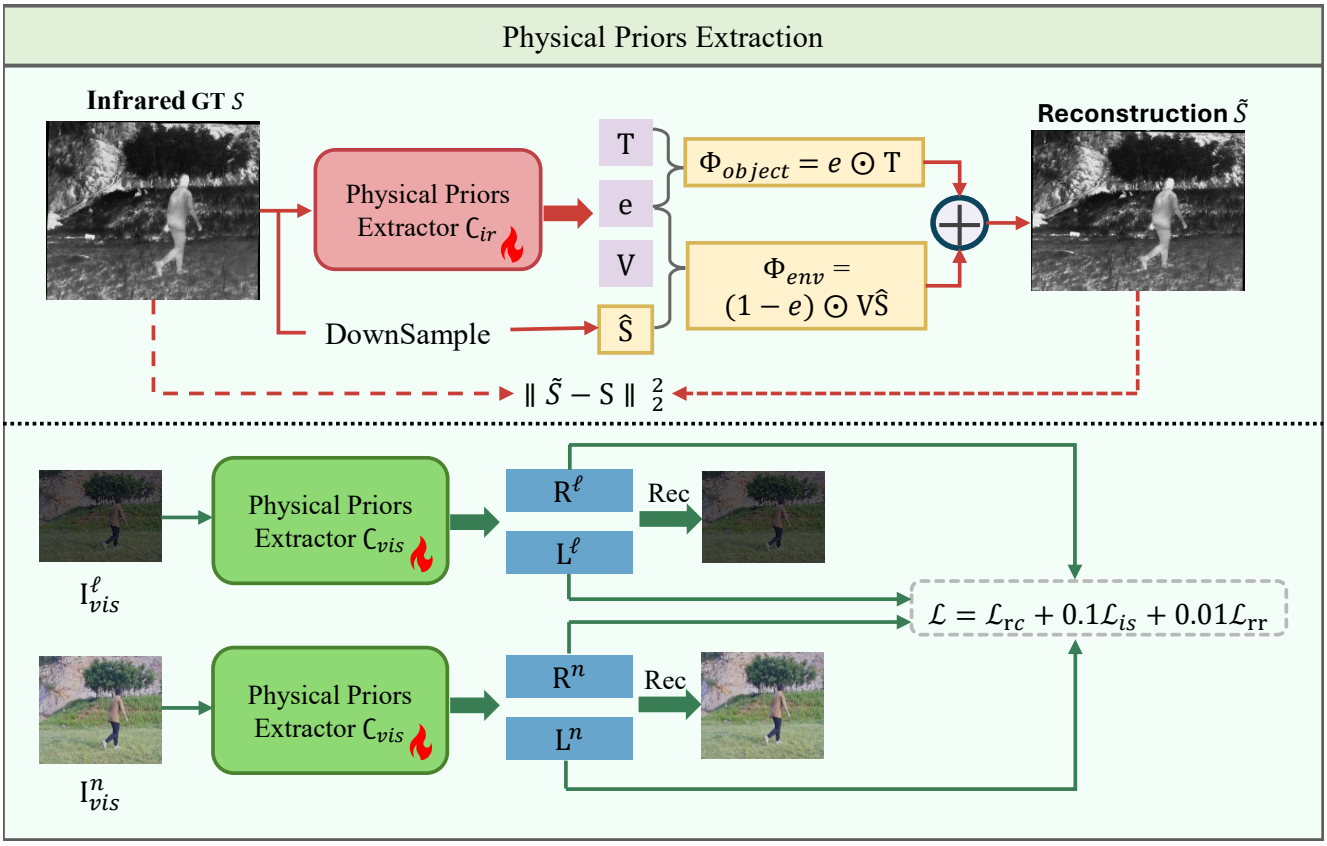

Figure 2. Training pipeline of the proposed physical prior extractors. The infrared and visible extractors are optimized through self-reconstruction and prior regularization to produce physically interpretable decompositions.

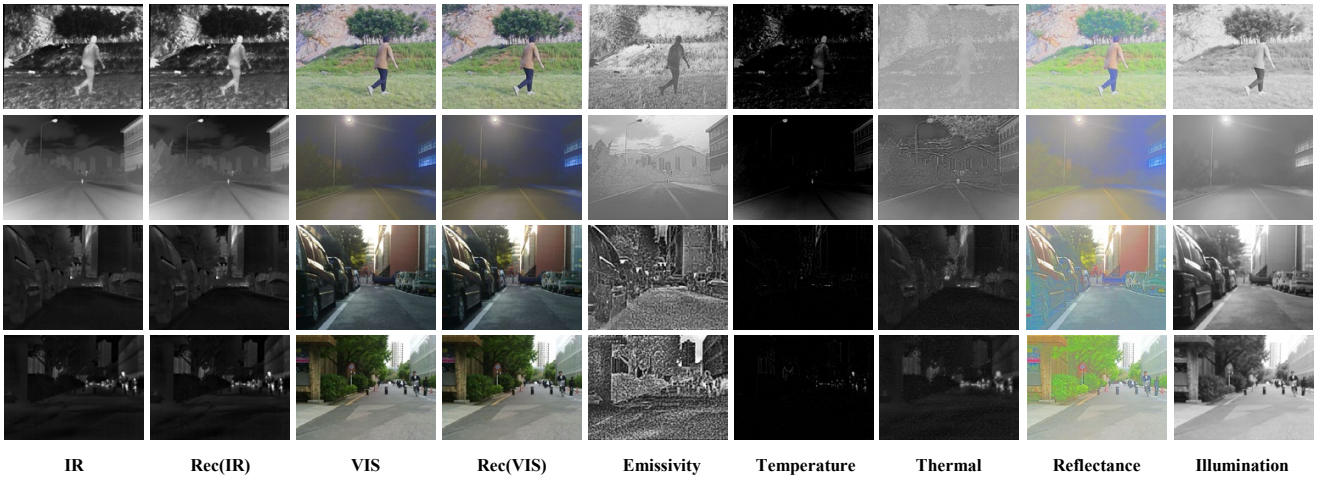

Figure 3. Visualization of decomposed physical priors. The extracted components reveal complementary physical attributes across infrared and visible modalities, including emissivity, temperature, thermal reflection, reflectance, and illumination.

relative reconstruction error (RelErr) and the gradient difference (gd). From the four example sets, we observe that Emissivity reflects the ability of a material to emit its thermal energy in the form of infrared radiation through variations in color intensity, while still preserving certain structural information. In contrast, Temperature only indicates the emitted radiation intensity corresponding to the object's temperature, and Thermal represents the ambient thermal

Model VIII Model IX Full Model

Figure 4. Ablation study on physical (e, $\mathbf{R}$) and $\mathcal{L}_{pc}$. Replacing e or removing $\mathcal{L}_{pc}$ leads to degraded structural fidelity and perceptual quality.

reflection. The characteristics of Emissivity make it particularly suitable for guiding the model to adaptively fuse infrared modality information.

Among the three physical priors, e describes the ability of an object surface to radiate thermal energy, serving as the primary source of thermal features in an infrared image and enhancing the saliency of thermal targets such as humans and vehicles. Meanwhile, in PhyFusion, it is precisely the non-explicit thermal feature representation of e that we leverage, allowing the fusion model to adaptively fuse thermal features based on the object surface's inherent ability to radiate thermal energy. To validate its effectiveness for PhyFusion, we replace e with $\mathbf{T}$ (Model IX), which represents the object's intrinsic thermal radiation intensity. The corresponding qualitative and quantitative results are shown in Fig. 4 and Tab 1, respectively. We observe that the fused images exhibit degradation in structural details, exposure quality, and quantitative metrics, indicating that the performance gains directly stem from physically meaningful priors rather than arbitrary decompositions.

Table 1. Ablation results of different physical prior configurations.

Model	EN	SF	AG	SD
Model VIII	7.212	10.311	2.631	58.847
Model IX	7.097	9.961	2.550	56.774
Full model	7.378	11.436	2.978	59.101

A.3. Visible Prior Decomposition Analysis

For the visible modality, we further analyze the decomposition quality under the Retinex assumption and examine whether the extracted reflectance and illumination improve detail preservation. To effectively decouple Reflectance and Illumination in the visible domain, we first perform low-light processing on 1,083 normal-light visible images from the MSRS dataset. Under Retinex theory, a low-light image and its normally illuminated counterpart can be regarded as observations sharing the same reflectance component but differing in illumination. Based on this assumption, we use normal-light images as references and simulate low-light imaging by attenuating their brightness, thereby constructing paired training samples.

As shown in Fig. 3, under the removal of illumination effects, the wrinkles of trousers (first group), building contours (second and third groups), and tree trunks (third and fourth groups) appear more clearly. Therefore, the details in an image that are affected by illumination become more clearly exposed, allowing the fusion model to focus more on preserving the details of the visible image under this condition. Note that when training the Physical Priors Extractor for Visible, we also require the decomposed physical quantities to be able to reconstruct the original image. Although the reconstruction results exhibit no noticeable visual difference from the original image, a certain yet acceptable deviation is observed when we compute the relative reconstruction error (RelErr) and the gradient difference (gd).

B. Why Physical Consistency Loss Works

To further explain the effectiveness of the proposed physical consistency loss, we provide both additional ablation evidence in the physical prior space and an analytical derivation. Through our ablation study (Model VIII), we further observe that our proposed $\mathcal{L}_{pc}$ and appearance-based loss functions are not mutually exclusive but complementary. When removing appearance-based loss functions (*e.g.*, $\mathcal{L}_{ssim}$and $\mathcal{L}_{int}$) and using $\mathcal{L}_{pc}$, PhyFusion yields physically plausible yet perceptually degraded results as shown in Fig. 4. Meanwhile, the metrics also exhibit a certain degree of degradation in Tab 1.

However, the quantitative and qualitative ablation analyses alone are insufficient to reveal the importance of $\mathcal{L}_{pc}$. We therefore provide a deeper theoretical analysis of $\mathcal{L}_{pc}$ in the following. Based on the infrared physical priors extractor $\mathbf{C}_{ir}$, the input infrared image $\mathbf{I}_{ir} \in \mathbb{R}^{H\times W\times C}$ can be decomposed into three physically interpretable components:

$$e_{ir}, \mathbf{T}_{ir}, \mathbf{V}_{ir} = C_{ir}(\mathbf{I}_{ir}), \tag{5}$$

Similarly, for the fused image $\mathbf{I}_f \in \mathbb{R}^{H\times W\times C}$, the same extractor is applied to obtain its corresponding physical decomposition in the infrared domain:

$$e_f, \mathbf{T}_f, \mathbf{V}_f = C_{ir}(\mathbf{I}_f), \tag{6}$$

Motivated by Eq. (2), we require that the fused image should be consistent with the source infrared image in the physical decomposition space, rather than merely at the pixel level. To this end, the physical consistency loss is defined as the mean squared error between the physical priors extracted from the fused image and those extracted from the infrared image:

$$\mathcal{L}_{pc} = \mathcal{L}_{MSE}\big(\mathbf{C}_{ir}(\mathbf{I}_f), \mathbf{C}_{ir}(\mathbf{I}_{ir})\big), \tag{7}$$

By substituting the outputs of $C_{ir}(\cdot)$ into the above expression, $\mathcal{L}_{pc}$ can be further written as:

$$L_{pc} = L_{MSE}((e_f, \mathbf{T}_f, \mathbf{V}_f), (e_{ir}, \mathbf{T}_{ir}, \mathbf{V}_{ir})), \tag{8}$$

Table 2. Quantitative comparison of object detection on the LLVIP dataset. Red indicates the best result, and Blue represents the second-best results.

Method	mAP@0.5	mAP@0.75	mAP@0.5:0.95
TarDAL	0.685	0.432	0.404
SHIP	0.742	0.467	0.442
Text-IF	0.751	0.475	0.446
SAGE	0.668	0.377	0.325
KDFuse	0.723	0.453	0.451
DCEvo	0.771	0.486	0.457
HCLFuse	0.726	0.459	0.420
Ours	**0.777**	**0.514**	**0.463**

Figure 5. Additional qualitative comparison on downstream object detection over LLVIP. The proposed PhyFusion achieves more reliable pedestrian detection under occlusion and low-visibility conditions.

Using the standard form of the mean squared error, we obtain

$$L_{pc} = \frac{1}{N}\|C_{ir}(\mathbf{I}_f) - C_{ir}(\mathbf{I}_{ir})\|_2^2, \tag{9}$$

where N denotes the total number of compared elements. Expanding the above equation in terms of the three physical components yields

$$L_{pc} = \frac{1}{N}\left(\|e_f - e_{ir}\|_2^2 + \|\mathbf{T}_f - \mathbf{T}_{ir}\|_2^2 + \|\mathbf{V}_f - \mathbf{V}_{ir}\|_2^2\right). \tag{10}$$

The fused output is explicitly constrained to preserve the emissivity distribution, temperature-related structure, and thermal radiation characteristics of the infrared modality in the learned physical space. In this way, L_{pc} further enforces radiative coherence during reconstruction and encourages the generated fused image to comply with the physical principles underlying infrared imaging.

C. Additional Object Detection Results on LLVIP

Beyond image-level evaluation, we further assess PhyFusion on the downstream object detection task to verify whether the physically grounded fusion improves high-level scene understanding. we fine-tuned a YOLOv5 [4] detector on the infrared and visible images of the LLVIP dataset. The qualitative and quantitative results are shown in Fig. 5 and Tab 2.

As shown in the Fig. 5, our detection results surpass those of the other methods, achieving the highest confidence score, and ours is the only approach that successfully detects, with the highest confidence, the person occluded by the three-wheeled vehicle in the center of the scene. Quantitatively, PhyFusion also achieves the best detection performance. Taken together, these qualitative and quantitative comparisons confirm that the proposed PhyFusion method delivers superior performance in the object-detection task.

References

[1] Fanglin Bao, Xueji Wang, Shree Hari Sureshbabu, Gautam Sreekumar, Liping Yang, Vaneet Aggarwal, Vishnu N Boddeti, and Zubin Jacob. Heat-assisted detection and ranging. *Nature*, 2023. 1

[2] Fangyuan Mao, Jilin Mei, Shun Lu, Fuyang Liu, Liang Chen, Fangzhou Zhao, and Yu Hu. Pid: physics-informed diffusion model for infrared image generation. *Pattern Recognition*, 2025. 1

[3] Max Planck. The theory of heat radiation. *Entropie*, 144(190): 164, 1900. 1

[4] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 779–788, 2016. 4