Hierarchical Contrastive Learning for Pattern-Generalizable Image Corruption Detection

Xin Feng* Yifeng Xu* Guangming Lu† Wenjie Pei†
Harbin Institute of Technology, Shenzhen

Figure 1: Visual comparison between our method and VCNet [42] (the state-of-the-art method for blind image inpainting) on four tasks: blind image inpainting (graffiti corruption) (**up left**), blind image inpainting (real-image-patch corruption) (**up right**), watermark removal (**bottom left**) and shadow removal (**bottom right**). The groundtruth mask and the predicted masks by different methods are presented.

Abstract

Effective image restoration with large-size corruptions, such as blind image inpainting, entails precise detection of corruption region masks which remains extremely challenging due to diverse shapes and patterns of corruptions. In this work, we present a novel method for automatic corruption detection, which allows for blind corruption restoration without known corruption masks. Specifically, we develop a hierarchical contrastive learning framework to detect corrupted regions by capturing the intrinsic semantic distinctions between corrupted and uncorrupted regions. In particular, our model detects the corrupted mask in a coarse-to-fine manner by first predicting a coarse mask by contrastive learning in low-resolution feature space and then refines the uncertain area of the mask by high-resolution contrastive learning. A specialized hierarchical interaction mechanism is designed to facilitate the knowledge propagation of contrastive learning in different scales, boosting the modeling performance substantially. The detected multi-scale corruption masks are then leveraged to guide the corruption restoration. Detecting corrupted regions by learning the contrastive distinctions rather than the semantic patterns of corruptions, our model has well generalization ability across different corruption patterns. Extensive experiments demonstrate following merits of our model: 1) the superior performance over other methods on both corruption detection and various image restoration tasks including blind inpainting and watermark removal, and 2) strong generalization across different corruption patterns such as graffiti, random noise or other image content. Codes and trained weights are available at https://github.com/xyfJASON/HCL.

1. Introduction

An essential yet challenging step for image restoration with large-size corruptions, such as blind image inpainting, watermark removal or shadow removal, is to detect the corruption region masks in pixel level precisely. The difficulties lie in the diverse shapes and appearance of corruptions, which can be hardly modeled in a uniform pattern. As a result, blind image inpainting remains a challenging task although image inpainting with known corruption masks [27, 11, 13, 32] has achieved remarkable progress.

A straightforward way for corruption detection is to for-

*These authors contributed equally to this work.
†Corresponding authors.

mulate it as a segmentation task and predict the corruption mask by pixel-wise binary classification. A prominent example following such modeling paradigm is VCNet [42], a state-of-the-art method for blind image inpainting, which regards corruptions as target objects and learns to recognize the semantic patterns of corruptions for detection. While VCNet can successfully detect the corruptions with uniform patterns, it has two potential limitations. First, the corruption may exhibit diverse patterns due to its potentially irregular appearance nature, which VCNet can hardly deal with. Second, it is challenging for VCNet to handle the corruption patterns that are distinct from those appearing in training data, thus it has limited generalization w.r.t. different corruption patterns.

In this work, we propose a novel model for image corruption detection to circumvent the aforementioned limitations of VCNet. Instead of recognizing the semantic patterns of corruptions as VCNet does, we apply metric learning to learn an embedding space in which our model can capture the contrastive semantic distinctions between the corrupted and uncorrupted regions. To this end, we design a hierarchical contrastive learning framework for detecting corrupted regions, which predicts the corruption mask in a coarse-to-fine manner. To be specific, it first predicts a coarse mask by lightweight contrastive learning in a low-resolution feature space. Then it refines the uncertain pixels with low confidence in the predicted mask by high-resolution contrastive learning based on fine-grained features. Note that only a small fraction of predicted masks need to be refined, thus the refining process can be performed quite efficiently. To facilitate the knowledge propagation and consistency of contrastive learning between different scales, we propose a specialized hierarchical interaction mechanism. As a result, the high-resolution contrastive learning can inherit useful knowledge from the low-resolution contrastive learning to achieve precise prediction of corruption masks quite efficiently. The detected multi-scale corruption masks are further used to guide the restoration of the corrupted regions, which also follows a coarse-to-fine generative process.

Unlike VCNet recognizing the semantic patterns of corruptions, our model focuses on capturing the contrastive semantic distinctions between corrupted and uncorrupted regions. Thus our model has well generalization ability across different corruption patterns, As shown in Figure 1. To conclude, we make following contributions.

- We design a hierarchical contrastive learning framework to detect multi-scale corruption masks by learning the contrastive semantic distinctions between corrupted and uncorrupted regions.
- Integrating the proposed hierarchical contrastive learning module, we develop a general-purpose blind image restoration model to perform high-quality image restoration without known corruption masks.
- Extensive experiments validate the effectiveness of our model both quantitatively and qualitatively in three aspects: 1) superior performance on corruption detection (Section 4.2), 2) favorable performance compared to the specialized methods on two challenging image restoration tasks including blind image inpainting and watermark removal (Section 4.3), and 3) strong generalizability across different corruption patterns (Section 4.2 and 4.3.1).

2. Related Work

Image Restoration with known Corruption masks. Image restoration with given corruption masks assumes that the corrupted regions are known, and conventional methods [2, 12, 5] leverage uncorrupted content to restore the corrupted regions by pixel diffusion or patch replacement. However, these methods usually generate significant artifacts in the restored image due to the lack of global semantic understanding and generalization ability. With the great success of convolutional neural networks(CNNs) in computer vision, recent methods leverage an encoder-decoder framework to restore corrupted images. A typical way of CNN-based methods [29, 44, 26, 27] employs known uncorrupted regions to infer corrupted content from the outside to the inside. To synthesize reasonable semantics, some methods [31, 35, 20, 30, 16, 28] leverage structure information to improve the quality of semantic structures in the corrupted regions. Promoted by generative models [14, 25], some recent methods [11, 32, 47] attempt to learn generative prior for improving the quality of synthesized images, and generating diverse content for corrupted regions.

Image corruption detection. Early methods [3, 33] for image corruption detection assume the corrupted content follows a simple distribution, such as constant values, Gaussian noise, etc. However, previous assumption simplifies the detection of corrupted regions, limiting the application scope of image corruption detection. Recently, Wang *et al.* [42] relax the definition of image corruption detection by increasing the diversity of corruption types, including watermarking, raindrop, or even random images. They also propose a two-stage framework *VCNet* for blind image restoration, which first locates corrupted regions in a manner of image segmentation and then fills in content following the typical encoder-decoder based methods for image restoration. However, it sometimes misuses content in corrupted regions while reconstructing the complete image due to the challenge of detecting diverse patterns of corruption.

Contrastive learning. Contrastive learning has been widely used in self-supervised representation learning [10, 17, 4]. Instead of matching an input image to a fixed target, contrastive learning maximizes mutual information in the representation space, keeping the query close to the positive sample and away from the negative sample. Previous works [17, 4, 41] have validated the effectiveness of contrastive

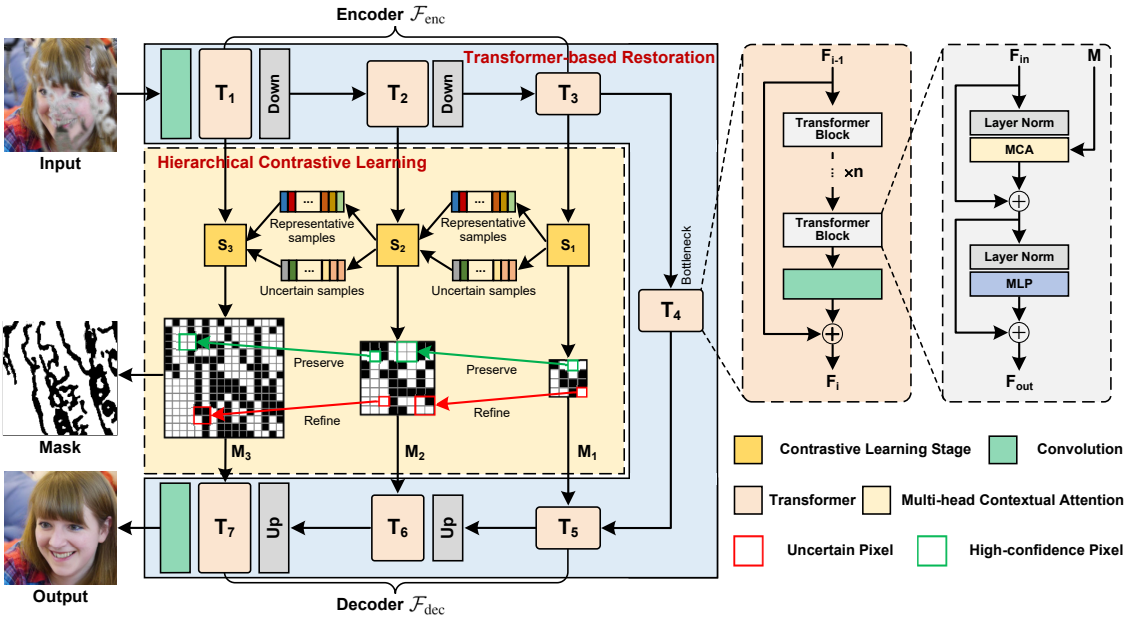


Figure 2: Architecture of our method. It leverages the designed Hierarchical Contrastive Learning to detect corruption masks and further synthesizes reasonable content for corrupted regions by the developed Transformer-based Restoration module.

learning in high-level vision tasks, due to the inherent suitability for modeling feature contrasts. Recently, some researchers [40, 43] have attempted to introduce contrastive learning into low-level vision tasks. However, pixel-wise contrastive learning on high-resolution images suffers from large computational cost and limited performance. In this paper, we design a hierarchical contrastive learning mechanism to solve above problems and significantly strengthen the performance of image corruption detection.

3. Approach

Detecting image corruptions by recognizing the semantic patterns of corruptions is difficult in that the corruptions can potentially exhibit diverse patterns. Our proposed hierarchical contrastive learning framework learns the contrastive semantic distinctions between corrupted and uncorrupted regions, and detects corruption masks in a coarse-to-fine manner. The obtained multi-scale corruption masks are then leveraged to guide the restoration of the corruptions in a coarse-to-fine generative process.

3.1. Overview

Problem formulation. Given an input image $\mathbf{I} \in \mathbb{R}^{C\times H\times W}$ which is corrupted from an intact groundtruth image $\mathbf{O}_{\text{gt}} \in \mathbb{R}^{C\times H\times W}$, the task of blind image restoration aims to first detect the corruption region mask $\mathbf{M} \in \mathbb{R}^{1\times H\times W}$ and then reconstruct a complete image $\hat{\mathbf{O}} \in \mathbb{R}^{C\times H\times W}$ by synthesizing realistic content $\hat{\mathbf{C}}$ for the corrupted regions:

$$\begin{aligned} \mathbf{I} &= \mathbf{O}_{\text{gt}} \odot \mathbf{M} + \mathbf{N} \odot (1 - \mathbf{M}), \\ \hat{\mathbf{O}} &= \mathbf{O}_{\text{gt}} \odot \mathbf{M} + \hat{\mathbf{C}} \odot (1 - \mathbf{M}). \end{aligned} \tag{1}$$

Herein, $\mathbf{M} \in \mathbb{R}^{1\times H\times W}$ is a binary map where the values of corrupted regions are 0 and uncorrupted regions are filled with 1. $\mathbf{N} \in \mathbb{R}^{C\times H\times W}$ is the noisy content in the corrupted regions. Precise detection of the corruption mask $\mathbf{M}$ is crucial to the performance of image restoration.

As shown in Figure 2, our model follows the encoder-decoder framework and consists of two core modules: Hierarchical Contrastive Learning module for detecting corruption masks and Transformer-based Restoration module for corruption restoration. The corrupted image is first encoded into multi-scale feature maps by Encoder $\mathcal{F}_{\text{enc}}$ of Transformer-based Synthesis module. Then the proposed Hierarchical Contrastive Learning module is employed to perform corruption detection in a coarse-to-fine manner from these feature maps, and predicts multi-scale corruption masks. Finally, the obtained masks are fed into Decoder $\mathcal{F}_{\text{Dec}}$ of the Transformer-based Synthesis module for restoration, which is also in a coarse-to-fine generative process. Both Encoder $\mathcal{F}_{\text{enc}}$ and Decoder $\mathcal{F}_{\text{dec}}$ are mainly composed of basic transformer blocks of MAT [39], while they are equipped with additional down-sampling layers and up-sampling layers respectively. We employ a shallow convolutional layer with 5×5 kernel before Encoder and after Decoder to perform basic feature transformation. Besides, we also employ another Conv-U-Net to refine high-frequency details of output results, leaning upon the local texture refinement capability and efficiency of CNNs.

3.2. Corruption Detection by Hierarchical Contrastive Learning

We design Hierarchical Contrastive Learning framework to guide the learning of multi-scale semantic embedding spaces of Encoder $\mathcal{F}_{\text{enc}}$ in such a way that the semantic dis-

tances between two arbitrary pixels both within uncorrupted or corrupted regions (intra-region distance) should be minimized while the distance between a corrupted pixel and an uncorrupted pixel (inter-region distance) should be maximized. As a result, our model can capture the intrinsic semantic distinctions between corrupted and uncorrupted regions. Then our model performs clustering on all pixels in this learned embedding space to separate them into two clusters, corresponding to the corrupted and uncorrupted regions, and thus achieves the corruption mask.

As shown in Figure 2, Encoder $\mathcal{F}_{\text{enc}}$ consists of three encoding stages and produces three scales of encoded feature maps. Accordingly, Our model performs three stages of pixel-level contrastive learning for the corresponding scale of feature maps to guide the learning of its embedding space. We will first describe how our model performs corruption detection in one stage with single-scale contrastive learning. Then we will elaborate on the proposed Hierarchical Interaction Mechanism which enables our model to perform hierarchical contrastive learning quite efficiently for coarse-to-fine multi-scale mask detection.

3.2.1 Single-Scale Contrastive Learning

Supervised Contrastive Learning. During each stage of contrastive learning, we construct positive pixel pairs by using intra-region pixels, i.e., both pixels are from either the uncorrupted region or the corrupted region. In contrast, each negative training pair consists of two inter-region pixels, one from the corrupted region and the other from the uncorrupted region. To be specific, for a query pixel q from a randomly selected query set Q, we construct the positive set P by randomly selecting pixels from the same region as q and construct the negative set N from the opposite region. Then we apply Circle loss [38] to maximize the cosine similarity of positive pairs while minimizing the similarities of negative pairs. Formally, the contrastive learning in the stage S_s is guided by the loss:

$$\mathcal{L}^s_{\text{CL}} = \sum_{q\in Q} \log\left[1 + \sum_{p\in P} \exp(-\mathbf{e}_q \cdot \mathbf{e}_p/\tau) \cdot \sum_{n\in N} \exp(\mathbf{e}_q \cdot \mathbf{e}_n/\tau)\right], \quad (2)$$

where τ is a scale factor. $\mathbf{e}_q$ is the features for pixel q projected from the corresponding embedding space of Encoder $\mathcal{F}^s_{\text{enc}}$ by a projection head in Stage-s:

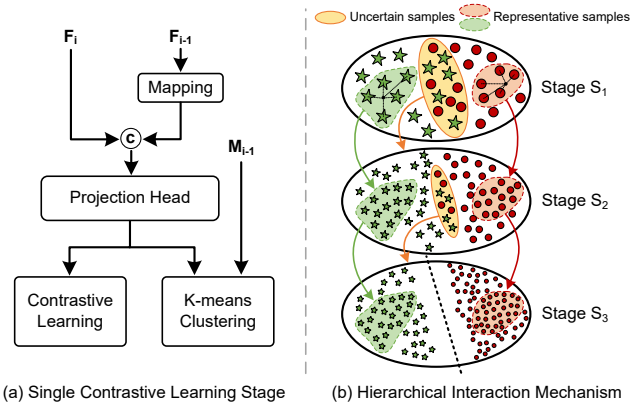
$$\mathbf{e}_q = \mathcal{F}^s_{\text{proj}}(\mathcal{F}^s_{\text{enc}}(\mathbf{I}_q)), \quad (3)$$

where the projection head $\mathcal{F}_{\text{proj}}$ comprises two fully connected layers and a *GELU* [18] layer in-between for nonlinear transformation.

Corruption Mask Detection. Under the supervision of contrastive learning in Equation 2, the pixels within the same region, either the corrupted or the uncorrupted regions, tend to have similar representations in each scale of embedding space of Encoder $\mathcal{F}_{\text{enc}}$ while the pixels from different regions would have dissimilar representations. Thus, we can perform clustering to separate them into two clusters. Specifically, our model adopts K-means algorithm for clustering, and the clustering in Stage-s is performed by:

$$[\mathbf{c}^s_1, \mathbf{c}^s_2], [\mathbf{M}^s_1, \mathbf{M}^s_2] = \text{K-means}(\mathcal{F}^s_{\text{proj}}(\mathcal{F}^s_{\text{enc}}(\mathbf{I}))), \quad (4)$$

where $[\mathbf{c}^s_1, \mathbf{c}^s_2]$ denotes the embeddings of two produced cluster centers and $[\mathbf{M}^s_1, \mathbf{M}^s_2]$ are the associated binary masks. To identify which cluster corresponds to the corrupted or uncorrupted regions, we train a lightweight binary classifier consisting of two fully connected layers with a *ReLU* layer, and use it to perform classification on two cluster centers. Thus, the associated mask for the cluster of the corrupted region is the predicted corruption mask.



(a) Single Contrastive Learning Stage (b) Hierarchical Interaction Mechanism

Figure 3: In the stages S_2 and S_3, both the selected representative samples and uncertain samples in the previous stage are used for contrastive learning in current stage while only the mask labels of uncertain samples are refined during mask detection.

3.2.2 Hierarchical Interaction Mechanism

Our model performs Hierarchical Contrastive Learning including three stages to detect the corruption mask in a coarse-to-fine manner. It first performs contrastive learning in the lowest-resolution stage (S_1) to predict a coarse mask for corruption. Then it refines the uncertain pixels of the mask with low confidence by higher-resolution contrastive learning in subsequent stages (S_2 and S_3). We propose Hierarchical Interaction Mechanism to facilitate the interaction and knowledge propagation between different stages of contrastive learning and guarantee the semantic consistency between them. In particular, higher-resolution contrastive learning can inherit useful knowledge from previous stages of contrastive learning, which can improve the learning performance substantially. Quadtree structure are used to build the positional correspondence between feature maps in adjacent stages considering that the resolution of feature maps is always scaled by four times between adjacent stages.

Selecting High-Quality Training Samples. During the contrastive learning in S_1, we construct the query set Q as well as the associated positive set P and negative set N

in Equation 2 by randomly selecting pixels from the whole feature map. To improve the efficiency of contrastive learning in the higher stages (S_2 and S_3), we select a small fraction of pixels from the entire feature maps, which are crucial for contrastive learning, as the high-quality candidate set for constructing Q, P and N.

To be specific, we first measure the prediction confidence of each pixel based on the clustering results of last stage. The prediction confidence of the pixel q in the stage S_s is calculated by:

$$z_q^s = \frac{\exp(-(\mathbf{e}_q \cdot \mathbf{c}_{y_q}^s/\tau))}{\sum_{i=1}^{2}\exp(-(\mathbf{e}_q \cdot \mathbf{c}_i^s/\tau))}, \tag{5}$$

where y_q denotes the cluster index q belongs to and τ is the scaling factor. Then we pick out two types of pixels as the high-quality query set for the stage S_{s+1}: 1) the pixels with high confidence which are regarded as representatives of two clusters and are typically close to the cluster centers; 2) the uncertain pixels with low confidence which are mostly near the boundaries between the corrupted and uncorrupted regions. Selecting pixels in such a way for constructing the positive and negative training pairs, our model is able to perform high-resolution contrastive learning in the stages S_2 and S_3 quite efficiently and effectively.

Inter-Stage Semantic Consistency. To guarantee the semantic consistency between different stages of contrastive learning, we reuse the learned features of the previous stage of contrastive learning in current stage. As shown in Figure 3, we concatenate the features in the previous stage with the features in the current stage, which are then fed into the projection head to produce input features for current stage. Accordingly, Equation 3 for high-resolution stages (S_2 and S_3) evolves into the following form:

$$\mathbf{e}_q = \mathcal{F}^s_{\text{proj}}(\text{Concat}(\mathcal{F}^s_{\text{enc}}(\mathbf{I}_q), \mathcal{F}^s_{\text{map}}(\mathcal{F}^{s-1}_{\text{enc}}(\mathbf{I}_q)))), \tag{6}$$

where the mapping function $\mathcal{F}^s_{\text{map}}$ is a linear transformation to regulate the feature dimension of previous stage and balance the effect between current and previous features. Besides the semantic consistency between adjacent stages, another merit of such feature reusing is that our model is able to inherit the learned semantics from previous stage.

Refining the Mask Prediction of Uncertain Pixels. During the high-resolution contrastive learning in the stages S_2 and S_3, we only refine the mask prediction by re-predicting the labels of uncertain pixels with low confidence in the previous stage, i.e., the type-2 samples of selected high-quality query set. The mask labels of other pixels with high confidence are directly inherited from the predictions in the previous stage according to the built quadtree, assuming that the predictions of these pixels are reliable. Thus, such refining process can be performed quite efficiently.

3.3. Transformer-based Restoration Module

As shown in Figure 2, the detected corruption masks are fed into Decoder of the Transformer-based Restoration module for corruption restoration, which consists of the designed mask-guided transformer block.

Mask-guided transformer block. As illustrated in Figure 2, our transformer block contains a multi-head contextual attention (MCA) module, followed by a two-layer MLP with *GELU* nonlinearity in between. Besides, a layer normalization (LN) [1] layer is applied before each attention and each MLP, and a residual learning connection is employed after each module:

$$\begin{aligned} F_i'^s &= \text{MCA}(\text{LN}(F^s_{i-1}), \mathbf{M}^s) + F^s_{i-1}, \\ F^s_i &= \text{MLP}(\text{LN}(F_i'^s)) + F_i'^s, \end{aligned} \tag{7}$$

where F_i^s is the output features of the i-th block in the stage S_s. To fully exploit the predicted masks by hierarchical contrastive learning, we employ the multi-head contextual attention module proposed in MAT [27], which follows the shifted window manner [34] and is formulated as:

$$\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^{\text{T}} + \hat{\mathbf{M}}^s}{\sqrt{d_k}})\mathbf{V}, \tag{8}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are the query, key, and value matrices, respectively. $\sqrt{d_k}$ is the scaling factor. $\hat{\mathbf{M}}$ is defined by

$$\hat{\mathbf{M}}^s_i = \gamma(\mathbf{M}^s_i - 1), \tag{9}$$

where γ is a large positive integer to reduce the values of corrupted pixels.

3.4. Joint Optimization for Parameter Learning

The proposed Hierarchical Contrastive Learning module and the Restoration module are integrated in a holistic way based on the encoding-decoding framework, forming a general-purpose blind image restoration model, as illustrated in Figure 2. The whole model is optimized jointly. Besides the loss of hierarchical contrastive learning (Equation 2), we also adopt three more supervision signals for corruption restoration, including pixel-wise reconstruction loss, perceptual loss, and adversarial loss.

Pixel Reconstruction Loss, which pushes the restored image $\hat{\mathbf{O}}$ as close as its groundtruth $\mathbf{O}_{\text{gt}}$:

$$\mathcal{L}_{\text{pixel}} = \left\| \hat{\mathbf{O}} - \mathbf{O}_{\text{gt}} \right\|_1. \tag{10}$$

Perceptual Loss [21], which performs semantic supervision on the restored images in the deep feature space:

$$\mathcal{L}_{\text{perc}} = \sum_{l=1}^{L} \frac{1}{C_l H_l W_l} \left\| f^l_{\text{vgg}}(\hat{\mathbf{O}}) - f^l_{\text{vgg}}(\mathbf{O}_{\text{gt}}) \right\|_1. \tag{11}$$

Herein $f^l_{\text{vgg}}(\hat{\mathbf{O}})$ and $f^l_{\text{vgg}}(\mathbf{O}_{\text{gt}})$ are the extracted features from $\hat{\mathbf{O}}$ and $\mathbf{O}_{\text{gt}}$ respectively from the l-th convolution layer of the pre-trained VGG-19 network [36].

Adversarial Loss, which employs WGAN-GP [15] to encourage $\hat{\mathbf{O}}$ to be as realistic as its groundtruth $\mathbf{O}_{\text{gt}}$:

$$\mathcal{L}_{\text{adv}} = -\mathbb{E}_{\hat{\mathbf{O}} \backsim \mathbb{P}_{\hat{\mathbf{O}}}} \left[D(\hat{\mathbf{O}}) \right], \tag{12}$$

where D is a discriminator. Overall, the whole model is optimized by:

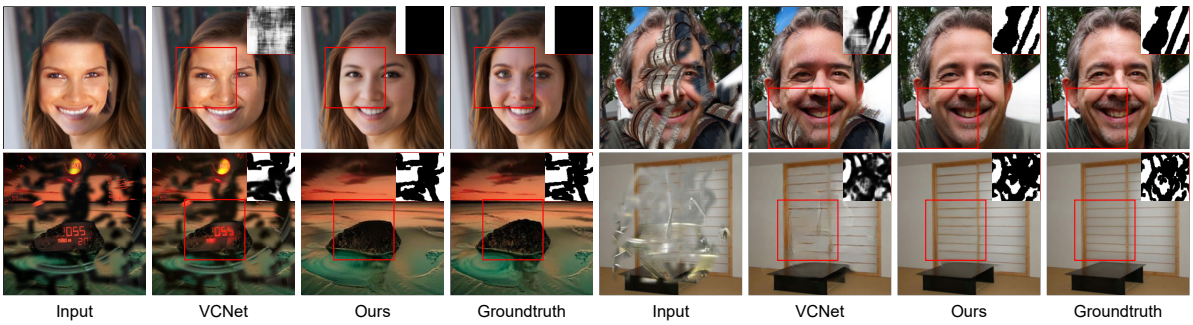


Figure 4: Visualization of detected masks and reconstructed images on four randomly selected samples from test set.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{pixel}} + \lambda_2 \mathcal{L}_{\text{perc}} + \lambda_3 \sum_{s=1}^{3} \mathcal{L}_{\text{CL}}^{s} + \lambda_4 \mathcal{L}_{\text{adv}}, \quad (13)$$

where λ_1, λ_2, λ_3, and λ_4 are hyper-parameters to balance between different losses.

We integrate the Contrastive Learning module between Encoder and Decoder with consistent coarse-to-fine refining stages for all three modules, which potentially enables synchronized optimization of three modules in each scale.

4. Experiments

We first evaluate the performance of our method on image corruption detection, then validates the effectiveness of our method on two challenging image restoration tasks including blind image inpainting and watermark removal. Finally, we conduct extensive ablation study to obtain more insights into our method.

4.1. Experimental Settings

Following the previous work [42] for blind image inpainting, we make two pre-processing operations for data generation: 1) we randomly select natural images from large-scale datasets rather than simple constant values or Gaussian noise as the noisy content for corruptions to increase the difficulty of mask detection in blind image inpainting; 2) we smooth the mask boundaries using alpha blending to avoid distinct boundaries between the corrupted and uncorrupted regions.

Four large-scale datasets are used in our experiments, including FFHQ (faces) [23], CelebA-HQ (faces) [22], ImageNet (objects) [9], and Places (scenes) [48]. For each dataset, images from different datasets are randomly selected as the noise content for corruptions. For instance, the images from CelebA-HQ and ImageNet are used as noise content for FFHQ. We employ the method [29] to generate irregular masks with mask shapes and corruption ratios. We perform experiments on two resolutions of images for comprehensive evaluation: 256×256 and 512×512.

Adam [24] is used as the optimizer by setting β_1, β_2, initial learning rate and batch size to be 0.9, 0.999, 0.0001, and 4, respectively. More experimental details are provided in the supplementary material.

Table 1: Performance comparison for mask detection on three benchmark datasets.

Dataset		FFHQ [23]		ImageNet [9]		Places [48]	
Mask ratio (%)		0-30	30-60	0-30	30-60	0-30	30-60
Acc ↑	VCNet	0.948	0.943	0.972	0.978	0.974	0.976
	Ours	**0.978**	**0.976**	**0.982**	**0.983**	**0.984**	**0.985**
F1 ↑	VCNet	0.967	0.948	0.982	0.980	0.984	0.978
	Ours	**0.986**	**0.978**	**0.989**	**0.985**	**0.991**	**0.987**
BCE ↓	VCNet	0.126	0.137	**0.073**	**0.055**	**0.064**	**0.060**
	Ours	**0.090**	**0.095**	0.083	0.075	0.068	0.064
IoU ↑	VCNet	0.931	0.900	0.966	0.962	0.969	0.959
	Ours	**0.975**	**0.959**	**0.978**	**0.965**	**0.982**	**0.974**

4.2. Image Corruption Detection

Accuracy of corruption detection. We first compare the accuracy of corruption detection between our model and VCNet, the state-of-the-art method for corruption detection in blind image inpainting. Four metrics are used for comprehensive evaluation: binary cross entropy (BCE), classification accuracy (Acc), F1 score, and intersection over union (IoU). Table 1 lists the results for different corruption ratios on three datasets in 256×256 resolution, which show that our model achieves superior performance of mask detection over VCNet in terms of all metrics except BCE. This is reasonable because VCNet is directly supervised by the BCE loss. Besides, we also make qualitative comparison in Figure 4, in which our model detects more precise corruption masks and restores higher-quality images than VCNet.

Image inpainting based on detected corruption masks by different methods. As an indirect evaluation of corruption detection, we perform image inpainting over the detected corruption masks from different methods, employing the same inpainting model MAT [27], and then compare the inpainting performance. To have a comprehensive evaluation, we also provide the inpainting performance over the detected masks from a state-of-the-art segmentation model Segmenter [37] and the groundtruth mask. The results on 512×512 images in Table 2 show that MAT achieves the highest inpainting performance with the mask detected by our method, which is consistent with the performance comparison of corruption detection.

Generalizability on unseen corruption patterns. To evaluate the generalizability of our model methods across different corruption patterns, we perform training and test over

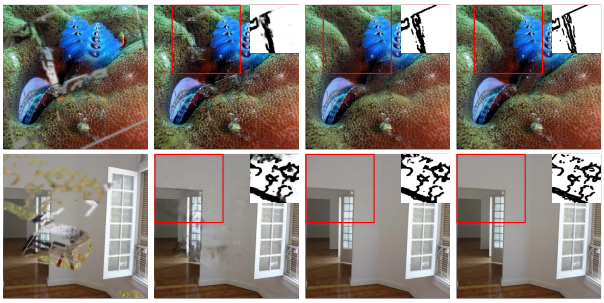


Figure 5: Visual results of blind image inpainting on 512×512 resolution images from the Places [48] dataset.

Table 2: Blind image inpainting based on detected corruption masks by different methods on 512 × 512 images from Places. MAT [27] is finetuned (MAT-F) for fair comparison.

Phase		Corruption Detection		Image Inpainting	
Corruption Detection	Image Inpainting	ACC ↑	IoU ↑	PSNR ↑	SSIM ↑
Groundtruth	MAT-F	–	–	25.34	0.852
Segmenter [37]	MAT-F	0.974	0.962	23.45	0.830
VCNet [42]	MAT-F	0.975	0.963	23.25	0.829
Ours	MAT-F	0.980	0.970	24.53	0.844
VCNet	VCNet	0.975	0.963	21.49	0.764
Ours	Ours	0.980	0.970	25.28	0.846

different corruption patterns. Table 4 presents the results of our model as well as VCNet by training them on Places dataset while performing testing directly on unseen corruption patterns such as random/constant noise and image content from CelebA-HQ [22]. It shows that our model consistently outperforms VCNet in all cases. which reveals the better generalizability of our model over VCNet.

4.3. Downstream Image Restoration Tasks

4.3.1 Blind Image Inpainting

In the experiments of blind image inpainting, we compare our model with VCNet and MPRNet [45], a state-of-the-art approach for image restoration. Table 3 presents the inpainting performance on images of 256×256 resolution. Our method achieves the best performance in terms of all metrics and outperforms the other two methods by a large margin. Besides, the qualitative comparison in Figure 4 also demonstrates that our model is able to restore higher-quality images than other two methods, benefiting from the precise corruption detection by the proposed hierarchical contrastive learning mechanism as well as the integrated restoration framework. Note that we also provide the user study in the supplementary material.

High-resolution blind image inpainting. We further conduct experiments on images of 512×512 resolution in Places [48] dataset for blind image inpainting. As shown in Table 2, our model outperforms VCNet substantially, consistent with the comparison in the case of 256×256 resolution. Besides, the qualitative comparison in Figure 5 also validates the superiority of our model. More visual results are presented in the supplementary material.

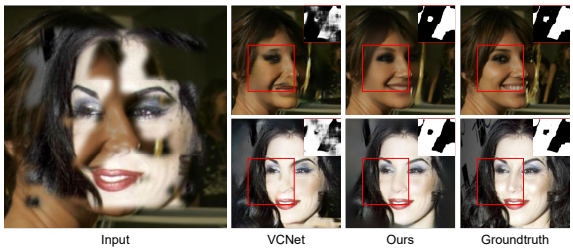


Figure 6: Visual results of bidirectional blind image inpainting on FFHQ [23]. More results are shown in the supplementary material.

Figure 7: Visual comparison of both corruption detection and blind image inpainting over unseen corruptions. More examples could be found in the supplementary material.

Table 3: Performance comparison for blind image inpainting in terms of four evaluation metrics on three benchmark datasets.

Dataset		FFHQ [23]			ImageNet [9]			Places [48]		
Mask ratio (%)		0-20	20-40	40-60	0-20	20-40	40-60	0-20	20-40	40-60
PSNR ↑	MPRNet	33.93	26.75	22.06	33.76	26.26	21.50	33.37	25.59	21.08
	VCNet	29.86	24.51	20.51	29.13	23.39	19.42	29.66	23.19	19.18
	Ours	**34.79**	**27.98**	**23.54**	**33.91**	**26.40**	**21.80**	**34.11**	**26.20**	**21.66**
SSIM ↑	MPRNet	0.968	0.898	0.799	0.965	0.882	0.762	0.963	0.872	0.743
	VCNet	0.933	0.849	0.736	0.924	0.818	0.678	0.942	0.833	0.688
	Ours	**0.971**	**0.910**	**0.824**	**0.966**	**0.884**	**0.767**	**0.969**	**0.887**	**0.746**
FID ↓	MPRNet	3.37	12.20	29.74	3.71	17.80	53.38	4.15	17.61	44.73
	VCNet	6.80	14.17	28.90	5.69	21.48	59.11	5.93	18.62	40.21
	Ours	**2.37**	**7.28**	**15.48**	**2.69**	**12.35**	**41.37**	**2.88**	**11.10**	**31.33**
LPIPS ↓	MPRNet	0.027	0.096	0.212	0.033	0.123	0.269	0.033	0.131	0.283
	VCNet	0.044	0.103	0.188	0.045	0.128	0.246	0.046	0.138	0.261
	Ours	**0.019**	**0.065**	**0.139**	**0.025**	**0.098**	**0.226**	**0.024**	**0.095**	**0.221**

Bidirectional blind image inpainting. When real natural images are used as the noise content for corruption, we can perform bidirectional image restoration: restore either a complete noise image or a complete background image by reversing the corruption mask. It is quite challenging to perform well in inpainting of both directions. The visual results in Figure 6 show that our model performs more robust than VCNet in both corruption detection and image inpainting.

Generalization on unseen corruption patterns. Using the same experimental settings as evaluating the generalizability for corruption detection, we also validate the generalizability of our model across unseen corruption patterns on image inpainting. Table 5 shows the image inpainting performance of both our model and VCNet on unseen corruption patterns during training. The experimental results indi-

Table 4: Generalizability over unseen corruptions for corruption detection. Models are trained on Places.

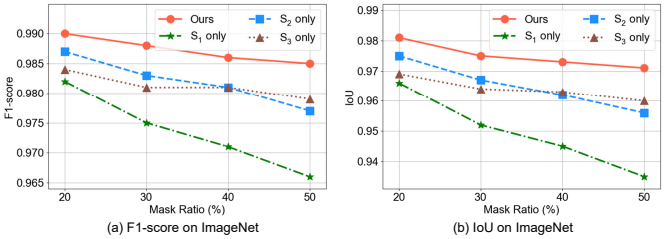
Corruption pattern		Random constant		CelebA-HQ [22]	
Mask ratio (%)		0-30	30-60	0-30	30-60
Acc ↑	VCNet	0.981	0.977	0.975	0.974
	Ours	**0.990**	**0.989**	**0.986**	**0.986**
F1 ↑	VCNet	0.988	0.978	0.984	0.976
	Ours	**0.994**	**0.990**	**0.992**	**0.988**
BCE ↓	VCNet	**0.050**	0.078	0.063	0.068
	Ours	**0.050**	**0.049**	**0.062**	**0.054**
IoU ↑	VCNet	0.977	0.960	0.969	0.954
	Ours	**0.988**	**0.981**	**0.983**	**0.976**

Table 5: Generalizability over unseen corruptions for blind inpainting. Models are trained on Places.

Corruption Pattern		Random constant			CelebA-HQ [22]		
Mask ratio (%)		0-20	20-40	40-60	0-20	20-40	40-60
PSNR ↑	VCNet	30.48	24.33	19.30	29.77	23.21	18.96
	Ours	**37.69**	**29.01**	**23.05**	**34.60**	**26.52**	**21.82**
SSIM ↑	VCNet	0.957	0.876	0.732	0.947	0.840	0.685
	Ours	**0.985**	**0.933**	**0.818**	**0.972**	**0.894**	**0.772**
FID ↓	VCNet	6.41	16.84	42.54	5.47	17.75	41.19
	Ours	**1.80**	**7.45**	**24.48**	**2.65**	**10.50**	**30.42**
LPIPS ↓	VCNet	0.051	0.131	0.268	0.041	0.128	0.256
	Ours	**0.013**	**0.063**	**0.182**	**0.021**	**0.089**	**0.214**

Table 6: Ablation study on hierarchical interaction mechanism.

Mask ratio (%)		0-30	30-60
Acc ↑	w/o inter-stage consistency	0.969	0.965
	w/o sample selection	0.975	0.974
	Complete model	**0.978**	**0.976**
F1 ↑	w/o inter-stage consistency	0.980	0.969
	w/o sample selection	0.984	0.976
	Complete model	**0.986**	**0.978**
BCE ↓	w/o inter-stage consistency	0.102	0.115
	w/o sample selection	0.094	0.101
	Complete model	**0.090**	**0.095**
IoU ↑	w/o inter-stage consistency	0.962	0.940
	w/o sample selection	0.972	0.957
	Complete model	**0.975**	**0.959**



(a) F1-score on ImageNet (b) IoU on ImageNet

Figure 8: Hierarchical vs single-stage contrastive learning.

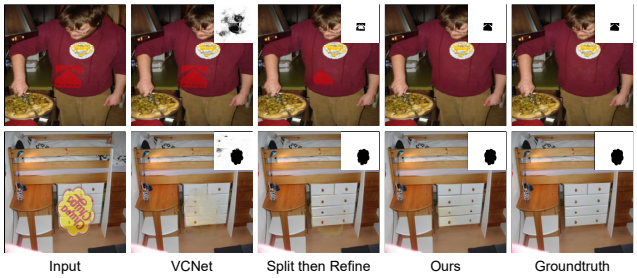


Figure 9: Visual comparison with the state-of-the-art model *Split then Refine* [7] and VCNet for watermark removal.

Table 7: Performance evaluation on watermark removal.

Metrics	PSNR↑	SSIM↑	FID↓
SIRF [46]	34.63	0.978	0.071
BS^{2}AM [6]	34.88	0.979	0.028
DHAN [8]	37.67	0.986	0.062
BVMR [19]	38.28	0.985	0.018
Split then Refine [7]	41.27	0.991	0.011
VCNet [42]	32.66	0.963	0.032
Ours	**41.88**	**0.992**	**0.007**

cate that our model performs much better than VCNet. Furthermore, the qualitative comparison in Figure 7 also validates such better generalizability of our model over VCNet on both corruption detection and image inpainting.

4.3.2 Image Watermark Removal

In this set of experiments, we further evaluate the performance of our model on LOGO-30K dataset [7] for watermark removal. The comparison in Table 7 shows that our model performs best in terms of all metrics and compares favorably with other specialized methods for watermark removal, which demonstrates the robustness of our model across different image restoration tasks. Moreover, the visual results in Figure 9 also validate the superiority of our model over other methods in watermark removal.

4.4. Ablation Study

Effect of hierarchical contrastive learning. To investigate the effectiveness of the proposed hierarchical contrastive learning framework, we compare its performance with that of single-stage contrastive learning. The results in Figure 8 demonstrate the distinct advantage of the proposed hierarchical contrastive learning framework.

Effect of hierarchical interaction mechanism. We further conduct experiments to investigate the effect of hierarchical interaction mechanism, especially the proposed 'inter-stage semantic consistency' and 'high-quality sample selection strategy'. Table 6 presents the concrete experimental results. While both techniques can boost the performance, 'inter-stage semantic consistency' is more crucial to the performance due to its essential merit: semantic propagation between stages by feature reusing.

5. Conclusion

In this work, we have designed a novel method, namely hierarchical contrastive learning framework, which can automatically detect corruption masks in pixel level and thus allows for blind image corruption restoration without known corruption masks. Extensive quantitative and qualitative comparisons have demonstrated the superior performance of our method over other methods for various corruption restoration tasks and its well generalization ability across different corruption patterns.

6. Acknowledgement

This work was supported in part by the NSFC fund (Grant No. U2013210, 62006060, 62176077), in part by the Shenzhen Key Technical Project under Grant 2022N001, 2020N046, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant (Grant No. 2022A1515010306), in part by Shenzhen Fundamental Research Program (Grant No. JCYJ20210324132210025, JCYJ20220818102415032), and in part by the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (Grant No. 2022B1212010005).

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[2] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009. 2

[3] Nian Cai, Zhenghang Su, Zhineng Lin, Han Wang, Zhijing Yang, and Bingo Wing-Kuen Ling. Blind inpainting using the fully convolutional neural network. *The Visual Computer*, 33(2):249–261, 2017. 2

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[5] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing*, 13(9):1200–1212, 2004. 2

[6] Xiaodong Cun and Chi-Man Pun. Improving the harmony of the composite image by spatial-separated attention module. *IEEE Transactions on Image Processing*, 29:4759–4771, 2020. 8

[7] Xiaodong Cun and Chi-Man Pun. Split then refine: stacked attention-guided resunets for blind single image visible watermark removal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1184–1192, 2021. 8

[8] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10680–10687, 2020. 8

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6, 7

[10] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 2

[11] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Incremental transformer structure enhanced image inpainting with masking positional encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11358–11368, 2022. 1, 2

[12] Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, pages 341–346, 2001. 2

[13] Xin Feng, Wenjie Pei, Fengjun Li, Fanglin Chen, David Zhang, and Guangming Lu. Generative memory-guided semantic reasoning model for image inpainting. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 1

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 2

[15] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017. 5

[16] Xiefan Guo, Hongyu Yang, and Di Huang. Image inpainting via conditional texture and structure dual generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14134–14143, October 2021. 2

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[18] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4

[19] Amir Hertz, Sharon Fogel, Rana Hanocka, Raja Giryes, and Daniel Cohen-Or. Blind visual motif removal from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6858–6867, 2019. 8

[20] Yong Shi Jie Yang, Zhiquan Qi. Learning to incorporate structure knowledge for image inpainting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12605–12612, 2020. 2

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 5

[22] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 6, 7, 8

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 6, 7

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[25] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[26] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020. 2

[27] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10758–10768, 2022. 1, 2, 5, 6, 7

[28] Xiaoguang Li, Qing Guo, Di Lin, Ping Li, Wei Feng, and Song Wnag. Misf: Multi-level interactive siamese filtering for high-fidelity image inpainting. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[29] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision*, pages 85–100, 2018. 2, 6

[30] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 725–741. Springer, 2020. 2

[31] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4170–4179, 2019. 2

[32] Qiankun Liu, Zhentao Tan, Dongdong Chen, Qi Chu, Xiyang Dai, Yinpeng Chen, Mengchen Liu, Lu Yuan, and Nenghai Yu. Reduce information loss in transformers for pluralistic image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11347–11357, 2022. 1, 2

[33] Yang Liu, Jinshan Pan, and Zhixun Su. Deep blind image inpainting. In *International Conference on Intelligent Science and Big Data Engineering*, pages 128–141. Springer, 2019. 2

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 5

[35] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *The IEEE International Conference on Computer Vision Workshops*, Oct 2019. 2

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[37] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021. 6, 7

[38] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2020. 4

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[40] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10581–10590, 2021. 3

[41] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 2

[42] Yi Wang, Ying-Cong Chen, Xin Tao, and Jiaya Jia. Vcnet: A robust approach to blind image inpainting. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020. 1, 2, 6, 7, 8

[43] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10551–10560, 2021. 3

[44] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480, 2019. 2

[45] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14821–14831, 2021. 7

[46] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4786–4794, 2018. 8

[47] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019. 2

[48] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 6, 7