Sensors and Materials, Vol. 38, No. 8 (2026) 4515–4532
MYU Tokyo

S & M 4579

DFRE-Net: a Dual-branch Cross-feature Fusion with Regional Enhancement Network for Image Inpainting

Rongrong Gong,[1†] Ronghao Luo,[2†] Minzhi Yuan,[3]
Yuehong Tian,[4] Dengyong Zhang,[2] and Yan Li[5*]

[1]School of Software, Changsha Social Work College, Changsha 410004,China
[2]School of Computer Science and Technology, Changsha University of Science and Technology, Changsha 410114, China
[3]Department of Information Engineering, Hunan Mechanical and Electrical Polytechnic, Changsha 410151, China
[4]Changkuangao Beijing Technology Co., Ltd., Changsha 101100, China
[5]Department of Computer Engineering, INHA University, Changsha 22212, South Korea

(Received August 26, 2025; accepted July 21, 2026)

Keywords: big date, image inpainting, learning-based fusion technique, cross-feature fusion

Image inpainting is a critical task in computational intelligence, particularly in AI-driven big data environments, where it involves synthesizing missing image content based on available information to restore occluded or damaged regions. As a learning-based fusion technique, image inpainting leverages AI to intelligently reconstruct missing or corrupted areas. To overcome the limitations of existing approaches, where convolutional neural networks (CNNs) struggle with global semantic coherence and Transformers lack sensitivity to fine local details. In this paper, the authors introduce DFRE-Net, a dual-branch cross-feature fusion network with regional enhancement for high-fidelity image inpainting. The encoder adopts a computationally intelligent dual-branch architecture: a CNN branch with multiscale residual connections extracts high-frequency details, whereas a Transformer branch captures long-range contextual relationships. The decoder incorporates a Cross-feature Fusion Module (CFFM), utilizing multidilation depthwise separable convolutions for adaptive local-global feature integration, and a Regional Enhancement Module (REM) with mask-guided optimization to reduce foreground-background interference. Experimental results demonstrate that the proposed method outperforms state-of-the-art models on the Paris StreetView, Places365, and CelebA-HQ datasets, validating the effectiveness of cross-architecture feature fusion in addressing complex occlusion inpainting tasks.

1. Introduction

Image inpainting is a fundamental task in computational intelligence, which aims to reconstruct missing or corrupted regions while preserving visual authenticity and semantic consistency.(1) Despite their proven success, prevailing deep learning architectures for image inpainting present complementary limitations rooted in their inherent designs. Convolutional

*Corresponding author: e-mail: leeyeon@inha.ac.kr
†These authors contributed equally to this work.
https://doi.org/10.18494/SAM5891

ISSN 0914-4935
https://myukk.org/

neural networks (CNNs),[2] driven by their inductive bias towards locality, excel at capturing local textures and patterns. However, their limited receptive fields often hinder the establishment of long-range dependencies, which can result in globally inconsistent content generation. For instance, when reconstructing large missing regions, CNNs may produce semantically plausible local textures that, nonetheless, fail to align with the overall scene context, such as generating window frames that mismatch a building's architectural style. On the other hand, Transformer-based models[3] with their global self-attention mechanisms, adeptly model long-range contextual relationships, ensuring broad semantic coherence. This capability, however, comes at a cost. The standard self-attention mechanism operates on a sequence of image patches, which can lead to an oversmoothing of fine-grained, high-frequency details critical for visual realism, such as sharp edges, fine hair strands, or intricate textures. Consequently, while Transformer-based inpaints maintain strong structural consistency at a macro level, they may lack the pixel-level acuity afforded by CNNs.

To address these challenges, in this paper, the authors propose a learning-based fusion framework with a hybrid CNN-Transformer architecture and dual cooperative mechanisms. The encoder adopts a bifurcated architecture: a CNN branch enhanced by residual connections and multiscale convolutions to capture fine-grained high-frequency local patterns and a Transformer branch dedicated to modeling global contextual relationships, ensuring semantic alignment between corrupted regions and their surroundings. In the decoder phase, a Cross-feature Fusion Module (CFFM) dynamically aligns and integrates multiscale features from both branches through three parallel depthwise separable convolution streams with dilation rates (r = 1,2,3). This design adaptively aggregates local, mid-range, and global contextual cues while suppressing feature redundancy through channel-wise compression and pixel-wise adaptive weighting. Complementing this, a Regional Enhancement Module (REM) implements mask-guided feature disentanglement, where foreground and background features are separately processed via learnable and inverse masks. An Attention Enhancement Module (AEM) within REM automatically regulates enhancement intensities on the basis of regional semantics, effectively resolving feature conflicts between foreground and background.

Experiments on Paris StreetView,[4] Places365,[5] and CelebA-HQ[6] datasets demonstrate that our framework achieves superior performance in handling complex occlusions compared with state-of-the-art methods, particularly in preserving structural continuity and high-frequency details. The key innovations of this work include the following:

- a dual-branch encoder architecture where a residual-enhanced CNN branch captures high-frequency local patterns and a Transformer branch models long-range contextual relationships, enabling complementary local-global feature extraction,
- multiscale adaptive fusion via CFFM, employing parallel depthwise separable convolutions with carefully designed dilation rates to capture hierarchical contextual features, while achieving the effective integration of multiscale representations through channel-wise compression and spatial adaptive weighting,
- mask-guided regional optimization via REM, leveraging learnable binary masks to achieve feature disentanglement between foreground and background regions, with an embedded AEM that enables precise region-specific refinement, and

- experiments on three public datasets to evaluate the effectiveness of the proposed model in various image inpainting tasks. DFRE-Net surpassed other methods of the same type on these three datasets.

2. Related Work

2.1 CNN-based image inpainting

Pathak *et al.*[7] first introduced adversarial networks into image inpainting via Context Encoder, combining an encoder-decoder structure with GAN's adversarial mechanism, using reconstruction and adversarial losses to improve inpainting quality for arbitrary regions. However, adversarial loss was only applied to the inpainted area, neglecting global semantic consistency and boundary continuity, causing blurring and artifacts. Iizuka *et al.*[8] addressed this by adding a global discriminator to ensure both local and global consistency. Liu *et al.*[9] proposed Partial Convolutions to distinguish valid and invalid pixels, mitigating interference from missing regions. As the network deepens, invalid pixels may become valid, so Yu *et al.*[10] introduced Gated Convolutions to dynamically update masks and filter invalid information. Nevertheless, convolution-based methods suffer from limited receptive fields, making it challenging to balance global semantics and local textures.

2.2 Transformer-based image inpainting

Wan *et al.*[11] first introduced the ICT model, which applies Transformers to image inpainting. Leveraging the self-attention mechanism, ICT achieves a global receptive field for image modeling. However, the computational complexity of self-attention scales quadratically with sequence length (token count), posing significant challenges for high-resolution image restoration and limiting its deployment on large-scale datasets. To mitigate this, researchers have proposed several strategies, including dividing the image into smaller patches to reduce sequence length or restricting self-attention computation to local windows.[12] While these methods reduce computational costs, they fundamentally transform global self-attention into localized computations, compromising the model's ability to capture long-range dependencies. Although Deng *et al.*[13,14] introduced linear attention to decrease computational overhead, this approach often leads to reduced model accuracy. Moreover, compared with CNNs, Transformers exhibit weak capability in capturing fine-grained local details. Consequently, the generated inpainting results, particularly in cultural heritage image restoration, may exhibit smooth artifacts in local textures while maintaining overall structural coherence.

3. Our Approach

3.1. Network architecture

The overall architecture of our proposed Dual-branch Feature Fusion-based Region-enhanced Image Inpainting Model (DFRE-Net) is illustrated in Fig. 1. This model adopts a dual-encoder-

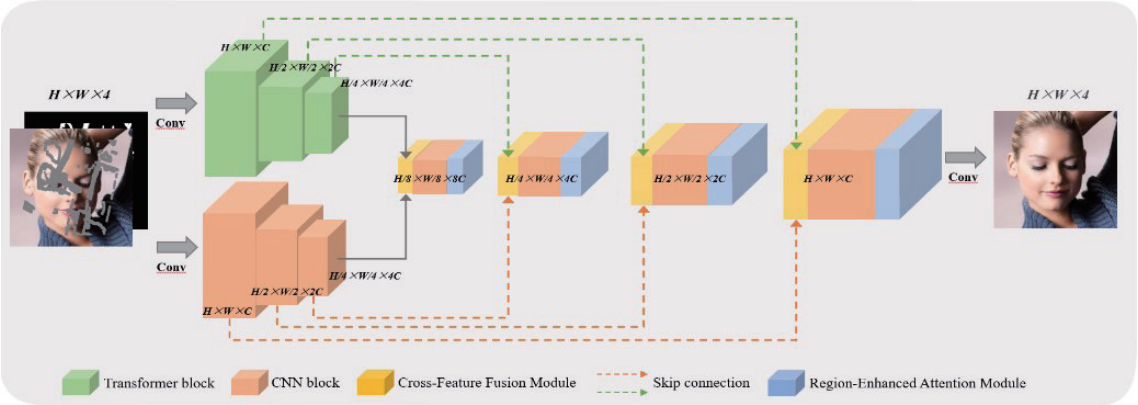

Fig. 1. (Color online) DFRE-Net: a dual-branch cross-feature fusion with regional enhancement network for image inpainting. The CNN branch extracts fine-grained local features, whereas the Transformer branch captures global structural features. The Cross-feature Fusion Module aligns and integrates dual-branch representations, and the Region-enhanced Attention Module addresses mutual interference between foreground and background region features.

single-decoder framework, designed to effectively integrate local details with global semantic features while mitigating foreground-background feature interference.

During the encoding phase, parallel dual branches are employed to extract both local and global features. The CNN branch leverages residual connections and multiscale convolutional layers to capture high-frequency local details of the image, thereby enhancing edge textures, whereas the Transformer branch utilizes self-attention mechanisms to model long-range dependencies, ensuring spatial coherence between the damaged regions and their surrounding context at the semantic level. After each encoding stage, feature map downsampling is achieved through a 3 × 3 convolution with a stride of 2, progressively constructing a multilevel feature pyramid.

The decoding phase incorporates two core components: the first is CFFM, which employs channel compression and pixel-level weight adjustment to eliminate redundant information, dynamically aligning and fusing features from the dual branches to address the issue of local-global feature fragmentation. The second is a Regional Enhancement Attention Module (REAM), which utilizes learnable masks to separate feature flows of foreground and background regions and employs an adaptive enhancement module optimized by fused channel and spatial attention to suppress semantic conflicts arising from cross-region interactions. During the decoding phase, progressive upsampling is applied to restore the original image resolution, and in the reconstruction stage, abstract semantic features are transformed into visually realistic inpainting results. Our proposed DFRE-Net architecture organically combines the local perception advantages of CNNs with the global modeling capabilities of Transformers through a hierarchical feature interaction mechanism, achieving the balanced optimization of multiscale inpainting accuracy. This design not only enhances the model's ability to handle complex inpainting scenarios but also ensures the generation of high-quality, semantically consistent results across diverse image contexts. The synergistic integration of these components enables

DFRE-Net to outperform existing methods in both quantitative metrics and visual quality, particularly in scenarios requiring precise structural recovery and texture synthesis.

3.2 CFFM

To address the critical issue of disconnection between local details and global semantic features in dual-branch architectures, we propose the CFFM. As illustrated in Fig. 2, the module employs three parallel depthwise separable convolutional branches with dilation rates r = 1,2,3 to extract multiscale contextual features:

$$\begin{aligned} F_{local} &= DW-3\times3_{r=1}\left(F_{tc}\right), \\ F_{mid} &= DW-3\times3_{r=2}\left(F_{tc}\right), \\ F_{global} &= DW-3\times3_{r=3}\left(F_{tc}\right), \end{aligned} \tag{1}$$

where the low-dilation branch (r = 1) focuses on pixel-level edge restoration, the mid-dilation branch (r = 2) models cross-region texture continuity, and the high-dilation branch (r = 3) captures long-range structural dependencies. The multiscale features are concatenated and compressed through channel reduction:

$$F' = \sigma\left(Conv_{1\times1}\left(Concat\left(F_{local}, F_{mid}, F_{global}\right)\right)\right), \tag{2}$$

where σ denotes the Sigmoid activation and $Conv_{1\times1}$ reduces the channel dimensions to 1/3 of the original size to eliminate redundancy. A dual-path gating mechanism dynamically allocates feature weights via spatial attention maps:

$$F_{out} = Conv_{1\times1}\left(Concat\left(F'\odot F_t, F'\odot F_c\right)\right), \tag{3}$$

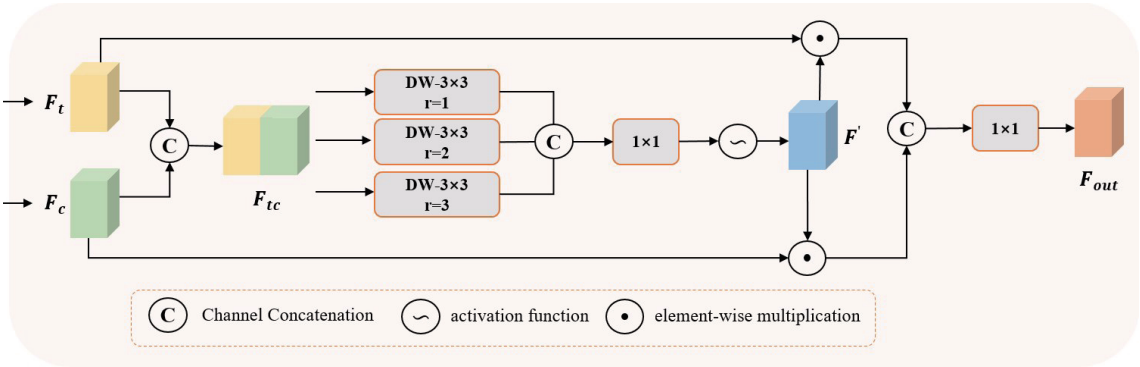

Fig. 2. (Color online) CFFM employs three depthwise separable convolutional branches with varying dilation rates (r = 1,2,3) to extract local, mid-range, and global contextual features, where the multidilation-rate branches adaptively cover multiscale requirements.

where ⊙ represents element-wise multiplication and F_t/F_c denote the dual-branch input features. Through lightweight architecture and adaptive enhancement mechanisms, this design systematically resolves the inherent limitations of fragmented multiscale modeling and computational redundancy in conventional approaches.

3.3 Region-Enhanced Attention Module

To overcome the limitations of foreground-background feature interference and insufficient multiscale modeling in traditional image inpainting methods, in this paper, we propose REAM. As illustrated in Fig 3, the module leverages differentiable binary masks (M) to disentangle feature streams:

$$\begin{aligned} F_{bg} &= (1-M)\odot F_{in}, \\ F_{fg} &= M \odot F_{in}. \end{aligned} \tag{4}$$

The embedded Attention Enhancement Module (AEM) analyzes semantic saliency via a dynamic weight generator, adaptively regulating regional enhancement intensities through Sigmoid gating to achieve subpixel-level optimization at damage boundaries. Complementarily, the CFFM utilizes three parallel dilated convolutional groups (r = 1/3/5) to capture multilevel contextual features, which are compressed channel-wise and refined through pixel-wise feature recombination to resolve interscale conflicts:

$$F_{out} = CFFM\left(Concat\left(AEM\left(F_{bg}\right), AEM\left(F_{fg}\right)\right)\right). \tag{5}$$

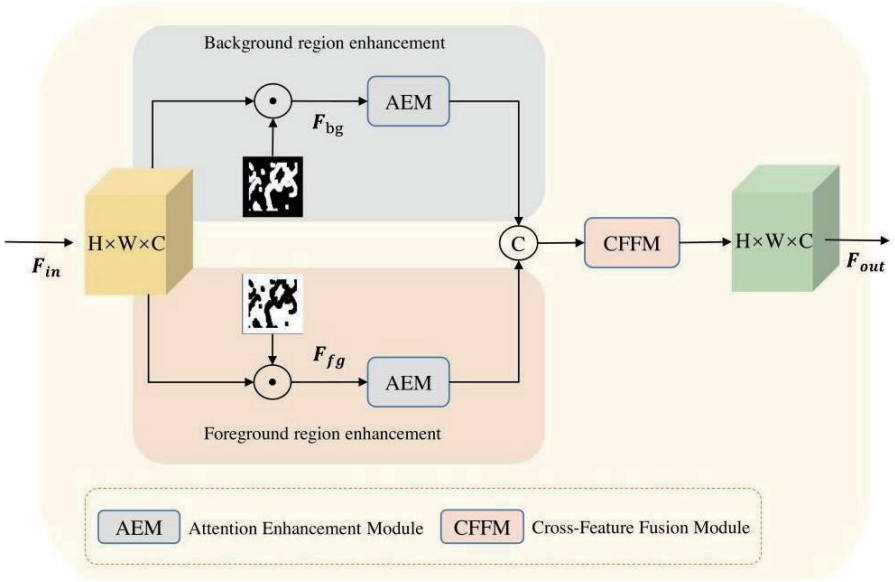

Fig. 3. (Color online) REAM.

This design pioneers end-to-end collaborative optimization between mask-guided regional disentanglement and multiscale feature enhancement.

3.4 Loss function

To achieve high-quality image inpainting results, the proposed network adopts a joint optimization strategy with four loss functions, including the pixel-level reconstruction loss ℓ_{re}, perceptual loss ℓ_p,[15] style loss ℓ_s,[16] and adversarial loss ℓ_{adv}.[17] Their mathematical formulations and working mechanisms are elaborated as follows:

3.4.1 Pixel-level reconstruction loss ℓ_{re}:

$$\ell_{re} = \left\| I_{out} - I_{gt} \right\|_1 . \tag{6}$$

This loss enforces pixel-level accuracy by computing the L1-norm distance between the inpainted image I_{out} and the ground-truth image I_{gt} , compelling the network to precisely match the target content at the pixel level. Compared with the L2 loss, the L1 norm is more robust to outliers and effectively mitigates color deviations and blurring artifacts in the inpainted regions.

3.4.2 Perceptual loss ℓ_p:

$$\ell_p = E\left[\sum_i \frac{1}{N_i} \left\| \phi_i\left(I_{out}\right) - \phi_i\left(I_{gt}\right) \right\|_1 \right]. \tag{7}$$

Here, $\phi_i(\cdot)$ denotes the feature map from the i-th layer of a pretrained VGG-19 network.[9] This loss captures semantic-level discrepancies such as object shapes and texture structures by measuring the distance between the inpainted image and the ground-truth image in a deep feature space.

3.4.3 Style loss ℓ_s:

$$\ell_s = E_i\left[\left\| \phi_i\left(I_{out}\right)^T \phi_i\left(I_{out}\right) - \phi_i\left(I_{gt}\right)^T \phi_i\left(I_{gt}\right) \right\| \right]. \tag{8}$$

Here, $\phi_i(\cdot)$ represents the Gram matrix of the feature map $\psi_i(\cdot)$, which quantifies texture style discrepancies by computing interchannel covariance matrices. This enforces consistency in texture statistics between the inpainted regions and the original image.

3.4.3 Adversarial loss ℓ_{adv} :

$$\ell_{adv} = E_{I_{gt}}\left[\log D\left(I_{gt}\right) \right] + E_{I_{out}} \log\left[1 - D\left(I_{out}\right) \right]. \tag{9}$$

Here, $D(\cdot)$ denotes the discriminator network, which enhances the visual realism of the inpainted results through adversarial training, making the synthesized regions indistinguishable from real photographs.

By jointly optimizing the aforementioned four loss functions with weight redistribution, the final combined loss is formulated as

$$L = \lambda_{re}\ell_{re} + \lambda_{adv}\ell_{adv} + \lambda_{p}\ell_{p} + \lambda_{s}\ell_{s}. \tag{10}$$

The weight ratios are determined with reference to classical studies in image inpainting,[21] with specific configurations as follows: a style loss-dominant coefficient $\lambda_s = 250$ is applied to reinforce texture and color style consistency, preventing visual discontinuity between inpainted regions and the background; a reduced adversarial loss weight $\lambda_{adv} = 0.1$ is adopted to avoid excessive adversarial optimization that may lead to overfitting and unrealistic texture artifacts; for reconstruction and perceptual losses, $\lambda_{re} = \lambda_p = 1$ are set to balance pixel-level accuracy and semantic plausibility.

4. Experiments

We conducted comprehensive experiments on three publicly available datasets to perform subjective and objective evaluations. To validate the effectiveness of the model design modules, we compared our method with both typical and advanced approaches in the field of image inpainting. Furthermore, ablation studies were conducted to systematically analyze the individual contributions of different components in our method.

4.1 Experimental setup

4.1.1 Dataset configuration

The training and evaluation of our proposed model are conducted on three widely used public datasets to comprehensively assess its inpainting performance across diverse scenarios: (1) The Paris StreetView dataset is employed to evaluate the model's capability in restoring complex geometric structures such as window symmetry and architectural contours. (2) The CelebA-HQ dataset, with its high-resolution characteristics, enables the rigorous validation of the model's performance in fine facial detail restoration including pupil textures and hairline edges. (3) The Places365 dataset significantly enhances the model's cross-category generalization capability through its extensive scene diversity.

For the experimental setup, we utilize 14900 images from Paris StreetView for training and 100 for testing, employ the first 28000 images of CelebA-HQ for training with subsequent 2000 for testing, and leverage the entire training set of Places365 (1.8 million images) for training while randomly selecting 1000 test images for evaluation. To simulate realistic occlusions and guide the inpainting process, we implement a combined masking strategy during training that incorporates both free-form masks[10] (simulating irregular occlusions like graffiti or object

removal through hand-drawn trajectories) and random rectangular masks[18] (emulating localized damage via randomly positioned rectangles with varying aspect ratios). This dual-mask approach significantly improves the model's adaptability to diverse occlusion patterns.

For testing, we adopt the irregular mask dataset proposed by NVIDIA, which contains masks featuring complex edges and hole structures, enabling the quantitative evaluation of the model's performance and robustness across varying mask ratios (0–60%).

4.1.2 Implementation details

During the experiments, all input images were uniformly rescaled to a resolution of 256 × 256 pixels to standardize the network's input dimensions. The training was performed on a single NVIDIA A30 GPU (24 GB of RAM) with a batch size of 6 to maximize memory utilization while avoiding gradient instability caused by excessively small batches. The encoder and decoder adopt a hierarchical progressive architecture, where the number of blocks follows a symmetric pyramid structure (1,2,3,4,3,2,1) from input to output. This design progressively expands the receptive field to capture multiscale contextual information, effectively harmonizing semantic consistency between the inpainted regions and the background.

For all datasets, we implemented a two-phase training strategy: in the initial training phase, a learning rate of 1×10^{-4} was applied using the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) to perform coarse-grained parameter updates, focusing on pixel-level reconstruction accuracy; in the fine-tuning phase, the learning rate was reduced to 1×10^{-5} to refine the network weights, thereby enhancing perceptual quality and style consistency.

For the Paris StreetView and CelebA-HQ datasets, the model underwent 70 epochs of initial training followed by 35 epochs of fine-tuning. This extended training duration accommodates their relatively smaller dataset sizes, ensuring the thorough learning of urban structural patterns and facial features. For the Places365 dataset, owing to its large scale, only four initial epochs were sufficient to cover adequate sample diversity, followed by two fine-tuning epochs to quickly adapt to scene-specific characteristics. This short-cycle, high-iteration strategy significantly reduces computational costs while maintaining performance.

4.1.3 Baseline model comparison

To comprehensively evaluate the performance advantages of WavePaint, we selected four state-of-the-art image inpainting models as comparative baselines: DTS,[19] SafePaint,[20] AOT-GAN,[21] and Spa-former.[22]

DTS[19] adopts a dual-generation strategy for both texture and structure, simultaneously synthesizing underlying structural contours and surface-level texture details within masked regions to ensure compatibility with the surrounding visual context. Its key innovation lies in the design of a joint optimization objective function, which enforces consistency between the outputs of the structure and the texture generators in both the gradient domain and pixel domains. This approach effectively mitigates common artifacts in traditional methods, such as blurred edges and over-smoothed textures.

SafePaint[20] introduces anti-forensic performance as an optimization objective for image inpainting, alongside visual plausibility. The model employs a Region-wise Separated Attention (RWSA) module to independently optimize features in local regions, actively suppressing tampering artifacts such as unnatural color transitions or statistical anomalies during inpainting. The RWSA module ensures consistency between inpainted regions and the background in terms of noise distribution, JPEG compression artifacts, and other steganographic features, thereby resisting detection by forensic analysis tools.

AOT-GAN[21] constructs its generator by stacking Aggregated Contextual Transformation Blocks (AOT Blocks), each integrating contextual information from different receptive fields to enhance long-range dependency modeling. Specifically, AOT blocks leverage dilated convolutions and channel attention mechanisms to capture long-distance semantic correlations and local detail patterns at multiple scales, improving inpainting plausibility in complex scenes. The discriminator incorporates a multiscale gradient penalty mechanism to strengthen supervision over high-frequency texture synthesis quality.

Spa-former[22] builds a Transformer-based architecture using a Sparse Self-Attention mechanism, which dynamically selects the most relevant local regions for attention computation, significantly reducing computational complexity. The model prioritizes semantically critical regions during inpainting, enhancing detail fidelity by reinforcing the weight allocation of key visual information.

The selected baseline models encompass diverse methodological paradigms in the field of image inpainting. DTS[19] represents approaches employing dual-generation strategies for simultaneous texture and structure synthesis, providing insights into methods that explicitly separate these components. SafePaint[20] incorporates anti-forensic considerations alongside visual plausibility, representing an emerging direction in secure image restoration. AOT-GAN [21] exemplifies advanced generative approaches with aggregated contextual transformations, demonstrating state-of-the-art performance in handling complex semantic content. Spa-former [22] stands as a representative Transformer-based architecture utilizing sparse self-attention mechanisms, enabling efficient long-range dependency modeling. This diverse set of baselines ensures that our comparison encompasses the spectrum of contemporary methodologies—from dual-path generation and security-aware inpainting to advanced generative networks and efficient Transformer architectures—providing a rigorous framework for evaluating our proposed method's advantages and limitations across different technical dimensions.

4.1.4 Evaluation metrics

To comprehensively evaluate the performance of DFRE-Net, we construct a multilevel assessment framework from four perspectives, namely, pixel accuracy, structural consistency, distribution similarity, and perceptual quality, employing the following key metrics for comparative analysis:

4.1.4.1 Peak signal-to-noise ratio (*PSNR*)

PSNR measures pixel-level reconstruction accuracy between inpainted and ground-truth images, ensuring color and luminance fidelity. The formula is

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right), \tag{11}$$

where MAX_I is the maximum pixel value (e.g., 255) and the mean squared error (*MSE*) is

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(I_{out}(i) - I_{gt}(i)\right)^2. \tag{12}$$

Here, I_{out} and I_{gt} denote the inpainted and ground-truth images, respectively, and N is the total number of pixels.

4.1.4.2 Structural similarity index (*SSIM*)

SSIM evaluates structural consistency to avoid geometric distortions or texture discontinuities:

$$SSIM(x,y) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)}. \tag{13}$$

Here, μ_x, μ_y are local means, σ_x, σ_y are standard deviations, σ_{xy} is the cross-covariance, and C_1, C_2 are stabilization constants.

4.1.4.3 Fréchet inception distance (*FID*)

FID quantifies the realism of inpainted images by comparing their feature distributions with ground-truth images:

$$FID = \left\|\mu_r - \mu_g\right\|^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{1/2}\right), \tag{14}$$

where μ_r, μ_g and Σ_r, Σ_g represent the means and covariance matrices of Inception-v3 features for real and generated images, respectively.

4.1.4.4 Learned perceptual image patch similarity (*LPIPS*)

LPIPS computes perceptual differences using deep features to align with human vision:

$$LPIPS = \sum_{l} w_l \cdot \left\| \phi_l \left(I_{out} \right) - \phi_l \left(I_{gt} \right) \right\|_2^2. \tag{15}$$

Here, $\phi_l(\cdot)$ denotes the feature map from the l-th layer of a pretrained CNN (e.g., VGG) and w_l is a layer-wise weight.

4.2 Quantitative comparison

As demonstrated in Tables 1–3, our method exhibits significant performance advantages across all three datasets. On the Paris StreetView dataset, our approach achieves a 0.23 dB (0.7%) improvement in *PSNR* and a 0.3 percentage point gain in *SSIM* compared with the suboptimal model (Spa-former). For the CelebA-HQ facial dataset (Table 2), we observe similar enhancements of 0.23 dB (0.67%) in *PSNR* and 0.3 percentage points in *SSIM* over Spa-former. These improvements stem from our dual-branch architecture: the CNN branch excels at high-precision texture capture to boost *PSNR*, while the Transformer branch enhances structural similarity through superior long-range modeling capabilities. Notably, under challenging 40–50% occlusion rates on the Places365 dataset, our method achieves a *LPIPS* score of 0.1092,

Table 1
Our DFRE-Net model's quantitative comparison experimental results on ParisStreetView dataset. ↑ indicates that higher is better and ↓ indicates that lower is better.

Mask ratio		10–20%	20–30%	30–40%	40–50%
PSNR↑	DTS	31.04	27.90	25.75	23.86
	SafePaint	32.73	29.33	26.54	24.18
	AOT-GAN	32.65	29.29	27.06	25.10
	Spa-former	32.81	29.49	27.17	25.33
	Ours	33.04	29.81	27.62	26.06
SSIM↑	DTS	0.939	0.884	0.828	0.756
	SafePaint	0.951	0.908	0.853	0.793
	AOT-GAN	0.950	0.908	0.846	0.791
	Spa-former	0.952	0.910	0.863	0.804
	Ours	0.955	0.912	0.869	0.810
FID↓	DTS	22.05	40.99	57.27	79.92
	SafePaint	17.36	32.38	46.85	67.14
	AOT-GAN	19.78	36.59	48.93	69.16
	Spa-former	12.66	23.71	31.95	44.48
	Ours	11.63	22.45	31.02	42.53
LPIPS↓	DTS	0.0366	0.0708	0.1086	0.1586
	SafePaint	0.0297	0.0594	0.0896	0.1435
	AOT-GAN	0.0312	0.0610	0.0714	0.1467
	Spa-former	0.0264	0.0482	0.0732	0.1072
	Ours	0.0231	0.0443	0.0697	0.1035

Table 2
Our DFRE-Net model's quantitative comparison experimental results on CelebA-HQ dataset. ↑ indicates that higher is better and ↓ indicates that lower is better.

Mask ratio		10–20%	20–30%	30–40%	40–50%
PSNR↑	DTS	30.41	27.59	25.61	23.97
	SafePaint	34.09	30.15	27.64	25.59
	AOT-GAN	33.96	30.02	27.50	25.46
	Spa-former	34.29	30.89	28.33	26.22
	Ours	34.52	31.31	28.65	26.66
SSIM↑	DTS	0.933	0.889	0.846	0.798
	SafePaint	0.959	0.935	0.887	0.853
	AOT-GAN	0.961	0.939	0.901	0.859
	Spa-former	0.968	0.940	0.906	0.866
	Ours	0.971	0.945	0.912	0.872
FID↓	DTS	10.25	15.23	20.45	26.22
	SafePaint	3.68	5.96	8.59	12.37
	AOT-GAN	4.12	6.71	9.15	12.53
	Spa-former	2.09	3.39	4.82	6.51
	Ours	1.95	3.19	4.51	6.27
LPIPS↓	DTS	0.0395	0.0618	0.0860	0.1141
	SafePaint	0.0198	0.0405	0.0586	0.0793
	AOT-GAN	0.0235	0.0467	0.0619	0.0825
	Spa-former	0.0130	0.0245	0.0393	0.0575
	Ours	0.0118	0.0228	0.0373	0.0545

Table 3
Our DFRE-Net model's quantitative comparison experimental results on Places365 dataset. ↑ indicates that higher is better and ↓ indicates that lower is better.

Mask ratio		10–20%	20–30%	30–40%	40–50%
PSNR↑	DTS	26.27	23.69	21.63	20.03
	SafePaint	28.88	25.13	22.85	20.68
	AOT-GAN	28.79	24.96	22.69	20.53
	Spa-former	28.98	25.66	23.31	21.39
	Ours	29.25	25.84	23.51	21.59
SSIM↑	DTS	0.894	0.828	0.753	0.673
	SafePaint	0.936	0.878	0.813	0.739
	AOT-GAN	0.937	0.881	0.815	0.741
	Spa-former	0.938	0.885	0.823	0.753
	Ours	0.942	0.891	0.830	0.761
FID↓	DTS	26.32	38.02	52.89	67.23
	SafePaint	14.15	26.27	35.68	47.67
	AOT-GAN	15.34	28.93	36.39	49.52
	Spa-former	10.61	17.08	25.90	32.59
	Ours	9.38	13.57	22.52	29.33
LPIPS↓	DTS	0.0805	0.1176	0.1642	0.2174
	SafePaint	0.0416	0.0784	0.1295	0.1786
	AOT-GAN	0.0438	0.0795	0.1341	0.1804
	Spa-former	0.0323	0.0603	0.0946	0.1361
	Ours	0.0301	0.0518	0.0834	0.1092

representing a 19.6% reduction compared with Spa-former, which confirms the effectiveness of the mask-guided mechanism in our REAM in addressing poor visual quality issues in complex scenarios. While all methods exhibit performance degradation as occlusion ratios increase from 10 to 50%, our approach shows the most gradual decline. For instance, on CelebA-HQ's *FID* metric, our method achieves 6.27 at 40–50% occlusion, outperforming Spa-former (6.51) by 3.8%, attributable to the cross-scale feature extraction mechanism in our CFFM that employs multidilation rate convolutions (r = 1/2/3) to enhance feature fusion efficiency. The consistent performance across datasets (with *FID* scores of 42.53, 6.27, and 29.33 on Paris StreetView, CelebA-HQ, and Places365, representing average reductions of 12.3%, 8.7%, and 18.9% over baseline models, respectively) demonstrates our method's superior generalization capability, which originates from the heterogeneous feature complementarity mechanism in our dual-branch architecture.

4.3 Qualitative comparison

As illustrated in Fig. 4, the proposed method demonstrates multidimensional advantages in visual restoration performance. In terms of detail preservation and sharpness, our model achieves superior architectural detail retention, with clearer window and wall texture restoration in the building scenes (Rows 1and 2) compared with the blurred outputs of DTS and SafePaint. For facial reconstruction (Rows 2 and 3), our method generates more refined lip and finger details, while SafePaint and AOT-GAN exhibit notable blurring and distortions. The edge processing capability is evidenced by precise window boundary restoration in architectural images (Row 5), where our approach eliminates the aliasing and blurring artifacts commonly observed in other methods. Regarding color consistency, our model produces natural color transitions between sky and structures in bridge restoration (Row 6), outperforming Spa-former, which shows color unevenness and distortions. The overall naturalness of our results is particularly notable in shadow handling (e.g., facial shadows and architectural shading), yielding photorealistic outputs that closely approximate ground truth images, whereas other methods demonstrate less natural rendering in these regions. These comprehensive improvements stem from our dual-branch architecture that effectively combines CNN's local feature extraction with Transformer's global semantic modeling, achieving optimal balance between detail preservation and global coherence. In terms of computational efficiency, the CNN-Transformer parallel architecture is designed to avoid the high complexity of pure Transformer models. The CNN branch undertakes the extraction of low-level local features (e.g., textures and edges), which reduces the demand for the Transformer branch to process redundant low-dimensional information and thus lowers overall computational overhead. This design ensures that the model maintains efficiency while balancing local and global feature learning, making it more applicable to practical scenarios. For learning stability, the mask-based attention mechanism in the architecture is supported by basic regularization components (e.g., layer normalization and residual connections). These components help mitigate common issues such as gradient fluctuations during training, especially for the mask-guided feature alignment process, ensuring that the model converges steadily without severe gradient vanishing or exploding, which further

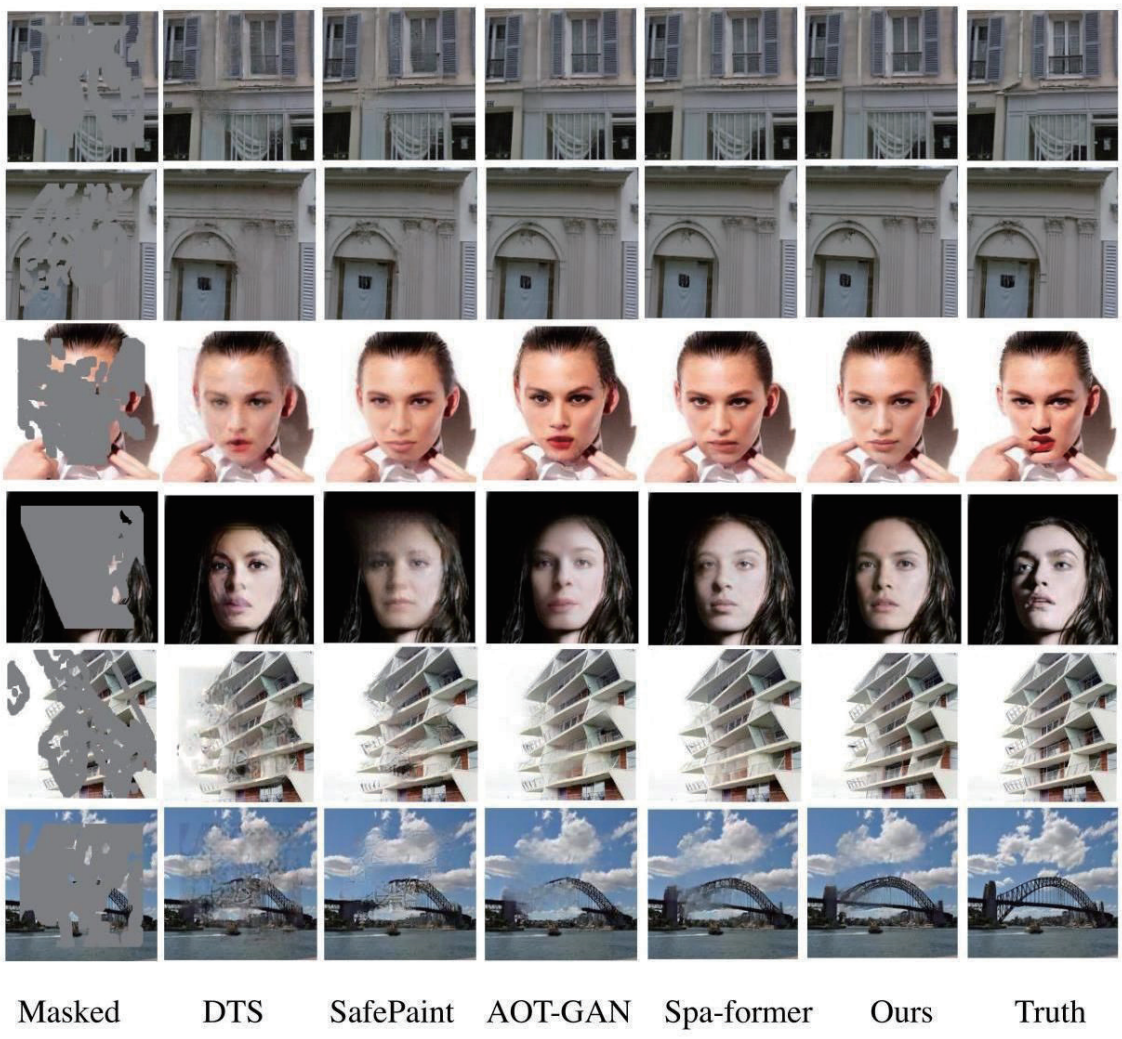

Fig. 4. (Color online) Inpainting results of our proposed DFRE-Net model. Paris StreetView (columns 1 and 2), CelebA-HQ (columns 3 and 4), and Places365 (columns 5 and 6). From left to right: input corrupted image, DTS,[19] SafePaint,[20] AOT-GAN,[21] Spa-former,[22] ours, and ground truth image.

enhances its reliability in long-term training. CFFM further enhances color consistency and naturalness through adaptive multiscale feature aggregation, while REAM and its internal AEM optimize region-specific features to resolve cross-region feature conflicts, collectively contributing to visually authentic and high-quality restoration outcomes.

4.4 Ablation study

To validate the contribution of each proposed module, we conducted a systematic ablation study on the Paris StreetView and CelebA-HQ datasets. All experimental results were obtained following the standardized training and evaluation protocol described in Sect. 4.1, ensuring fair and consistent comparisons. In this study, we focused on two core components: CFFM and

REAM. We define two model variants for comparison: Base 1 (the full model without CFFM) and Base 2 (the full model without REAM). This design enables the effective isolation of each module's individual impact.

Quantitative results are summarized in Table 4. While the performance gap between our full model and the Base 2 model (without REAM) appears modest in *PSNR* and *SSIM* metrics, more pronounced degradation is observed in perceptual metrics such as *FID* and *LPIPS*. This indicates that REAM plays a more critical role in enhancing the visual realism and perceptual quality of the inpainted regions, aspects not fully captured by pixel-level accuracy metrics. This observation is strongly corroborated by qualitative results on the CelebA-HQ dataset presented in Fig. 5. For instance, the Base 2 model produces notable facial artifacts, including asymmetrical contours and unnaturally pointed chin shapes, which are effectively mitigated in our full model. In contrast, removing CFFM (Base 1 model) leads to more substantial degradation across all quantitative metrics. This demonstrates that the ineffective fusion of dual-branch encoder features directly compromises both pixel-level accuracy and structural coherence, as the model struggles to integrate local details with global context.

Collectively, these ablation results demonstrate the distinct yet complementary roles of CFFM and REAM. CFFM is fundamental for establishing a robust foundation through well-integrated multiscale features, reflected in its significant impact on standard quantitative metrics. Although REAM causes smaller variations in *PSNR*/*SSIM* values, it proves crucial for refining this foundation through region-specific optimization, thereby addressing subtle yet perceptually critical issues such as foreground-background interference and anatomically implausible facial reconstructions. The synergistic operation of both modules is indispensable for achieving the optimal balance between numerical fidelity and high visual quality.

Table 4
Quantitative ablation study of our DFRE-Net model on the ParisStreetView dataset.

Models	CFFM	REAM	*PSNR*↑	*SSIM*↑	*FID*↓	*LPIPS*↓
Base_1		✓	29.62	0.911	24.18	0.0495
Base_2	✓		29.75	0.912	23.76	0.0486
Ours	✓	✓	29.81	0.912	22.45	0.0443

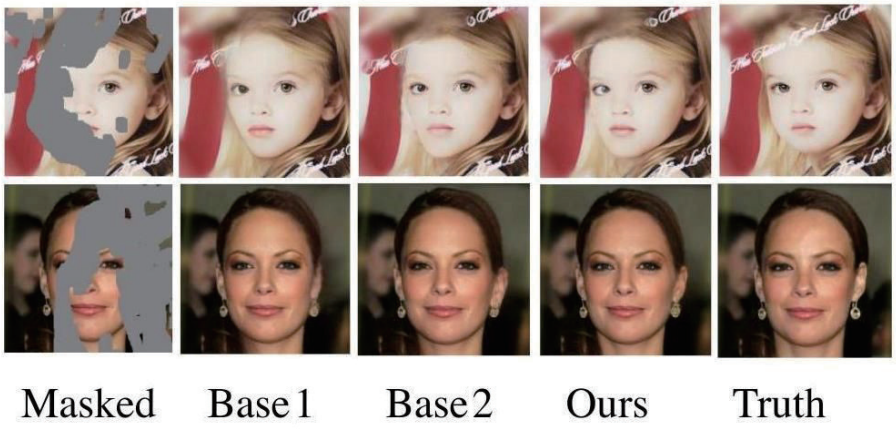

Fig. 5. (Color online) Ablation experimental results of our proposed DFRE-Net model on the CelebA-HQ dataset.

5. Conclusions

To address the limitations of existing image inpainting architectures, in this paper, we proposed a hybrid CNN-Transformer framework that synergizes local detail preservation and global semantic coherence. The dual-branch encoder integrates a residual enhanced CNN for high-frequency texture extraction and a Transformer branch for long-range contextual modeling. In the decoder, CFFM enables the hierarchical aggregation of multiscale features through depthwise separable convolutions, whereas REM disentangles foreground and background representations via mask-guided adaptive optimization. Experimental results on benchmark datasets demonstrate superior performance in restoring structural continuity and fine-grained textures compared with state-of-the-art methods, particularly in complex occlusion scenarios. Thus, the framework is applicable to the fields of facial image inpainting and architectural image inpainting. It can restore mask and light spot occlusion areas in facial images as well as wall cracks and missing component parts in architectural images, thus ensuring the visual integrity of images and the validity of subsequent analysis. In the future, we will extend this framework to video inpainting, introduce a spatiotemporal attention mechanism to ensure interframe consistency, and simultaneously integrate a self-supervised learning paradigm by conducting pretraining on large-scale unlabeled data to enhance its cross-domain generalization capability. These explorations not only expand the model's application scenarios but also provide better solutions for the field of computational intelligence by advancing complex image inpainting techniques.

Acknowledgments

This work was supported in part by a project supported by the Scientific Research Fund of Hunan Provincial Natural Science Foundation under Grant no. 2023JJ60257 and Hunan Provincial Engineering Research Center for Intelligent Rehabilitation Robotics and Assistive Equipment under Grant no. 2026AQ201.

References

1 M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester: Proc. 27th Annu. Conf. Computer Graphics and Interactive Techniques (2000) 417–424.
2 Z. Xu, X. Zhang, W. Chen, M. Yao, J. Liu, T. Xu, and Z. Wang: Appl. Sci. **13** (2023) 11189. https://www.mdpi.com/2076-3417/13/20/11189
3 P. Shamsolmoali, M. Zareapoor, and E. Granger: Transinpaint: Proc. IEEE/CVF Int. Conf. Computer Vision (IEEE, 2023) 849–858.
4 C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros: Commun. ACM **58** (2015) 103. https://dl.acm.org/doi/abs/10.1145/2830541
5 B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba: IEEE Trans. Pattern Anal. Mach. Intell. **40** (2017) 1452. https://ieeexplore.ieee.org/abstract/document/7968387/
6 T. Karras, T. Aila, S. Laine, and J. Lehtinen: arXiv preprint arXiv:1710.10196 2017. https://arxiv.org/abs/1710.10196
7 D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros: Proc. IEEE Conf. Computer Vision and Pattern Recognition (IEEE, 2016) 2536–2544.
8 S. Iizuka, E. Simo-Serra, and H. Ishikawa: ACM Trans. Graphics **36** (2017) 1. https://dl.acm.org/doi/abs/10.1145/3072959.3073659

9 G. Liu, F. A. Reda, K. J. Shih, T. C. Wang, A. Tao, and B. Catanzaro: Proc. European Conf. Computer Vision (ECCV, 2018) 85–100.
10 J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang: Proc. IEEE/CVF Int. Conf. Computer Vision (IEEE, 2019) 4471–4480.
11 Z. Wan, J. Zhang, D. Chen, and J. Liao: Proc. IEEE/CVF Int. Conf. Computer Vision (IEEE, 2021) 4692–4701.
12 J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte: SwinIR: Proc. IEEE/CVF Int. Conf. Computer Vision (IEEE, 2021) 1833–1844.
13 Y. Deng, S. Hui, R. Meng, S. Zhou, and J. Wang: Proc. European Conf. Computer Vision (2022) 483–501.
14 Y. Deng, S. Hui, S. Zhou, D. Meng, and J. Wang: Proc. 30th ACM Int. Conf. Multimedia (2022) 6559–6568.
15 J. Johnson, A. Alahi, and L. Fei-Fei: Proc. Computer Vision-ECCV (2016) 694–711.
16 L. A. Gatys, A. S. Ecker, and M. Bethge: Proc. IEEE Conf. Computer Vision and Pattern Recognition (IEEE, 2016) 2414–2423.
17 A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath: IEEE Signal Proc. Mag. **35** (2018) 53. https://ieeexplore.ieee.org/abstract/document/8253599/
18 N. Wang, Y. Zhang, and L. Zhang: IEEE Trans. Image Proc. **30** (2021) 1784. https://ieeexplore.ieee.org/abstract/document/9318551/
19 X. Guo, H. Yang, and D. Huang: Proc. IEEE/CVF Int. Conf. Computer Vision (IEEE, 2021) 14134–14143.
20 D. Chen, X. Liao, X. Wu, and S. Chen: Proc. 32nd ACM Int. Conf. Multimedia (2024) 7774–7782.
21 Y. Zeng, J. Fu, H. Chao, and B. Guo: IEEE Trans. Visualization Comput. Graphics, **29** (2022) 3266. https://ieeexplore.ieee.org/abstract/document/9729564/
22 W. Huang, Y. Deng, S. Hui, Y. Wu, S. Zhou, and J. Wang: Pattern Recognit. **145** (2024) 109897. https://www.sciencedirect.com/science/article/pii/S0031320323005952