Supplementary: Multi-Modal Image Fusion via Intervention-Stable Feature Learning

Xue Wang[1,2], Zheng Guan[1*], Wenhua Qian[1*], Chengchao Wang,[1] Runzhuo Ma[3]
[1] School of Information Science and Engineering, Yunnan University
[2] School of Artificial Intelligence, Nanyang Normal University
[3] Department of Electrical and Electronic Engineering, Hong Kong Polytechnic University
gz_627@sina.com whqian@ynu.edu.cn

1. Why Intervention-Based Learning?

The fundamental challenge in MMIF lies in distinguishing meaningful cross-modal relationships from spurious statistical regularities. Consider a typical training scenario: infrared sensors consistently capture strong thermal signatures from vehicles, while visible cameras simultaneously record their metallic surfaces and headlights. During training, these features co-occur with perfect correlation. However, this correlation conflates two distinct phenomena: genuine complementarity (thermal signature reveals vehicle presence in darkness while visible provides shape details) and spurious association (both modalities happen to observe the same object).

Traditional correlation-based methods cannot disentangle these relationships. They optimize to preserve all statistical dependencies, treating spurious correlations as informative signals. This leads to three critical failure modes:

I. Distribution Shift Vulnerability: Models learn dataset-specific biases rather than fundamental fusion principles. For instance, if training data predominantly contains well-lit scenes where infrared and visible modalities are equally informative, the model may learn to simply average both inputs. When deployed in low-light conditions where visible information degrades, such models fail catastrophically as they never learned true cross-modal compensation.

II. Feature Selection Ambiguity: Without principled criteria for feature importance, models resort to frequency-based selection. Features appearing frequently in training receive high weights regardless of their actual contribution to fusion quality. This explains why many fusion methods produce visually pleasing but semantically incomplete results: they optimize for common patterns while missing rare but crucial details.

III. Architectural Over-Engineering: Lacking theoretical guidance, researchers compensate through increasingly complex architectures. Each new method introduces additional modules, attention mechanisms, or loss functions to capture elusive modal relationships. Yet without understanding what makes features truly complementary, these additions often capture more spurious correlations rather than addressing the root problem.

Our intervention-based framework addresses these issues by actively testing modal dependencies rather than passively observing them. Drawing inspiration from causal reasoning [1], we recognize that genuine relationships should remain stable under systematic perturbations while spurious correlations break down. This principle guides our design of three complementary interventions:

***i*) Complementary masking** tests whether modalities can genuinely compensate for each other's missing information, revealing true complementarity versus redundant encoding. ***ii*) Random masking** identifies features that remain informative under partial observability, distinguishing robust patterns from fragile correlations. ***iii*)Modality dropout** quantifies each modality's unique contribution, preventing over-reliance on dominant signals.

By training under these interventions, our model learns to identify intervention-stable features, those maintaining importance across perturbations. These features represent fundamental fusion patterns rather than dataset artifacts, enabling robust generalization beyond training distributions.

2. Parameter Analysis

We conduct comprehensive experiment to analyze key hyperparameters in our framework. All experiments are performed on the validation dataset with other parameters fixed at default values.

Gate Activation Target η. The parameter η in the gate regularizer $\mathcal{R}(\bar{\mathcal{G}}) = \|\mu(\bar{\mathcal{G}} - \eta)\|_1 - \mathbf{H}(\bar{\mathcal{G}})$ controls the target activation ratio of invariance gates. As shown in the first section of the table 1, η significantly impacts fusion performance.

Table 1. Ablation experiment results for hyperparameters. The best value is highlighted with **Bold**.

Performance of η					
η	**AG**	**SF**	**PSNR**	**CC**	$\mathcal{Q}_{abf}$
0.1	5.823	6.076	62.16	0.580	0.449
0.3	**6.136**	**6.244**	**63.62**	**0.605**	**0.467**
0.5	5.803	5.542	62.93	0.549	0.405
0.7	4.579	4.882	60.58	0.609	0.411
Performance of r					
r	**AG**	**SF**	**PSNR**	**CC**	$\mathcal{Q}_{abf}$
4	3.227	3.482	57.49	0.588	0.284
8	**6.136**	**6.244**	**63.62**	**0.605**	**0.467**
12	4.275	4.915	57.85	0.562	0.345
16	5.293	5.561	60.95	0.612	0.429
Transformer	6.049	6.159	61.85	0.603	0.461
Performance of $\mathcal{M}$ size					
Size	**AG**	**SF**	**PSNR**	**CC**	$\mathcal{Q}_{abf}$
4×4	6.106	6.112	61.81	0.532	0.422
8×8	5.095	5.545	58.63	0.530	0.363
16×16	**6.136**	**6.244**	**63.62**	**0.605**	**0.467**
32×32	4.795	5.123	58.79	0.599	0.377

Setting $\eta = 0.1$ causes the model to identify very few intervention-stable regions, missing important features and degrading overall performance. Conversely, 0.5 and 0.7 lead to excessive gate activation where most regions are deemed stable, reducing the discriminative power of the gating mechanism. The optimal value $\eta = 0.3$ achieves balanced activation, identifying genuinely stable features while maintaining selectivity.

Attention Pooling Size r. The pooling size r in CFI controls the computational efficiency and receptive field of cross-modal attention. Results in the second section of the table 1 reveal a clear performance pattern.

Small pooling sizes $r = 4$ severely limit the receptive field, preventing effective cross-modal reasoning and causing substantial performance degradation across all metrics. Larger values $r = 12$ and $r = 16$ improve performance but plateau beyond $r = 8$, suggesting diminishing returns from increased spatial context. Notably, replacing pooled attention with full Transformer attention yields comparable performance to $r = 16$ but with significantly higher computational cost. Our choice of $r = 8$ optimally balances performance and efficiency, achieving superior results while maintaining reasonable computational complexity.

Intervention Mask Size. The mask size determines the granularity of interventions during training. The third sec-

Algorithm 1 Intervention-Based Multi-Modal Fusion Training

Require: Visible images $\{I_{vi}\}$, Infrared images $\{I_{ir}\}$
Ensure: Trained fusion network parameters Θ
Initialize:
  Siamese encoder: $\mathcal{E}(\cdot;\theta_e)$
  Causal Feature Integrator: CFI$(\cdot;\theta_c)$
  Decoder: $\mathcal{D}(\cdot;\theta_d)$
Hyperparameters:
  Loss weights: $\alpha \leftarrow 0.1$, $\beta \leftarrow 0.05$, $\lambda_1 \leftarrow 1.0$
  Gate target: $\eta \leftarrow 0.3$, Pooling size: $r \leftarrow 8$
  Training epochs: K, Batch size: $B \leftarrow 16$, Learning rate: $\gamma \leftarrow 10^{-4}$
for epoch $= 1$ to K **do**
  for each batch $\mathcal{B} = \{(I_{vi}^j, I_{ir}^j)\}_{j=1}^{B}$ **do**
    % Generate intervention variants
    for each pair $(I_{vi}, I_{ir}) \in \mathcal{B}$ **do**
      $\mathcal{M}_v^c, \mathcal{M}_i^c \leftarrow$ ComplementaryMask(H, W)
      $\mathcal{M}^r \leftarrow$ RandomMask(H, W)
      $\mathcal{S} \leftarrow \{(I_{vi}, I_{ir})\}$ % Original pair
      $\mathcal{S} \leftarrow \mathcal{S} \cup \{(I_{vi} \odot \mathcal{M}_v^c, I_{ir} \odot \mathcal{M}_i^c)\}$ % Complementary
      $\mathcal{S} \leftarrow \mathcal{S} \cup \{(I_{vi} \odot \mathcal{M}^r, I_{ir} \odot \mathcal{M}^r)\}$ % Random
      $\mathcal{S} \leftarrow \mathcal{S} \cup \{(\mathbf{0}, I_{ir}), (I_{vi}, \mathbf{0})\}$ % Modality dropout
    end for
    % Forward pass for all intervention variants
    for each $(I_{vi}^p, I_{ir}^p) \in \mathcal{S}$ **do**
      $\{\Phi_k^v\}_{k=1}^3 \leftarrow \mathcal{E}(I_{vi}^p;\theta_e)$ % Encode visible
      $\{\Phi_k^i\}_{k=1}^3 \leftarrow \mathcal{E}(I_{ir}^p;\theta_e)$ % Encode infrared
      % Hierarchical fusion with CFI
      $\Psi_3 \leftarrow \text{CFI}(\Phi_3^v, \Phi_3^i;\theta_c)$
      $\Psi_2 \leftarrow \text{CFI}(\Phi_2^v, \Phi_2^i;\theta_c) + \mathcal{D}(\Psi_3;\theta_d)$
      $\Psi_1 \leftarrow \text{CFI}(\Phi_1^v, \Phi_1^i;\theta_c) + \mathcal{D}(\Psi_2;\theta_d)$
      $I_f^p \leftarrow \sigma(\text{Conv}_{1\times1}(\Psi_1))$ % Generate fused output
    end for
    % Compute losses
    $\mathcal{L}_f \leftarrow$ FusionLoss(I_f, I_{vi}, I_{ir}) using Eq. (10)
    $\mathcal{L}_{inv} \leftarrow$ InterventionLoss$(I_f, I_f^c, I_f^r, \bar{\mathcal{G}})$ using Eq. (11)
    $\mathcal{L}_{nec} \leftarrow$ NecessityLoss(I_f, I_f^i, I_f^v) using Eq. (12)
    $\mathcal{L}_{total} \leftarrow \mathcal{L}_f + \alpha \cdot \mathcal{L}_{inv} + \beta \cdot \mathcal{L}_{nec}$
    Update $\{\theta_e, \theta_c, \theta_d\}$ via backpropagation with learning rate γ
  end for
end for
return Trained network parameters $\Theta = \{\theta_e, \theta_c, \theta_d\}$

Table 2. Comparison of model complexity and efficiency. The best value is highlighted with **Bold**.

Metric	TIMFusion	SAGE	MUFusion	LUT-Fuse	LRRNet	IGNet	DCEvo	Conti	A^2RNet	Ours
Para (M)	0.62	0.14	0.55	**0.01**	0.05	7.87	2.01	1.66	10.6	0.23
Flops (G)	116	4.3	118	**0.5**	3.3	54.1	195	23	37	6.0
FPS	0.52	17	3.64	30	8.33	7.36	1.43	3.20	0.16	**36**

tion of the table 1 demonstrates its crucial impact on learning intervention-stable features.

Small masks (4×4) create minimal perturbations that fail to challenge the model, resulting in limited robustness learning. Conversely, excessive masking (32×32) corrupts too much information, hindering effective feature learning and cross-modal compensation. The intermediate size (16×16) provides optimal perturbation strength: sufficient to test feature robustness without destroying essential spatial context. This size also aligns with typical receptive fields in our encoder, ensuring interventions meaningfully probe learned representations.

3. Implementation Details

Following parameter analysis, we set both complementary and random masks to 16×16 pixels. The number of masks per image is randomly sampled from one to six, with total masked area constrained not to exceed half the training patch to preserve sufficient information for learning. We set $\eta = 0.3$ for balanced gate activation and $r = 8$ for efficient cross-modal attention. These parameters remain fixed across all experiments, demonstrating the robustness of our framework to different datasets and fusion scenarios.

Notably,to align the input data modalities, we utilize the YCbCr color space to separate the luminance and chrominance components of VI, and restore the VI chrominance of the fused image after fusion. The overall training strategy can be found in Algorithm 1. To ensure fairness, all comparison models were provided by their original authors.

4. Efficiency Analysis

Beyond fusion quality, practical deployment hinges on computational efficiency. Table 2 compares metrics across methods on 256×256 images under identical hardware. Our approach excels in parameter efficiency with just 0.23M parameters, far below most competitors, while LUT-Fuse and LRRNet achieve smaller sizes at the expense of fusion quality, and parameter-heavy models like A^2RNet and IGNet yield disproportionate gains. This stems from intervention-based training, which fosters robust fusion patterns with minimal overhead via the CFI module and feature-reusing invariance gating. In FLOPs, our method demands only 6.0G operations, outpacing attention-intensive TIMFusion and MUFusion, thanks to linear-complexity spatially pooled cross-attention in CFI and discriminative feature learning without elaborate correlation modeling. For real-time use, it delivers the top inference speed at 36 FPS, surpassing optimized LUT-Fuse, enabled by streamlined architecture that fuses intervention-stable features through simple gating, unlike multi-branch methods such as A^2RNet and TIMFusion. Our intervention framework effectively resolves the efficiency-performance trade-off that characterizes existing approaches: lightweight methods such as LRRNet compromise output quality in pursuit of computational efficiency, whereas sophisticated architectures like DCEvo achieve performance improvements at the expense of substantial computational overhead.

References

[1] Judea Pearl. Causal inference. *Causality: objectives and assessment*, pages 39–58, 2010. 1