More Than Meets the Eye: A Unified Image Fusion Framework via Semantic-Pixel Entropy Trade-off for Zero-Shot Generalization

Supplementary Material

1. Mathematical Derivation of Dual-Entropy Optimization Objective

In this section, we analyze the mathematical derivation from the free energy principle to the dual-entropy expectation optimization objective. The variational free energy aims to establish a unified optimization framework that balances semantic fidelity and pixel-level richness in image fusion tasks. The formal definition is given by:

$$F(\tilde{s}(\alpha),\mu)=\mathbb{E}_{q(\mu)}[-\log p(\tilde{s}(\alpha)|\mu)]+\mathrm{KL}[q(\mu)\|p(\mu)] \tag{1}$$

where:

- $\tilde{s}(\alpha)$ represents sensory inputs dependent on action α, corresponding to pixel-level images I_{pix} in the fusion context
- μ denotes internal states, corresponding to semantic representations in the fusion framework
- $q(\mu)$ is the variational posterior distribution over μ
- $p(\mu)$ is the prior distribution over μ
- The first term represents the prediction error (expectation of negative log-likelihood)
- The second term is the complexity term (KL divergence between posterior and prior)

1.1. Bayesian Decomposition of Prediction Error

We begin by applying Bayes' theorem to the prediction error term. The posterior distribution $p(\mu|\tilde{s}(\alpha))$ relates to the likelihood $p(\tilde{s}(\alpha)|\mu)$ and prior $p(\mu)$ through:

$$p(\mu|\tilde{s}(\alpha))=\frac{p(\tilde{s}(\alpha)|\mu)p(\mu)}{p(\tilde{s}(\alpha))} \tag{2}$$

Taking the negative logarithm of both sides:

$$-\log p(\tilde{s}(\alpha)|\mu)=-\log p(\mu|\tilde{s}(\alpha))-\log p(\tilde{s}(\alpha))+\log p(\mu) \tag{3}$$

Computing the expectation with respect to the variational distribution $q(\mu)$:

$$\begin{aligned}\mathbb{E}_{q(\mu)}[-\log p(\tilde{s}(\alpha)|\mu)]=\mathbb{E}_{q(\mu)}[-\log p(\mu|\tilde{s}(\alpha))]\\ -\log p(\tilde{s}(\alpha))+\mathbb{E}_{q(\mu)}[\log p(\mu)]\end{aligned} \tag{4}$$

Note that $p(\tilde{s}(\alpha))$ is independent of μ and is treated as constant in the expectation operation.

1.2. Reformulation of Variational Free Energy

Substituting Equation 4 into Equation 1 yields:

$$\begin{aligned}F=\mathbb{E}_{q(\mu)}[-\log p(\mu|\tilde{s}(\alpha))]-\log p(\tilde{s}(\alpha))\\ +\mathbb{E}_{q(\mu)}[\log p(\mu)]+\mathrm{KL}[q(\mu)\|p(\mu)]\end{aligned} \tag{5}$$

Expanding the KL divergence term:

$$\mathrm{KL}[q(\mu)\|p(\mu)]=\mathbb{E}_{q(\mu)}[\log q(\mu)-\log p(\mu)] \tag{6}$$

Substituting Equation 6 into Equation 5:

$$\begin{aligned}F=\mathbb{E}_{q(\mu)}[-\log p(\mu|\tilde{s}(\alpha))]-\log p(\tilde{s}(\alpha))\\ +\mathbb{E}_{q(\mu)}[\log p(\mu)]+\mathbb{E}_{q(\mu)}[\log q(\mu)-\log p(\mu)]=\\ \mathbb{E}_{q(\mu)}[-\log p(\mu|\tilde{s}(\alpha))]-\log p(\tilde{s}(\alpha))+\mathbb{E}_{q(\mu)}[\log q(\mu)]\end{aligned} \tag{7}$$

The terms $\mathbb{E}_{q(\mu)}[\log p(\mu)]$ and $-\mathbb{E}_{q(\mu)}[\log p(\mu)]$ cancel each other during the simplification process.

Recognizing that $\mathbb{E}_{q(\mu)}[\log q(\mu)]=-H(q(\mu))$, where $H(q(\mu))$ denotes the entropy of distribution $q(\mu)$:

$$F=-\log p(\tilde{s}(\alpha))+\mathbb{E}_{q(\mu)}[-\log p(\mu|\tilde{s}(\alpha))]-H(q(\mu)) \tag{8}$$

1.3. KL Divergence and Conditional Entropy Relationship

The combination $\mathbb{E}_{q(\mu)}[-\log p(\mu|\tilde{s}(\alpha))]-H(q(\mu))$ corresponds precisely to the KL divergence between $q(\mu)$ and the posterior distribution $p(\mu|\tilde{s}(\alpha))$:

$$\begin{aligned}\mathrm{KL}[q(\mu)\|p(\mu|\tilde{s}(\alpha))]&=\mathbb{E}_{q(\mu)}[\log q(\mu)-\log p(\mu|\tilde{s}(\alpha))]\\ &=-H(q(\mu))-\mathbb{E}_{q(\mu)}[\log p(\mu|\tilde{s}(\alpha))]\end{aligned} \tag{9}$$

Therefore:

$$\mathbb{E}_{q(\mu)}[-\log p(\mu|\tilde{s}(\alpha))]-H(q(\mu))=\mathrm{KL}[q(\mu)\|p(\mu|\tilde{s}(\alpha))] \tag{10}$$

Substituting Equation 10 into Equation 8:

$$F=-\log p(\tilde{s}(\alpha))+\mathrm{KL}[q(\mu)\|p(\mu|\tilde{s}(\alpha))] \tag{11}$$

In the context of image fusion, the sensory input $\tilde{s}(\alpha)$ corresponds to the pixel image I_{pix} generated by action α, leading to:

$$F = -\log p(I_{\text{pix}}) + \text{KL}[q(\mu)\|p(\mu|I_{\text{pix}})] \tag{12}$$

1.4. Semantic and Pixel Entropy Approximation

From an information-theoretic perspective, the entropy of pixel information $H(I_{\text{pix}})$ is defined as $H(I_{\text{pix}}) = \mathbb{E}[-\log p(I_{\text{pix}})]$, where the expectation is taken over the distribution of pixel information. Consequently, the term $-\log p(I_{\text{pix}})$ represents the self-information of the pixel configuration. Maximizing the expected self-information corresponds to maximizing the entropy $H(I_{\text{pix}})$, which indicates richer details.

In our image fusion framework, the reconstruction path aims to maximize pixel entropy $\mathbb{E}[H(I_{\text{pix}}|\alpha)]$, where α controls pixel generation. We therefore establish the approximation:

$$-\log p(I_{\text{pix}}) \propto -\mathbb{E}[H(I_{\text{pix}}|\alpha)] \tag{13}$$

Here, the expectation $\mathbb{E}[\cdot]$ is taken over both the action α and the data distribution. The minimization of $-\log p(I_{\text{pix}})$ corresponds to maximizing pixel diversity, aligning with the objective of the reconstruction path.

The term $\text{KL}[q(\mu)\|p(\mu|I_{\text{pix}})]$ quantifies the discrepancy between the variational posterior $q(\mu)$ and the true posterior $p(\mu|I_{\text{pix}})$, which relates directly to semantic uncertainty. Higher semantic uncertainty results in increased KL divergence. The image fusion task aims to minimize semantic entropy $\mathbb{E}[H(I_{\text{sem}}|\mu)]$, where $H(I_{\text{sem}}|\mu)$ represents the conditional entropy of semantic information given μ. When $q(\mu)$ accurately captures the semantic state, the KL divergence decreases, indicating reduced semantic entropy. We thus formulate the approximation:

$$\text{KL}[q(\mu)\|p(\mu|I_{\text{pix}})] \propto \mathbb{E}[H(I_{\text{sem}}|\mu)] \tag{14}$$

The expectation $\mathbb{E}[\cdot]$ in this context is taken over both the variational distribution $q(\mu)$ and the semantic data distribution.

1.5. Dual-Entropy Optimization Objective

Combining Equations 13 and 14, the free energy F can be approximated as:

$$F \propto \mathbb{E}[H(I_{\text{sem}}|\mu)] - \mathbb{E}[H(I_{\text{pix}}|\alpha)] + C \tag{15}$$

where C represents constant terms incorporating contributions from prior distributions, approximation errors, and the marginal probability $p(I_{\text{pix}})$.

Based on this mathematical analysis, the perception-generation symmetry inherent in the free energy principle can be effectively translated into a dual-entropy collaborative optimization framework. This formulation provides a principled foundation for balancing semantic certainty and pixel diversity in unified image fusion, enabling zero-shot generalization across diverse fusion tasks while maintaining robust visual-semantic integrity.

2. Details of Network Architecture and Training Strategy

The Dual-Entropy Collaborative Constraints (DECC) network proposed in this paper is a unified image fusion framework designed for cross-task zero-shot generalization. At its core, the network features a symmetric architecture comprising the Perception Path (PP) and the Reconstruction Path (RP). This design decouples semantic and pixel-level features by integrating a frequency dynamic encoder. Furthermore, it effectively achieves the primary optimization objectives—semantic entropy minimization and pixel entropy maximization—through a dual-path alternating training mechanism. To clearly explain the design concept and implementation details of the proposed method, this subsection provides a detailed analysis and illustration of both the frequency dynamic encoder and the dual-path alternating training mechanism.

2.1. Frequency Dynamic Encoder

Semantic features are primarily concentrated in abstract, low-frequency information, whereas pixel features tend to contain more complex, high-frequency signals. These two types of features exhibit distinct distributions in the frequency domain. To enable the model to simultaneously capture robust semantic and pixel features, we design an encoder based on frequency-domain dynamic perception. This encoder performs frequency-aware decomposition, cross-modal fusion, and cross-frequency fusion of input images, serving as the core shared module for the dual paths.

2.1.1. Frequency-Aware Decomposition

The network design for frequency-aware decomposition is illustrated in Figure 1. In this module, the input infrared and visible images are each decomposed into three frequency-aware feature groups. These groups correspond to different fusion information and enable frequency-domain decoupling of semantic and pixel features. The semantic-oriented group focuses on the global structure and semantic information of images, providing semantic representations for the perception path. Its relevant parameters are optimized exclusively during the training of the perception path. The pixel-oriented group focuses on the edges, textures, and fine details of images, providing pixel-level representations for the reconstruction path. Its relevant parameters are optimized exclusively during the training of the reconstruction

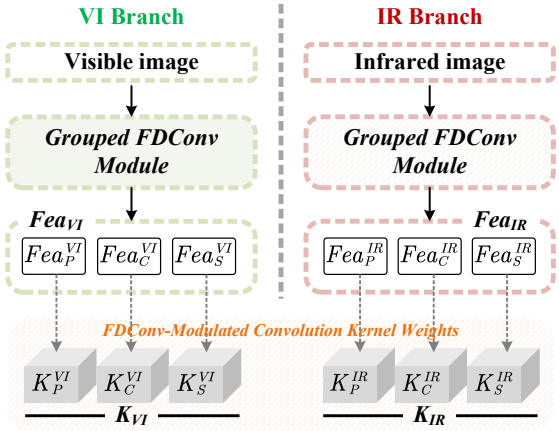

Figure 1. The specific network architecture of frequency-aware decomposition.

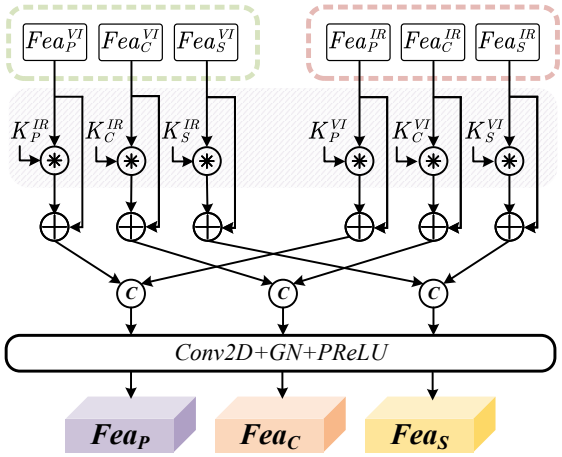

Figure 2. The specific network architecture of cross-modal fusion. ⊛ denotes the convolution operation.

path. The common-oriented group focuses on the transition frequency bands between semantic and pixel features, ensuring feature correlation across paths and maintaining the integrity of information throughout the full frequency band. In addition, to achieve interactive fusion of cross-modal features in the next stage, we adopt the design of Frequency Dynamic Convolution (FDConv). This approach uses infrared and visible features to weight and modulate two sets of convolution kernels, thereby obtaining infrared-modulated and visible-modulated convolution parameters.

2.1.2. Cross-Modal Fusion

After frequency-aware decomposition, infrared and visible images are separated into features corresponding to different frequency bands. To achieve cross-modal fusion, we employ a cross-modal frequency interactive perception approach, enabling interactive integration of features

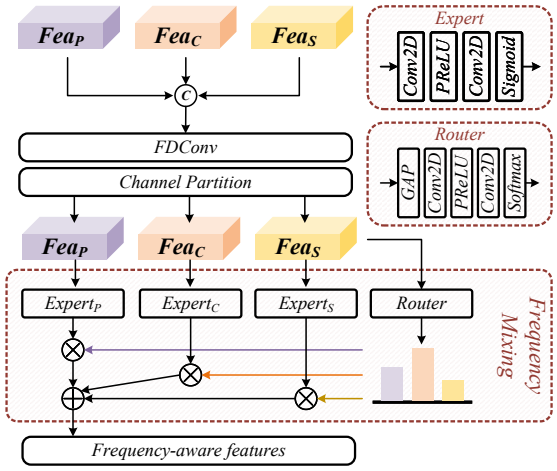

Figure 3. The specific network architecture of cross-frequency fusion. The GAP denotes global average pooling operation.

from both modalities within the dynamic frequency domain. The specific design is illustrated in Figure 2: the three groups of visible features are each convolved with infrared-modulated convolution parameters, while simultaneously, the three groups of infrared features are convolved with visible-modulated convolution parameters. The original features are then added to the resulting outputs to preserve information integrity. Subsequently, features from the same frequency band across different modalities are concatenated and passed through a convolutional block for fusion, producing three groups of cross-modally fused features: Fea_S (semantic-oriented), Fea_P (pixel-oriented), and Fea_C (common-oriented).

2.1.3. Cross-Frequency Fusion

To enhance the extracted features with greater frequency adaptability, we perform additional cross-frequency dynamic integration on the three groups of features extracted during the cross-modal fusion stage. As illustrated in Figure 3, these three feature groups are first concatenated and fed into the FDConv module for frequency-domain dynamic perception. Subsequently, the initially combined features are split and passed through a frequency mixing module based on Mixture of Experts (MoE) attention. This process further weights the information from different frequency bands, producing the final frequency-aware features.

2.2. Dual-Path Alternating Training Mechanism

Since semantic information and pixel details in image fusion tasks exhibit fundamentally different feature distributions and tend to interfere with each other during optimization, a single training path struggles to balance high-level semantic fidelity with low-level visual richness. To address

this challenge, we propose a dual-path alternating training mechanism. The perception path utilizes object detection supervision to minimize semantic entropy and enhance semantic representation, while freezing pixel-related parameters in the encoder to prevent interference from pixel-level features. Conversely, the reconstruction path employs random masking to reconstruct images, maximizing pixel entropy to improve detail reconstruction, while freezing semantic-related parameters in the encoder to avoid excessive constraints from semantic information. Both paths share the same encoder, achieving feature decoupling through alternating training. Additionally, Elastic Weight Consolidation (EWC) regularization is introduced to mitigate catastrophic forgetting, enabling collaborative optimization between semantic and pixel features, thereby establishing a foundation for zero-shot generalization across various fusion tasks. The specific process of this mechanism is shown in Algorithm 1.

3. Loss Function

During training, the total loss for each path consists of four components: a supervised base loss ($\mathscr{L}_{\text{base}}$), an unsupervised information expectation loss ($\mathscr{L}_{\text{info}}$), a Kullback-Leibler divergence-based trade-off loss ($\mathscr{L}_{\text{trade}}$), and an Elastic Weight Consolidation (EWC) loss ($\mathscr{L}_{\text{ewc}}$) to prevent catastrophic forgetting during iterative optimization.

The overall loss for each path is defined as:

$$\mathscr{L}_{\text{path}} = \mathscr{L}_{\text{base}} + \alpha\mathscr{L}_{\text{info}}^{\text{path}} + \beta\mathscr{L}_{\text{trade}}^{\text{path}} + \mathscr{L}_{\text{ewc}}, \tag{16}$$

where path $\in \{\text{PP}, \text{RP}\}$. α and β are adjustable parameters.

The base loss is composed of intensity and gradient terms:

$$\mathscr{L}_{\text{base}} = \mathscr{L}_{\text{int}} + \gamma\mathscr{L}_{\text{grad}}. \tag{17}$$

where γ is an adjustable parameter. The intensity loss $\mathscr{L}_{\text{int}}$ enforces content consistency between the fused image and source images:

$$\mathscr{L}_{\text{int}} = \mathscr{L}_{\text{MSE}}(x_f, x_1) + \mathscr{L}_{\text{MSE}}(x_f, x_2), \tag{18}$$

where $\mathscr{L}_{\text{MSE}}$ is the Mean Squared Error loss defined as:

$$\mathscr{L}_{\text{MSE}}(A,B) = \frac{1}{N}\sum_{i=1}^{N}(A_i - B_i)^2, \tag{19}$$

with N being the number of pixels.

The gradient loss $\mathscr{L}_{\text{grad}}$ constrains gradient consistency:

$$\mathscr{L}_{\text{grad}} = \mathscr{L}_1(\nabla x_f, \max(\nabla x_1, \nabla x_2)), \tag{20}$$

where ∇ denotes the Sobel gradient operator. $\mathscr{L}_1$ is the L1 norm loss, and its specific formula is given by:

$$\mathscr{L}_1(A,B) = \sum_{i=1}^{N}|A_i - B_i|, \tag{21}$$

Algorithm 1 Dual-Path Alternating Training Mechanism of DECC

Require:
Training dataset $\mathscr{D}$; Total epochs E; Frequency dynamic encoder $\mathscr{E}$ with cross-path sharing parameter θ; Perception head $\mathscr{H}_{PP}$ with parameter ϕ; Reconstruction head $\mathscr{H}_{RP}$ with parameter ψ; Fusion head $\mathscr{H}_{FU}$ with parameter η;
Base loss ($\mathscr{L}_{\text{base}}$); Unsupervised information expectation loss ($\mathscr{L}_{\text{info}}$); Kullback-Leibler divergence-based trade-off loss ($\mathscr{L}_{\text{trade}}$); Elastic Weight Consolidation loss ($\mathscr{L}_{\text{ewc}}$); Optimizer $\mathscr{O}$.
Ensure: Well-trained DECC network

Initialize parameters of θ, ϕ, ψ, and η
for epoch $= 1$ to E **do**
  Stage 1: Perception Path (PP) Training
  Freeze pixel-oriented parameters of $\mathscr{E}$
  for each batch $(x_1, x_2) \sim \mathscr{D}$ **do**
    Extract features:
    $\{Fea_S, Fea_P, Fea_C\} \leftarrow \mathscr{E}(x_1, x_2; \theta)$
    Predict detection results and fused image:
    $\hat{y} \leftarrow \mathscr{H}_{PP}(Fea_S, Fea_P, Fea_C; \phi)$
    $\hat{x_f} \leftarrow \mathscr{H}_{FU}(Fea_S, Fea_P, Fea_C; \eta)$
    Compute perception path loss:
    $\mathscr{L}_{pp} = \mathscr{L}_{base} + \alpha\mathscr{L}_{info}^{PP} + \beta\mathscr{L}_{trade}^{PP} + \lambda\mathscr{L}_{ewc}$
    Update θ, ϕ and η via $\mathscr{O}(\nabla\mathscr{L}_{pp})$
  end for
  Update EWC parameters and importance weights
  Stage 2: Reconstruction Path (RP) Training
  Freeze semantic-oriented parameters of $\mathscr{E}$
  for each batch $(x_1, x_2) \sim \mathscr{D}$ **do**
    Generate masked images $\tilde{x}_1, \tilde{x}_2$
    Extract features:
    $\{Fea_S, Fea_P, Fea_C\} \leftarrow \mathscr{E}(\tilde{x}_1, \tilde{x}_2; \theta)$
    Predict reconstruction results and fused image:
    $\hat{x}_{rec} \leftarrow \mathscr{H}_{RP}(Fea_S, Fea_P, Fea_C; \psi)$
    $\hat{x_f} \leftarrow \mathscr{H}_{FU}(Fea_S, Fea_P, Fea_C; \eta)$
    Compute reconstruction path loss:
    $\mathscr{L}_{rp} = \mathscr{L}_{base} + \alpha\mathscr{L}_{info}^{RP} + \beta\mathscr{L}_{trade}^{RP} + \lambda\mathscr{L}_{ewc}$
    Update θ, ψ and η via $\mathscr{O}(\nabla\mathscr{L}_{rp})$
  end for
  Update EWC optimal parameters and importance weights
end for
return Trained $\mathscr{E}$, $\mathscr{H}_{PP}$, $\mathscr{H}_{RP}$, and $\mathscr{H}_{FU}$

where $|.|$ denotes the absolute value operation.

Perception Path: The information expectation loss minimizes semantic entropy through cross-entropy:

$$\mathscr{L}_{\text{info}}^{\text{PP}} = -\sum_i y_i \log \hat{y}_i, \tag{22}$$

where y_i is the ground-truth label and $\hat{y}_i$ is the predicted

Figure 4. The training data samples for multi-exposure fusion, and the reference images are derived from the visible images in the M3FD dataset.

class probability. This aligns the detection output with semantic labels.

The trade-off loss enforces consistency between the fused result and perceptual features:

$$\mathscr{L}_{\text{trade}}^{\text{PP}} = \mathscr{L}_{\text{KL}}(x_f, Fea_{\text{per}}), \tag{23}$$

where Fea_{per} denotes features fed into the perception head, and $\mathscr{L}_{\text{KL}}$ is the Kullback-Leibler divergence defined as:

$$\mathscr{L}_{\text{KL}}(P,Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}. \tag{24}$$

Reconstruction Path: The information expectation loss encourages accurate pixel-level reconstruction:

$$\mathscr{L}_{\text{info}}^{\text{RP}} = \frac{1}{N} \sum_j \|\hat{I}_j - I_j\|_2^2, \tag{25}$$

where I_j is the original pixel value in the occluded region, $\hat{I}_j$ is the reconstructed value, and N is the total number of occluded pixels.

The corresponding trade-off loss constraints the fused result to match the high-dimensional pixel distribution:

$$\mathscr{L}_{\text{trade}}^{\text{RP}} = \mathscr{L}_{\text{KL}}(x_f, Fea_{\text{rec}}), \tag{26}$$

where Fea_{rec} represents features input to the reconstruction head.

EWC Regularization: The EWC term helps preserve important parameters from previous tasks:

$$\mathscr{L}_{\text{EWC}} = \frac{1}{2} \sum_i \lambda_i (\theta_i - \theta_i^*)^2, \tag{27}$$

where θ_i is the current parameter, θ_i^* is its value from the previous path, and λ_i is the importance weight. This regularization enhances stability and mitigates catastrophic forgetting in dual-path iterative learning.

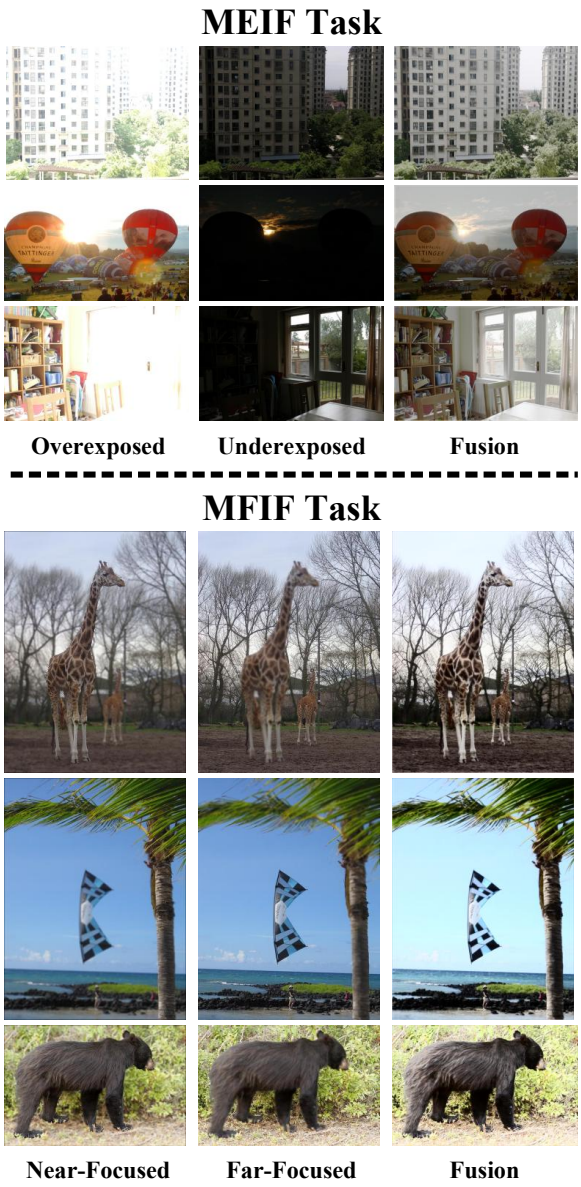

Figure 5. Fusion results of the Multi-Exposure and Multi-Focus Image Fusion task.

4. Zero-shot Generalization Validation via Digital Photography Fusion Training

To further validate the zero-shot generalization capability of DECC, we conducted experiments where the model is trained on a multi-exposure task and generalized zero-shot to other tasks. RGB images from the M3FD dataset were processed to generate underexposure and overexposure image pairs, which served as the training data for the model. Figure 4 shows a set of reference images, along with the corresponding processed underexposure and overexposure

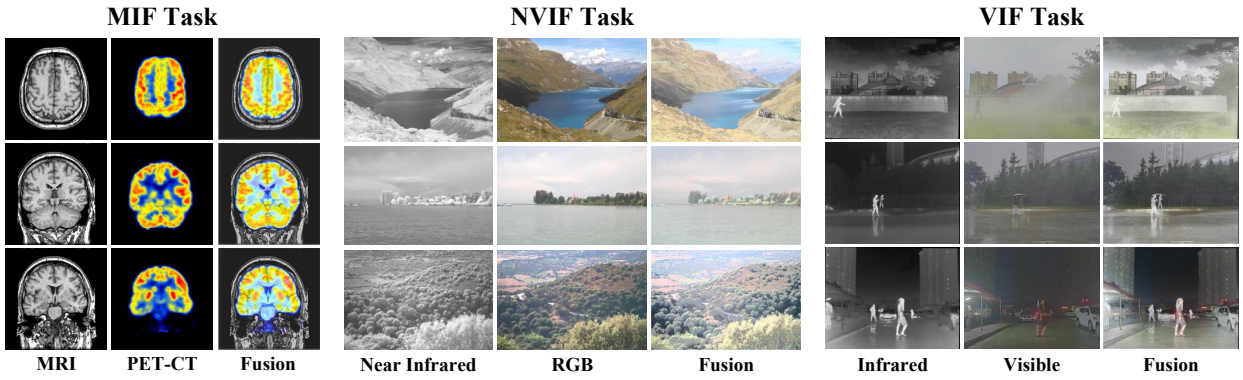

Figure 6. Visual results of zero-shot generalization on cross-modal fusion tasks

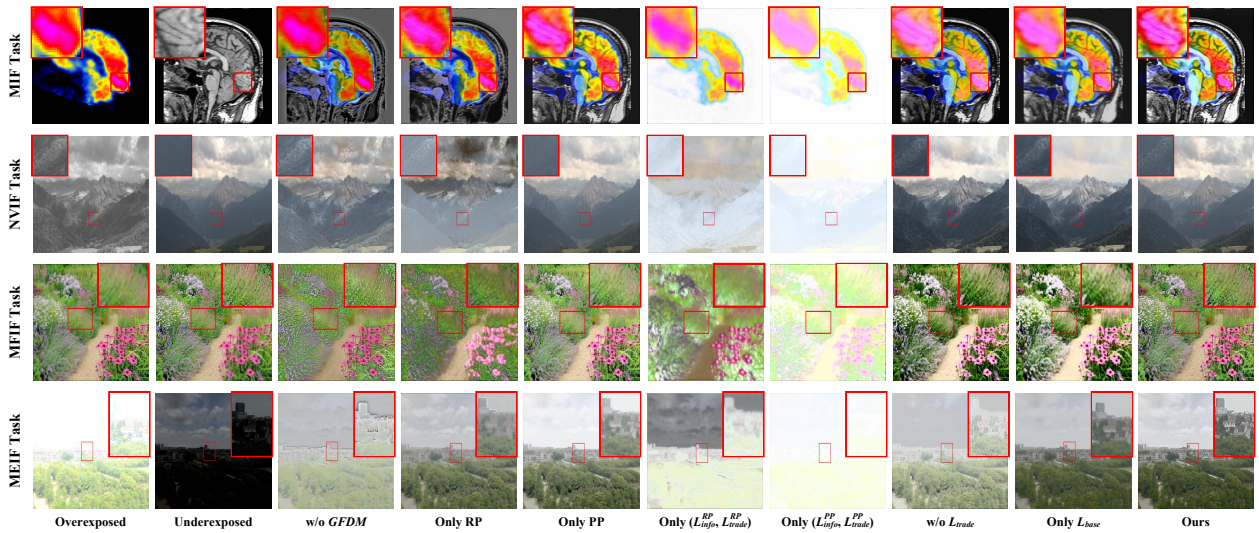

Figure 7. Comparison of fusion results under different ablation settings, including the perception-only path, reconstruction-only path, and without trade-off loss, to verify the impact of each component on zero-shot generalization performance.

images. The base loss design is adjusted as follows:

$$\mathscr{L}_{\text{base}} = \mathscr{L}_{\text{int}}^{\text{MF}} + \mathscr{L}_{\text{grad}}. \tag{28}$$

The intensity loss $\mathscr{L}_{\text{int}}^{\text{MF}}$ enforces content consistency between the fused image x_f and the reference image x_r:

$$\mathscr{L}_{\text{int}}^{\text{MF}} = \mathscr{L}_{\text{MSE}}(x_f, x_r), \tag{29}$$

where $\mathscr{L}_{\text{MSE}}$ is the Mean Squared Error loss defined as:

$$\mathscr{L}_{\text{MSE}}(A,B) = \frac{1}{N}\sum_{i=1}^{N}(A_i - B_i)^2, \tag{30}$$

with N being the number of pixels. All other training details, including the dual-path optimization and frequency-aware feature decoupling, remain consistent with the main paper.

As illustrated in Figure 5, the model successfully generalizes to both multi-exposure and multi-focus fusion tasks, despite being trained only on exposure-adjusted data. The fused results exhibit well-balanced brightness, preserved structural details, and natural texture rendering. The results presented in Figure 6 provide further validation of the model's robust generalization capabilities across a diverse set of cross-modal fusion tasks. It can be observed that our method successfully preserves critical features from all input sources while generating a fused image of high visual fidelity and informational integrity. These results demonstrate that the dual-entropy trade-off strategy enables the model to learn a task-agnostic fusion principle that transfers effectively across domains.

5. Extended Ablation Study on Zero-shot Generalization

To further validate the contribution of each component in zero-shot generalization, we conduct extended ablation

studies on unseen tasks including near-infrared and visible image fusion (NVIF), medical image fusion (MIF), multi-focus image fusion (MFIF) and multi-exposure image fusion (MEIF).

As shown in Figure 7, the results demonstrate that the model trained solely with the perception path (*only PP*) produces outputs with clearer contour structures but exhibits noticeable deficiencies in texture details. In contrast, the model trained exclusively with the reconstruction path (*Only RP*) achieves richer detail reconstruction but shows degradation in overall structural integrity. The variant trained with only the combination of information expectation loss and trade-off loss (*only* $\mathscr{L}_{info}$ + $\mathscr{L}_{trade}$) maintains a reasonable balance between structure and detail. However, removing the trade-off loss (*w/o* $\mathscr{L}_{trade}$) leads to observable degradation in both contour precision and background details. Furthermore, replacing the frequency dynamic convolution module with conventional convolution (*w/o GFDM*) causes significant deterioration in visual quality, manifested as blurred details and discontinuous edges, particularly evident in multi-exposure tasks.

Notably, under extreme conditions such as overexposure/underexposure, the complete model maintains stable fusion performance, while all ablation results exhibit varying degrees of degradation. This further demonstrates the robustness of the DECC framework in complex scenarios. The visual results consistently indicate that the collaborative optimization of dual paths, frequency-aware feature decoupling, and multi-objective trade-off loss collectively form the foundation for zero-shot generalization capability.