Deep Image Prior

Haozheng Wang
College of Engineering, Peking University
2400011002@stu.pku.edu.cn

Zhonghe Liu
College of Engineering, Peking University
2400011220@stu.pku.edu.cn

Yihui Lan
School of Electronic Engineering and Computer Science, Peking University
2400013178@stu.pku.edu.cn

Abstract

This work explores the *Deep Image Prior* (DIP) framework, which demonstrates that the structure of a randomly initialized convolutional network itself encodes a powerful prior for natural images, enabling high-quality image restoration without external data. We first reproduce core DIP experiments on denoising, super-resolution, and inpainting. We then propose a multi-scale DIP extension that enhances structural coherence in inpainting and further adapt the framework to image blending, where it outperforms classical fusion methods. Both quantitative metrics and visual assessments confirm the effectiveness of our approach.

1 Introduction

Deep Image Prior (DIP) demonstrates that the structure of a convolutional network itself, even when untrained, provides a strong implicit prior for natural images, enabling effective solutions to inverse problems such as denoising, super-resolution and inpainting without external data. Building on this idea, we evaluate the reproducibility and extensibility of the DIP framework. Specifically, we reproduce its core results on standard restoration tasks and personal images, propose a multi-scale reconstruction strategy that improves coherence in image inpainting, and extend DIP to image blending, where it outperforms traditional methods.

2 Related work

In 2018, the proposal of DIP by Dmitry Ulyanov et al. [1] breaks the reliance of traditional supervised deep learning on large-scale annotated datasets, pioneering a new direction of image restoration that is training-free and data-independent. Its core insight lies in the fact that the structure of convolutional neural networks (CNNs) inherently contains implicit priors of natural images. Through hierarchical convolutions and spatial constraints, the network naturally tends to generate outputs with statistical characteristics of natural images such as local smoothness and edge continuity.

Further research has shown strong interest in integrating DIP with existing methods to develop a more powerful DIP framework. For instance, total variation regularization has been reintroduced to the DIP framework to improve image restoration [2], while another line of work combines DIP with Regularization by Denoising (RED) to boost stability and detail preservation [3]. However, we feel that the potential of DIP's inherent architectural design remains underexplored. Drawing inspiration from the U-Net [4] architecture—which serves as the backbone of DIP—we realize that its capacity for multi-scale feature integration may constitute a fundamental reason for DIP's effectiveness. By explicitly formalizing this multi-scale principle within the loss function, its advantages can be further amplified, which will be elaborated in the subsequent sections of this work.

3 Data

The DIP method operates on a single degraded observation without requiring any external training dataset. For test images, we employ a combination of benchmark images referenced in the original paper [1] and additional images from personal collections, encompassing both natural scenes and structured man-made content to ensure overall applicability.

To simulate realistic degradation scenarios, each input undergoes tailored preprocessing: synthetic Gaussian noise (typically $\sigma = 25$) is introduced for denoising tasks; images are downsampled to lower resolutions to assess super-resolution capability; and structured or randomly generated masks are applied to simulate inpainting challenges. These controlled modifications enable a systematic evaluation of DIP's effectiveness across varied and practical image restoration contexts.

4 Methods

4.1 Original DIP

We employ a U-Net-style encoder-decoder architecture with skip connections that generates an output image $f_\theta(z)$ from a fixed random latent tensor z. The restoration problem is formulated as an energy minimization problem:

$$\theta^* = \arg\min_\theta \; E\big(f_\theta(z);\, x_0\big),$$

where x_0 denotes the degraded observation and E is a task-specific data term. In this framework, the implicit prior induced by the network architecture replaces explicit regularizers (e.g., total variation), which is the key innovation of DIP. Optimization is performed over the network parameters θ using the Adam optimizer without any pre-training on external datasets. The network is fitted to a single degraded observation, exploiting its structural bias to recover the clean signal before overfitting to noise or corruptions.

4.2 Multi-scale DIP and its application in image blending

The multi-scale DIP enhances the original framework by reformulating its loss function to explicitly leverage and strengthen its inherent capacity for multi-scale feature representation. Let $\mathbf{I} \in [0,1]^{C\times H\times W}$ be the target image and $\mathbf{O}$ the network output. For inpainting, we use a binary mask $\mathbf{M} \in \{0,1\}^{C\times H\times W}$ indicating missing pixels.

The multi-scale loss at iteration t is:

$$\mathcal{L}_{\text{ms}}^{(t)}(\mathbf{O},\mathbf{I},\mathbf{M}) = \sum_{s\in\{1,2,4\}} w_s \cdot \mathcal{L}_s^{(t)}(\mathbf{O},\mathbf{I},\mathbf{M})$$

where $\mathcal{L}_s^{(t)}$ is the masked MSE at scale s:

$$\mathcal{L}_s^{(t)}(\mathbf{O},\mathbf{I},\mathbf{M}) = \frac{\|\mathbf{M}_s \odot (\mathbf{O}_s - \mathbf{I}_s)\|_2^2}{\|\mathbf{M}_s\|_1}$$

Here $\odot$ denotes element-wise multiplication, $\mathbf{O}_s = \text{Downsample}_s(\mathbf{O})$, and similarly for $\mathbf{I}_s$ and $\mathbf{M}_s$. Downsampling uses average pooling with stride s. Weights are typically set to $w_1 = 1.0, w_2 = 0.5, w_4 = 0.25$ in the following experiments.

Given a foreground image, a background image, and a coarse binary mask, the following approach based on multi-scale DIP generates a naturally fused result without solving Poisson equations or using external training data. We employ an encoder-decoder network with skip connections, take random noise as input and directly output the blended image. Instead of a global pixel-wise loss, we design a loss function that separately constrains three regions derived from the mask: the foreground, the background, and a transition zone. The loss comprises:

- **Background region:** A pixel-level MSE constraint, supplemented by illumination and low-frequency consistency terms to prevent global brightness drift.
- **Foreground region:** A multi-scale gradient consistency loss to preserve structural edges and contours while allowing color adaptation.
- **Transition zone:** An anisotropic texture propagation constraint that uses a soft weight map and a structure tensor to encourage smooth blending along dominant texture directions.

For mathematical depiction of the loss above, please refer to supplementary material.

To avoid artifacts from hard mask boundaries, the binary mask is smoothed to create a continuous soft weight map. Furthermore, an alpha annealing strategy is applied during optimization to progressively refine the blending boundary, thereby improving convergence stability and final visual quality. The optimization is stopped early to prevent overfitting to noise, leveraging DIP's tendency to first recover natural image structures.

5 Experiments

5.1 Reproduction of denoising and super-resolution

5.1.1 Metric

For denoising, super-resolution and inpainting, we use peak signal-to-noise ratio (PSNR) as our primary quality metric. Given an $m \times n$ RGB image I and its approximation K, this metric is calculated using the following formula:

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}_I^2}{\frac{1}{3nm}\sum_{R,G,B}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I_{color}(i,j) - K_{color}(i,j)]^2}\right)$$

where MAX_I denotes the maximum possible value of the image.

5.1.2 Denoising

Experiments are performed on images *F16_GT.png* and *snail.jpg* from [1], along with *delicious.jpg* from personal collections. The metric *psrn_noisy/gt* measures PSNR between the output and noisy/ground truth images, while *srn_gt_sm* evaluates PSNR after smoothing. For *snail.jpg*, lacking ground truth, the latter two metrics are inapplicable. The outcomes demonstrate that DIP effectively restores both artificial and natural structures.

Table 1: Denoising

Iteration	F16_GT.png			snail.jpg	delicious.jpg		
	psrn_noisy	psrn_gt	psrn_gt_sm	psrn_noisy	psrn_noisy	psrn_gt	psrn_gt_sm
0	10.818	11.295	11.295	7.714	13.479	14.043	14.043
300	18.701	23.523	23.290	16.306	18.655	22.049	21.425
600	19.498	26.692	27.309	20.265	19.872	25.703	26.305
900	19.730	27.938	29.315	23.027	20.359	28.089	28.965
1200	19.883	28.903	30.484	25.165	20.565	29.581	30.613
1500	19.986	29.538	31.345	25.931	20.672	30.452	31.600
1800	20.077	30.121	32.000	26.816	20.718	31.025	32.222
2100	20.154	30.455	32.414	27.360	20.770	31.510	32.663
2400	20.221	30.718	32.649	-	-	-	-
2700	20.248	30.282	32.712	-	-	-	-

5.1.3 Super-Resolution

On the Blake_Lively_00.png test image, the PSNR rises rapidly from 21.447 to 28.263 within the first 900 iterations, then continues to increase steadily and reaches 29.964 at 2,100 iterations.

Notably, this PSNR value has already surpassed that of bicubic interpolation (29.8901) by 2,100 iterations; while the final result (at 3,900 iterations) reaches 30.935, which is 1.0449 higher than that of bicubic interpolation and 4.3077 higher than that of nearest-neighbor interpolation. This result further validates that the DIP method consistently demonstrates superior performance over conventional non-trained upsampling techniques in super-resolution tasks across different types of images.

Table 2: Super-Resolution

Iteration	zebra_GT.png		Blake_Lively_00.png	
	PSNR_HR	PSNR_LR	PSNR_HR	PSNR_LR
300	20.924	24.361	25.093	26.682
600	22.326	27.568	27.209	29.912
900	23.033	29.446	28.263	32.062
1200	23.480	30.399	28.928	33.042
1500	23.767	31.624	29.399	34.441
1800	23.879	32.403	29.686	35.660
2100	24.119	32.842	29.964	36.523
2400	24.216	33.947	30.155	37.206
2700	24.361	34.396	30.257	37.879
3000	24.404	34.869	30.485	37.898
3300	24.502	35.789	30.749	38.296
3600	24.559	36.251	30.713	38.621
3900	24.647	37.002	30.935	40.107

5.2 Image inpainting with original DIP and multi-scale DIP

We conducted comparative experiments between the original DIP and the multi-scale DIP on three test images: *kate.png*, *vase.png*, and *library.png*. Two region-specific PSNR metrics were employed for evaluation: *psnr_masked*, computed exclusively within the inpainted mask region, and *psnr_coarse*, measured on the masked region after downsampling the image by a factor of 4. Note that for *vase.png* and *library.png*, the latter two metrics are not applicable due to the intrinsic properties of the inpainting task.

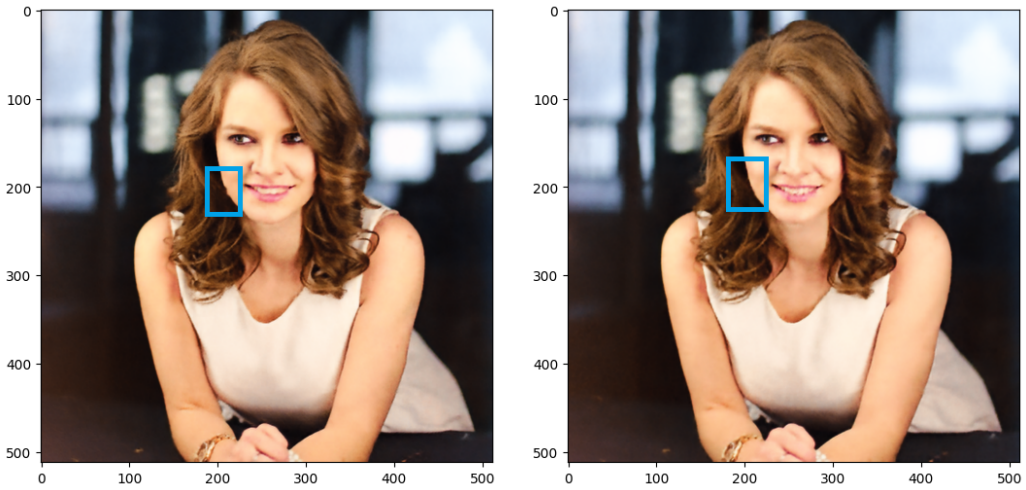
The experimental data demonstrate that the multi-scale DIP consistently outperforms the original DIP, achieving higher PSNR values upon convergence. This performance gap becomes even more pronounced when examining the visual results. As highlighted in the blue region of Figure 1, the multi-scale DIP produces locally smoother and more coherent inpainting results compared to the original DIP.

Table 3: Comparison between original DIP and multi-scale DIP: *kate.png*

Iteration	Original DIP			Multi-scale DIP		
	psnr	psnr_masked	psnr_coarse	psnr	psnr_masked	psnr_coarse
0	8.8386	4.6847	7.1461	8.9174	4.9836	7.4939
600	29.0963	21.1456	25.4868	28.9473	20.8022	24.8590
1200	31.8359	22.6755	27.7126	31.6311	22.6331	27.7803
1800	33.6385	24.3045	30.0019	33.2704	24.1710	30.0681
2400	35.0535	25.2139	31.1267	34.2900	24.8937	30.8459
3000	34.9377	25.3944	30.9944	34.7598	25.5820	31.5321
3600	36.7344	26.0520	32.1317	36.2608	26.3815	33.4666
4200	36.9862	26.0794	31.9815	36.4430	26.3816	32.6985
4800	37.9309	26.6463	33.3572	37.5931	27.2485	34.5113
5400	38.1235	26.4713	32.7123	38.1450	27.4124	34.9180
6000	38.5044	26.6183	33.1859	38.5234	27.5851	35.0679

Table 4: Comparison between original DIP and multi-scale DIP: *vase.png* and *library.png*

Iteration	*vase.png* (PSNR)		*library.png* (PSNR)	
	Original DIP	Multi-scale DIP	Original DIP	Multi-scale DIP
0	12.5253	12.8969	9.4005	9.3677
600	21.4033	21.8555	18.0442	17.5735
1200	25.4794	23.8434	19.1009	18.7700
1800	27.3330	26.8672	19.2487	19.1232
2400	24.5847	27.9522	19.4136	19.3303
3000	27.2474	28.5666	19.4083	19.4313
3600	28.1106	26.4479	19.4949	19.5459
4200	28.0698	23.9175	19.4229	19.5958
4800	28.8467	28.0888	19.4398	19.5495
5400	28.4320	29.1012	19.3819	19.5964
6000	28.5350	28.9807	19.3372	19.5045



(a) Original DIP (b) Multi-scale DIP

Figure 1: Multi-scale DIP demonstrates stronger continuity and smoothness in local inpainting.

5.3 Image blending with multi-scale DIP

Figure 2: The child and the lake.

Figure 3: The bear and the penguin.

Table 5: Quantitative comparison of different image fusion methods on *the child and the lake*.

Method	**GFF**	**GFP**	**BSI**	**SSIM**
Alpha	0.666876	0.840638	15.171874	0.363135
Poisson	0.608026	0.842703	22.493957	0.679560
Ours	0.665686	9.351012	29.016716	0.834993
Laplacian	0.658930	1.861390	27.954209	0.334541

Table 6: Quantitative comparison of different image fusion methods *the bear and the penguin*.

Method	**GFF**	**GFP**	**BSI**	**SSIM**
Alpha	0.591284	0.817655	515.180007	0.570862
Poisson	0.611293	2.090503	585.931454	0.837563
Ours	0.709274	2.785332	133.374406	0.848493
Laplacian	0.593106	1.932817	562.232692	0.660616

Table 7: Comparison of different blending methods on four evaluation dimensions.

Method	**structural retention**	**Directional Texture**	**light consistency**	**transition naturalness**
Alpha	Poor	None	Medium	Poor
Laplacian	Medium	Fragile	Fine	Medium
Poisson	Medium	None	Instable	Medium
Ours	Excellent	Excellent	Excellent	Excellent

From left to right, the fused results generated by Alpha blending, Poisson blending, Laplacian pyramid blending, and the proposed multi-scale DIP blending are displayed. Quantitative evaluations for these two sets of blended images are summarized in Table 5 and Table 6 (the detailed definitions of these metrics are provided in the supplementary materials due to space constraints). A qualitative assessment of visual characteristics across different fusion methods is presented in Table 7. Collectively, these results demonstrate that the multi-scale DIP blending method achieves superior performance compared to the traditional blending techniques.

6 Conclusion

In this paper, we have validated the effectiveness of DIP in various image restoration tasks and extended it through a multi-scale formulation. Our experiments show that the multi-scale DIP not only improves inpainting quality by enhancing structural continuity but also provides a robust solution for image blending, surpassing traditional methods in both quantitative metrics and visual smoothness. The ability to achieve these results without any training data underscores the strength of architectural priors in convolutional networks. Future work may explore adaptive weighting schemes across scales and applications to more complex image synthesis tasks.

References

[1] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 9446-9454.

[2] J. Liu, Y. Sun, X. Xu, and U. S. Kamilov. Image restoration using total variation regularized deep image prior. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7715-7719.

[3] Gary Mataev, Peyman Milanfar, and Michael Elad. DeepRED: deep image prior powered by RED. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 0-0.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234-241.

Supplementary material

A1 Runnable source code

Runnable source code of this work can be obtained here.

A2 Supplementary results

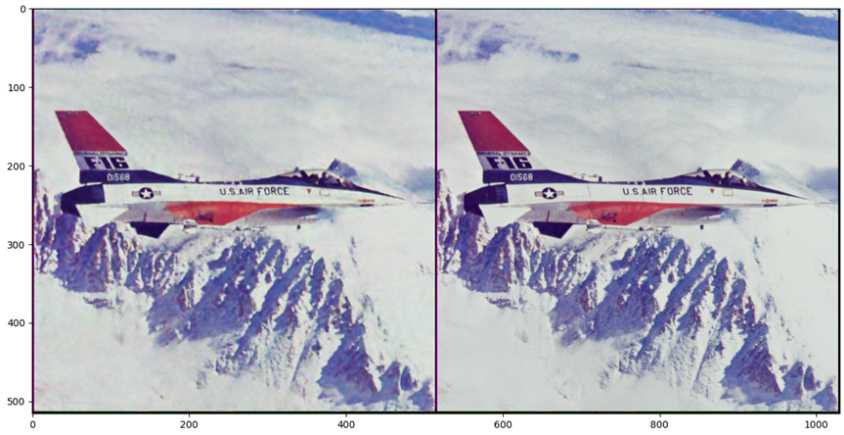


Figure 4: *F16* reproduced from [1]. Left: DIP-denoised image; Right: ground truth.

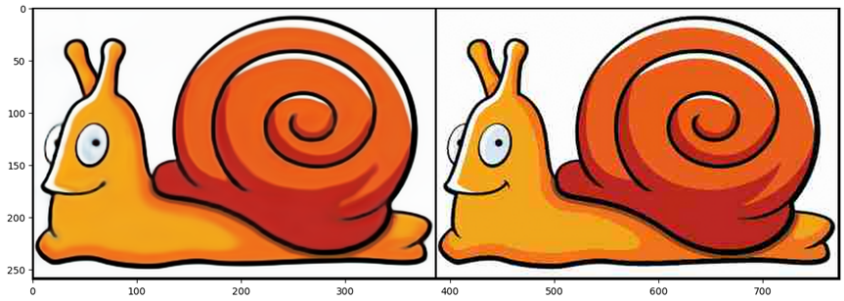


Figure 5: *Snail* Reproduced from [1]. Left: DIP-denoised image; Right: ground truth.

Figure 6: *Tachibana Sherry* from personal collections. Left: DIP-denoised image; Right: ground truth.

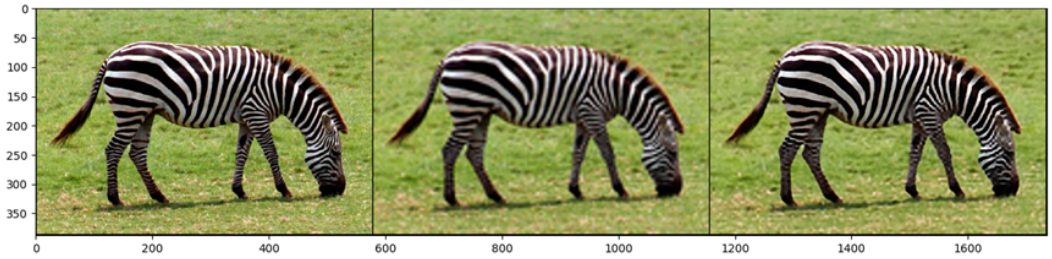


Figure 7: *Zebra* reproduced from [1]. Left: ground truth; Middle: super-resolution by bicubic interpolation; Right: super-resolution by DIP.

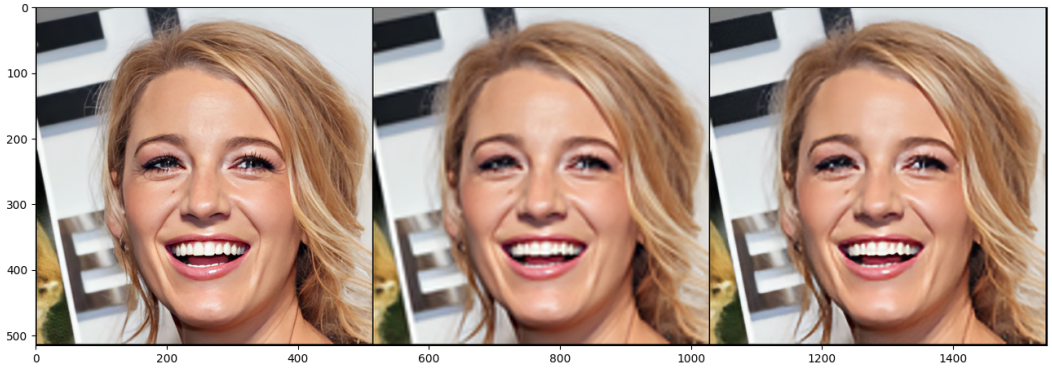


Figure 8: *Blake Lively* from personal collections. Left: ground truth; Middle: super-resolution by bicubic interpolation; Right: super-resolution by DIP.

Apple image

Orange Image

(a) Original images to be blended

(b) Ours

(c) Laplacian blending

Figure 9: Another comparison experiment between multi-scale DIP blending and Laplacian blending. (a)(c) are collected from here.

A3 Mathematical depiction of the loss in multi-scale DIP blending

Soft Transition Band Modeling

Instead of directly using a hard binary mask, we construct a soft transition weight map ($\alpha(x) \in [0, 1]$) by applying repeated low-pass filtering to the original mask. Pixels inside the foreground region have weights close to 1, background pixels close to 0, and pixels near the boundary form a smooth transition band.

To further improve boundary sharpness while maintaining early-stage flexibility, we adopt an annealing strategy during optimization:

$$\alpha_t(x) = \text{clip}(\alpha(x)^{\gamma(t)}, 0, 1), \quad \gamma(t) = 1 + 4\left(1 - \frac{t}{T}\right)^c$$

where t denotes the current iteration and T the total number of iterations. This allows the transition region to be wide and smooth at early stages, and gradually become sharper as optimization converges.

Background Appearance Consistency

To preserve the original background structure and illumination, we enforce both pixel-level and low-frequency consistency outside the foreground region:

$$\mathcal{L}_{\text{pixel}} = \mathbb{E}_x[(1-\alpha_t(x))\|F(x)-B(x)\|_2^2]$$

$$\mathcal{L}_{\text{low-freq}} = \mathbb{E}_x[(1-\alpha_t(x))\|G(F)(x)-G(B)(x)\|_2^2]$$

where $G(\cdot)$ denotes a large-kernel average filter capturing low-frequency components.

Additionally, an illumination consistency term is introduced by extracting large-scale grayscale illumination:

$$\mathcal{L}_{\text{illum}} = \mathbb{E}_x[(1-\alpha_t(x))\|J(F)(x)-J(B)(x)\|_2^2]$$

Foreground Structure Preservation

To prevent foreground texture degradation, we preserve multi-scale gradient information inside the foreground region:

$$\mathcal{L}_{\text{grad}} = \frac{1}{|S|}\sum_{s\in S}\mathbb{E}_x\left[\alpha_t^{(s)}(x)\|\nabla F^{(s)}(x)-\nabla A^{(s)}(x)\|_1\right]$$

We further enforce gradient direction consistency to preserve structural orientation:

$$\mathcal{L}_{\text{dir}} = \mathbb{E}_x\left[\alpha_t(x)\left(1-\frac{\nabla F(x)\cdot\nabla A(x)}{\|\nabla F(x)\|\|\nabla A(x)\|}\right)\right]$$

Gradient Blending Consistency in Transition Band

To explicitly model smooth gradient propagation across the boundary, we enforce the fused gradient to approximate a weighted combination of foreground and background gradients:

$$\mathcal{L}_{\text{blend}} = \mathbb{E}_x[w(x)\|\nabla F(x)-\big(\alpha_t(x)\nabla A(x)+(1-\alpha_t(x))\nabla B(x)\big)\|_1]$$

where $w(x)=4\alpha_t(x)(1-\alpha_t(x))$ activates this constraint mainly within the transition band.

Anisotropic Texture Propagation

Natural scenes such as water surfaces or roads exhibit strong directional texture continuity. To exploit this property, we estimate a local structure tensor from the background image and derive its tangential direction ($\mathbf{t}(x)$). We then enforce texture continuity along this direction:

$$\mathcal{L}_{\text{aniso}} = \mathbb{E}_x[w(x)|\nabla_{\mathbf{t}(x)}F(x)-\nabla_{\mathbf{t}(x)}B(x)|]$$

$$\mathbf{t}(x)=\begin{bmatrix}-\sin\theta(x)\\ \cos\theta(x)\end{bmatrix},\quad \theta(x)=\frac{1}{2}\arctan\left(\frac{2J_{xy}}{J_{xx}-J_{yy}}\right).$$

Overall Objective and Optimization

The final optimization objective is a weighted sum of all loss terms:

$$\mathcal{L}_{\text{total}} = \lambda_p\mathcal{L}_{\text{pixel}}+\lambda_{lf}\mathcal{L}_{\text{low-freq}}+\lambda_{\text{ill}}\mathcal{L}_{\text{illum}}+\lambda_g\mathcal{L}_{\text{grad}}+\lambda_d\mathcal{L}_{\text{dir}}+\lambda_b\mathcal{L}_{\text{blend}}+\lambda_a\mathcal{L}_{\text{aniso}}$$

$$(\lambda_p, \lambda_{lf}, \lambda_{\text{ill}}, \lambda_g, \lambda_d, \lambda_b, \lambda_a) = (50, 25, 20, 6, 4, 12, 18)$$

We optimize this objective using a Deep Image Prior (DIP) framework, where a randomly initialized convolutional network is trained to generate the fused image. Input noise perturbation is applied at each iteration to improve robustness and prevent overfitting.

A4 More metrics for image blending

Gradient-based Fusion Factor (GFF)

The GFF metric evaluates the preservation of salient structural information in source images by fusion results, considering both gradient magnitude and direction. At each pixel location, the source image with greater gradient magnitude is selected as the reference.

$$G_S(x) = \max(G_A(x), G_B(x)),$$

$$\theta_S(x) = \begin{cases} \theta_A(x), & G_A(x) \geq G_B(x), \\ \theta_B(x), & \text{otherwise}, \end{cases}$$

The definition of fused gradient consistency is

$$Q(x) = \frac{2G_F(x)G_S(x)}{G_F(x)^2 + G_S(x)^2} \cdot \frac{1 + \cos(\theta_F(x) - \theta_S(x))}{2},$$

The GFF is averaged across the transition region.

$$\text{GFF} = \frac{1}{|\Omega_T|} \sum_{x \in \Omega_T} Q(x)$$

Gradient Fusion Performance (GFP)

This metric evaluates whether the fused image adequately captures the primary variations in the source image through gradient magnitude.

$$\text{GFP} = \frac{1}{N} \sum_{x} \frac{G_F(x)}{\max(G_A(x), G_B(x)) + \varepsilon}$$

A higher GFP value indicates that the fusion results retain critical variation information from the source images at the gradient level.

Boundary Smoothness Index (BSI)

To quantify whether the fusion boundary exhibits ringing or unnatural abrupt changes, this paper introduces a boundary smoothness metric. The method first computes the gradient magnitude of the fused image and compares it with its Gaussian-smoothed version:

$$\Delta G(x) = |G_F(x) - G(G_F)(x)|,$$

$$\text{BSI} = \frac{1}{|\Omega_T|} \sum_{x \in \Omega_T} \frac{1}{\Delta G(x) + \varepsilon}$$

A higher BSI indicates a more gradual gradient change in the transition region, resulting in a more natural fusion.

Masked SSIM

To evaluate the preservation of background structures in the transition zone during fusion, this study calculates the Structural Similarity Index (SSIM) within the Ω_T region.

$$\text{SSIM}_T = \frac{1}{|\Omega_T|} \sum_{x \in \Omega_T} \text{SSIM}(F(x), B(x))$$

The index is used to measure the consistency of the fusion results with the background at the structural level.