Image Deraining using Degradation-aware CLIP

Pullabhotla Bhuvana Chandra
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, India
200070063@iitb.ac.in

***Abstract*—Image deraining is helpful for humans' visual perception on multiple levels and is necessary for many tasks in computer vision(e.g., object detection). Vision-language models such as CLIP have greatly impacted diverse downstream tasks for zero-shot or label-free predictions. However, when it comes to low-level vision, such as image restoration, their performance deteriorates dramatically due to corrupted inputs. In this paper, we present a degradation-aware vision language model (DA-CLIP) to better transfer pre trained vision-language models to low-level vision tasks as a universal framework for image restoration, specifically experimenting on image deraining tasks. Mean reverting SDE(Stocasthic Differential Equations) is used as the base model for image restoration task. The key construction consists in a mean-reverting SDE that transforms a high quality image into a degraded counterpart as a mean state with fixed Gaussian noise. Then, by simulating the corresponding reverse-time SDE, we can restore the origin of the low-quality image without relying on any task-specific prior knowledge.**

***Index Terms*—Diffusion models, stochastic differential equations, vision language models(VLMs), image deraining, contrastive learning**

I. Introduction

The goal of single image deraining algorithms is to generate sharp images from a rainy image input. Image deraining can potentially benefit both the human visual perception quality of images, and many computer vision applications, such as outdoor surveillance systems and intelligent vehicles, basically helping in tasks like object detection, remote sensing, autonomous driving, and semantic segmentation. Large-scale pre-trained vision-language models (VLMs) such as CLIP [1] have recently garnered significant attention partly because of their wide-reaching usefulness on many fundamental computer vision tasks. Oftentimes, image restoration methods learn to generate images pixel-by-pixel using an l_1 or l_2 loss without leveraging knowledge of the degradation type.

In this paper, we combine large-scale pre-trained vision-language models with image restoration networks and present an effective framework for universal image restoration, experimenting with image-deraining task. Specifically, aiming at addressing feature mismatching between corrupted inputs and clean captions, we propose an Image The controller adapts the VLM's image encoder to output high-quality (HQ) content embeddings aligned with clean captions. Meanwhile, the controller itself also predicts a degradation embedding to match the real degradation types. For image restoration, diffusion models have shown impressive performance in various image generation tasks, based on modeling a diffusion process and then learning its reverse. Among the commonly used formulations, we adopt the use of diffusion models defined via stochastic differential equations (SDEs), specifically mean-reverting SDE. Figure.1

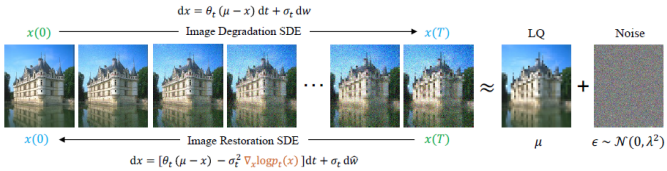

Fig. 1. Overview of Image Restoration from Mean-reverting SDE

shows the overview of the process. It adapts the forward process to model the image degradation itself, from a high-quality image to its low-quality counterpart. By simulating the correspond ing reverse-time SDE, high-quality images can be restored. Importantly, no task-specific prior knowledge is required to model the image degradation at test time, just a set of image

II. Background

A. Vision-Language Models

Recent works have demonstrated the great potential of applying pretrained VLMs to improve downstream tasks with generic visual and text representations. A

classic VLM usually consists of a text encoder and an image encoder and tries to learn aligned multimodal features from noisy image-text pairs with contrastive learning. Although VLMs provide a strong capability of zero shot and label-free classification for downstream tasks, they have so far had limited effect on image restoration. A noteworthy approach for finetuning vision-language models is so-called prompt learning [2] where the prompt's context words are represented by learnable vectors that are then optimized for the downstream task.

B. Reverse-time SDEs

In this section, we briefly review the key concepts underlying SDE-based diffusion models and show the process of generating samples with reverse-time SDEs. Let p_0 be the initial distribution representing the data and $t\epsilon[0,T]$ be the continuous time variable, we consider a diffusion process $x(t)_{t=0}^{T}$ defined by SDE of the form

$$dx = f(x,t)dt + g(t)dw, \tag{1}$$

where f and g are the drift and dispersion functions, respectively,$x(0) = p_0(x)$,w is a standard Wiener process, and $x(0)\epsilon R^d$ is the initial condition. Typically $x(T)$ follows a Gaussian distribution with fixed mean and variance. The general idea is to design such an SDE that gradually trans forms the data distribution into fixed Gaussian noise.

Now We can then reverse the process to sample data from noise by simulating the SDE backward in time [3]. Anderson (1982) shows that a reverse-time representation of the SDE 1 is given by

$$dx = \left[f(x,t) - g(t)^2\nabla_x \log p_t(x)\right] dt + g(t)d\hat{w} \tag{2}$$

where $x(T) \sim p_T(x)$. Here, $\hat{w}$ is a reverse-time Wiener process, and $p_t(x)$ stands for the marginal probability density function of $x(t)$ at time t. The score function $\nabla_x \log p_t(x)$ is in general intractable, and thus SDE-based diffusion models approximate it by training a time-dependent neural network $s_\theta(x,t)$ under a so-called score matching objective.

III. Degradation-aware CLIP

Figure.2 gives a brief overview of the modified CLIP model(DA CLIP) where an image controller is added to get the degradation embeddings from the image. At the core of our approach is the idea of controlling a pre-trained CLIP model to output the highquality image feature from a corrupted image while simultaneously predicting the degradation type. The image content embedding e_c^I matches the clean caption embeddinge_c^T.

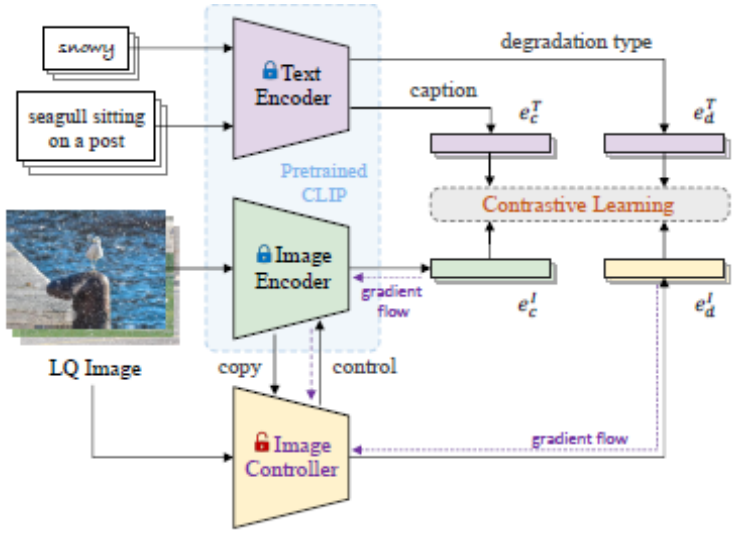

Fig. 2. Overview of DA-CLIP [4]

Moreover, the image degradation embedding e_d^I predicted by the controller specifies the corruption type of the input, i.e., the corresponding degradation embedding e_d^T from the text encoder. These features can then be integrated into other image restoration models to improve their performance.

A. Image controller

The image controller is a copy of the CLIP image encoder but wrapped with a few zero-initialized connections to add controls to the encoder. It manipulates the outputs of all encoder blocks to control the prediction of the image encoder. This paper uses ViT [5] as the default backbone for both the encoder and the controller.Figure.3 illustrates the controlling procedure, where the output of the controller consists of two parts: an image degradation embedding e_d^I and hidden controls h_c.

We freeze all weights of the pretrained CLIP model and only fine-tune the image controller. We use a contrastive objective to make the degradation-embedding spaces discriminative and well separated to learn the embedding matching process. Let N denote the number of paired embeddings (from text encoder and image encoder/controller) in a training batch. The contrastive loss is defined as:

$$\mathcal{L}_{\text{con}}(x,y) = -\frac{1}{N}\sum_{i=1}^{N}\frac{\exp(x_i^T \cdot y_i/t)}{\sum_{j=1}^{N}\exp(x_i^T \cdot y_j/t)} \tag{3}$$

where x and y are normalised vectors, and t is a learnable temperature parameter that controls the contrastive

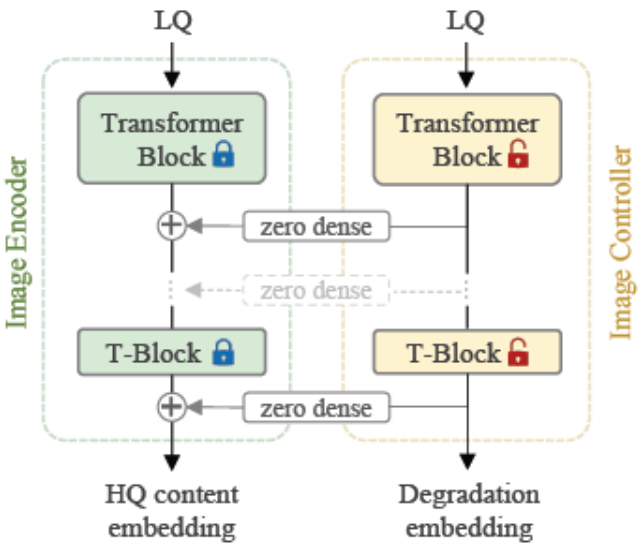

Fig. 3. Controller for ViT-based image encoder

strength. To optimize both content and degradation embeddings, we use the following joint objective:

$$\mathcal{L}_c(\omega) = \mathcal{L}_{\text{con}}(e_c^I, e_c^T; \omega) + \mathcal{L}_{\text{con}}(e_d^I, e_d^T; \omega) \tag{4}$$

IV. Image Restoration from Mean-Reverting SDE

The key idea of our proposed image restoration approach is to combine a mean-reverting SDE with a maximum likelihood objective for neural network training.

A. Forward SDE for Image degradation

We construct a special case of the SDE (1) whose score function is analytically tractable, as follows:

$$dx = \theta_t(\mu - x)dt + \sigma_t dw, \tag{5}$$

where μ is the state mean, and θ_t, σ_t are time-dependent positive parameters that characterize the speed of the mean-reversion and the stochastic volatility, respectively.

In general, the starting state can be set to any pair $(\mu, x(0))$ of different images. The forward SDE (5) then transfers one image to the other as a kind of noisy interpolation. To carry out image degradation, we let μ and $x(0)$ be the ground truth high-quality (HQ) image and its degraded low-quality (LQ) counterpart, respectively (see Fig.1). It is worth noting that while μ thus depends on $x(0)$ (as they are paired HQ and LQ images of the same object or scene), $x(0)$ is independent of the Brownian motion and the SDE is therefore still valid in the Itˆo sense.

For our SDE (5) to have a closed-form solution, we set $\frac{\sigma_t^2}{\theta_t} = 2\lambda^2$, where λ^2 is the stationary variance.

$$\begin{aligned} x(t) = \mu + (x(s) - \mu)\, e^{-\bar{\theta}_{s:t}} \\ + \int_s^t \sigma e^{-2\bar{\theta}_{z:t}}\, dW(z), \end{aligned} \tag{6}$$

where

$$\bar{\theta}_{s:t} = \int_s^t \theta(z)\, dz$$

is known, and the transition kernel

$$p(x(t)|x(s)) = \mathcal{N}\left(x(t) \,|\, m_{s:t}(x(s)), v_{s:t}\right) \tag{7}$$

is a Gaussian with mean $m_{s:t}$ and variance $v_{s:t}$, given by:

$$\begin{aligned} m_{s:t}(x(s)) &= \mu + (x(s) - \mu)\, e^{-\bar{\theta}_{s:t}}, \\ v_{s:t} &= \lambda^2 \left(1 - e^{-2\bar{\theta}_{s:t}}\right). \end{aligned} \tag{8}$$

Note that as $t \to \infty$, the mean m_t converges to the low-quality image μ and the variance v_t converges to the stationary variance λ^2 (hence the qualifier "mean-reverting"). In other words, the forward SDE (5) diffuses the high-quality image into a low-quality image with fixed Gaussian noise.

B. Reverse-Time SDE for Image Restoration

To recover high quality image(HQ) from terminal state x(T), we reverse the SDE (5) according to (2) to derive an image restoration IR-SDE given by

$$dx = \left[\theta_t(\mu - x) - \sigma_t^2 \nabla_x \log p_t(x)\right] dt + \sigma_t d\hat{w} \tag{9}$$

At test time, the only unknown part is the score $\log p_t(x)$ of the marginal distribution at time t. But during training, the ground truth, high-quality image is available and thus we can train a neural network to estimate the conditional score $\log p_t(x|x^{(0)})$. Specifically, we can use

$$\nabla_x \log p_t(x|x(0)) = -\frac{x(t) - m_t(t)}{v_t} \tag{10}$$

V. Experiment

A. Dataset Construction

We evaluate the image deraining(image restoration) task on the synthetic raining dataset- Rain100H [6], where there are 1800 HQ-LQ image pairs for training and 100 HQ-LQ image pairs for testing

B. Image Restoration using DA-CLIP

We use IR-SDE as the base framework for image restoration. It adapts a U-Net architecture similar to DDPM [7] but removes all self-attention layers. To inject clean content embeddings into the diffusion process, we introduce a cross-attention mechanism to learn semantic guidance from pre-trained VLMs. Considering the varying input sizes in image restoration tasks and the increasing cost of applying attention to high-resolution features, we only use cross-attention in the bottom blocks of the U-Net for sample efficiency.

On the other hand, the predicted degradation embeddings are useful for unified image restoration, where the aim is to process low-quality images of multiple degradation types with a single model. Moreover, to make use of the degradation embeddings, we combine them with a prompt learning [2] module to improve the results further, as shown in Figure.(4). Generally, we can use cross-attention to integrate content embedding into networks to improve their performance on an image restoration task. In contrast, the prompt module combined with degradation embedding specifically aims to improve the classification of the degradation type in the context of unified image restoration. The prompt module helps the model dynamically adjust its behavior based on the provided textual context, enhancing its capability to adapt to different conditions or environments. We use BLIP [8] for generating clean captions from HQ images; we also give a degradation type attached to these captions by specifying them in that degradation directory, and next, we train them with the Image controller where the image and text encoder are trained already. Here, DA-CLIP model pre-trained weights are already taken, and the image restoration is performed

C. Results

After performing the IR-SDE task with and without DA-CLIP on the Rain100H training dataset and testing it on the test dataset, the obtained results are as follows, it is observed that we get a significant improvement in the PSNR and SSIM values when DA-CLIP is added to the image restoration task, the training time is improved for same number of epochs used for training the dataset

Method	PSNR	SSIM	LPIPS
IR-SDE	26.212	0.8520	0.0845
IR-SDE with DA-CLIP	29.980	0.8867	0.0505

TABLE I
OBTAINED RESULTS IN RGB SPACE

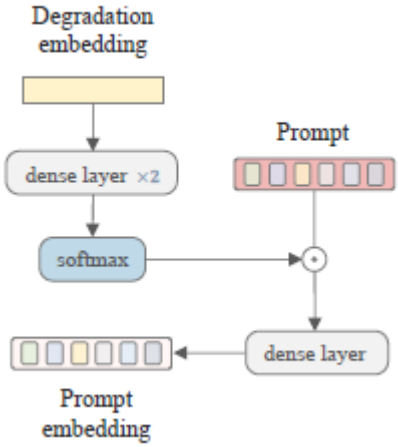

Fig. 4. Prompt with degradation embeddings

Method	PSNR	SSIM
IR-SDE	27.7002	0.8707
IR-SDE with DA-CLIP	31.3411	0.8997

TABLE II
OBTAINED RESULTS IN YCBCR SPACE

Fig. 5. Rain image, IR-SDE only, DA-CLIP added, in similar order

Fig. 6. Rain image, IR-SDE only, DA-CLIP added, in similar order

Fig. 7. Rain image, IR-SDE only, DA-CLIP added, in similar order

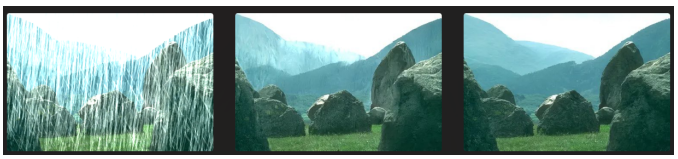

Fig. 8. Rain image, IR-SDE only, DA-CLIP added, in similar order

First two images are real rain images taken from the internet and not from the test data. The test data consists of synthetic rain images

VI. Conclusions

At the core of our approach is a controller that accurately predicts the degradation embeddings for low-quality images and also controls the CLIP image encoder to output high-quality content embeddings that improve the image restoration from a low-quality image of multiple degradation types. Using VLMs(text embeddings) significantly improved the training time and results for the dataset's same number of training rotations. Including the prompt module modifies the time variable in the SDE image restoration part, significantly affecting the training time and the amount of memory the task uses. using VLMs for image restoration is an efficient task and it need not have a task specific previous knowledge.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[2] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, p. 2337–2348, Jul. 2022. [Online]. Available: http://dx.doi.org/10.1007/s11263-022-01653-1

[3] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," 2021.

[4] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön, "Controlling vision-language models for multi-task image restoration," 2023.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[6] W. Yang, R. T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan, "Deep joint rain detection and removal from a single image," 2017.

[7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020.

[8] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," 2022.