Diffusion-Based Adaptation for Classification of Unknown Degraded Images

Dinesh Daultani[1], Masayuki Tanaka[1], Masatoshi Okutomi[1], Kazuki Endo[2]
[1]Tokyo Institute of Technology, [2]Teikyo Heisei University
ddaultani@ok.sc.e.titech.ac.jp, {mtanaka,mxo}@sc.e.titech.ac.jp, k.endo@thu.ac.jp

Abstract

Classification of unknown degraded images is essential in practical applications since image-degraded models are usually unknown. Diffusion-based models provide enhanced performance for image enhancement and image restoration from degraded images. In this study, we use the diffusion-based model for the adaptation instead of restoration. Restoration from the degraded image aims to restore the degrade-free clean image, while adaptation from the degraded image transforms the degraded image towards a clean image domain. However, the diffusion models struggle to perform image adaptation in case of specific degradations attributable to the unknown degradation models. To address the issue of imperfect adapted clean images from diffusion models for the classification of degraded images, we propose a novel Diffusion-based Adaptation for Unknown Degraded images (DiffAUD) method based on robust classifiers trained on a few known degradations. Our proposed method complements the diffusion models and consistently generalizes well on different types of degradations with varying severities. DiffAUD improves the performance from the baseline diffusion model and clean classifier on the Imagenet-C dataset by 5.5%, 5%, and 5% with ResNet-50, Swin Transformer (Tiny), and ConvNeXt-Tiny backbones, respectively. Moreover, we exhibit that training classifiers using known degradations provides significant performance gains for classifying degraded images.[1]

1. Introduction

Computer vision models have transformed human life in critical ways. Computer vision applications include autonomous driving [1, 13], facial recognition [42], healthcare [10], robotics [19, 30], and agriculture [20]. However, in all differing computer vision applications, there is scope for input images/videos to be degraded or corrupted with diverse types of degradations. A few recurring degradations include Gaussian noise due to inadequate illumination, fog due to weather conditions, blur caused by sudden camera movement, and low image quality due to lossy compression techniques such as JPEG. Many studies acknowledge [8, 9, 21, 34, 37] that computer vision models gradually decline in performance with increased severity levels of degradation. In our study, we use the terms degradation or corruption analogously.

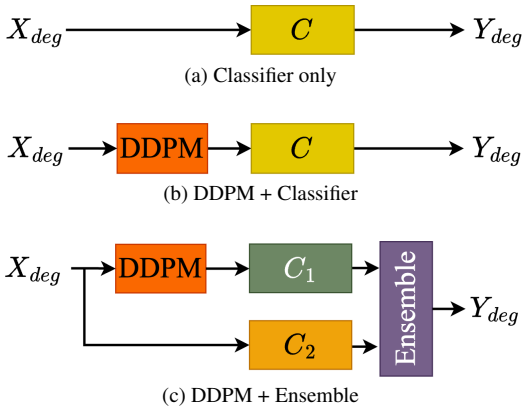


Figure 1. Different categories of methods, where X_{deg} and Y_{deg} represent input degraded image and its associated inferred label. C represents typical classifiers to predict a given image and DDPM is denoising diffusion probalisitic model. C_1, and C_2 represents two classifiers trained with different dataset.

Image restoration methods such as AirNet [25] and SwinIR [27] are notable for their performance. AirNet [25] handles noise, haze, and rain degradations well, while SwinIR [27], based on the Swin Transformer [28], demonstrates decent performance in super-resolution, denoising, and JPEG artifacts reduction. On the other hand, diffusion models have shown remarkable performance on generation tasks, as shown by Dhariwal and Nichol [7, 32]. In addition, the diffusion-driven adaptation (DDA) [11] method applies the denoising diffusion probabilistic model (DDPM) to adapt any unknown degraded images and then classify adapted images. However, DDA and DDPM can not correctly adapt all different types of degraded images back to the source domain. Especially in the case of blurred images,

[1]Our source code will be available here after publication: https://github.com/dineshdaultani/DiffAUD

adapted images often need to be more perfectly deblurred. Hence, the classifier trained on clean images would not naturally perform well on these imperfect images. Our study tries to resolve this limitation by training separately robust classifiers using a few known degradations and knowledge distillation [18]. Since our trained classifiers are more robust towards a few known degradation images, they perform remarkably well on diverse types of unknown corruptions with varying severity levels. Moreover, our proposed classifiers can complement any DDPM and adaptation process.

Concurrently, several studies [5, 9, 37, 41] use a model trained on synthetically degraded images to classify degraded images. Our study merges the domain of diffusion models and image degradation research to classify unknown degraded images. As diffusion models become more efficient in terms of model parameters and low latency inferencing, it has immense opportunities for practical application utilization, such as autonomous driving applications where extreme diversity in climate and environment conditions can lead to a substantial reduction in the performance of vision models [12, 16]. In this paper, we focus on the image classification task. However, our study can bridge the gap by integrating imperfectly adapted images from diffusion models into different downstream computer vision tasks such as object detection and semantic segmentation.

Fig. 1 shows three methods considered in our study for classifying degraded images. First, the "Classifier only" method is a typical classification method. Next, "DDPM + Classifier" combines DDPM and classifier, where we first adapt the given degraded image and then classify it using a typical classifier. Next is the "DDPM + Ensemble" method, where two classifiers are employed to combine the output with an ensemble. In the case of DDA [11], C_1 and C_2 share their model parameters which are pre-trained on clean images. We base our proposal also on "DDPM + Ensemble" architecture, where the difference between our proposal and DDA [11] lies in the classifiers. In our proposal, we use distilled classifiers trained on adapted and degraded images as further explained in Section 3.

The primary contributions of our study are summarized as follows:

1. We propose a diffusion-based classifier training method, *i.e.*, DiffAUD, based on known degradations and distillation that work complementary with diffusion models to classify unknown degraded images.
2. We demonstrate that training a classifier with few synthetically degraded images can enhance generalization and robustness for various types of single and sequential degradations.
3. We provide comprehensive evidence that our proposed method consistently outperforms other existing methods on different datasets, backbones, and single/sequential degradations with varying levels of severity.

2. Related work

2.1. Classification of Degraded Images

Several works have explored how to train and evaluate models on single known degradations [5, 9, 37, 41]. Specifically, Roy *et al*. [37] proposed a CapsuleNet [39] based architecture that performs better than other commonly known convolution architectures on different degradations. Endo *et al*. [9] proposed an architecture on restoration and ensemble while employing degradation and degradation levels information that perform well on different levels of known degradations. Daultani *et al*. [5] proposed a knowledge distillation [18] based network with cutout [6] data augmentation where they show that increasing the depth of the network architecture is not required but instead applying data augmentation can improve the robustness of the network. Recently, Sasaya *et al*. [41] proposed a distillation-based architecture without needing paired clean and degraded images to classify Gaussian noise images. However, in all the above cases, an individual architecture is utilized to predict a particular degradation image. Conversely, we handle a wide variety of unknown degraded images.

Simultaneously, Laugros *et al*. [24] has studied how augmenting the images with known corruptions can improve the robustness for unknown corruptions. Daultani *et al*. [4] has shown that training the classification models on one known degradation is insufficient for different degradations. Hence, they provide a framework to utilize several known degradation images to improve the performance of respective known degradations. Our work is inspired by FusionDistill [4] since we incorporate several degradations for training a classifier that can classify images with different types of degradations. However, among others, some significant differences in our proposed approach are: (1) The FusionDistill method utilizes paired clean and degraded images; however, we utilize only unpaired degradation images in training. (2) FusionDistill focuses only on known single degradations; in contrast, we focus on a wide variety of unknown single and sequential degradations.

2.2. Diffusion Models

Diffusion models have shown tremendous potential in versatile generation tasks [47]. Concretely, among others they are utilized for tasks such as super resolution [26, 40], medical image reconstruction [3, 46], text-to-image generation [35, 36], and text-to-video generation [2]. Especially, denoising diffusion restoration models (DDRM) [22] have considerably improved image restoration by using diffusion models to convert known degradation operator images to clean images. However, using it requires knowledge about the degradation operator, which limits its application for unknown corruptions.

Particularly, Nie *et al*. [33] establish how diffusion mod-

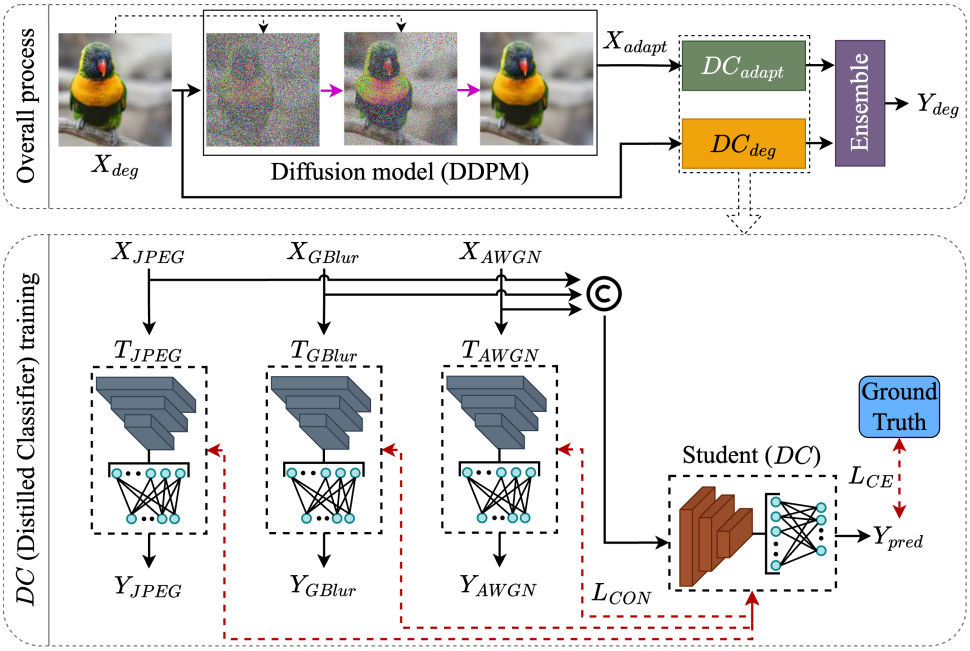


Figure 2. Architecture diagram of the proposed method, where the top figure shows the overall inferencing process and the bottom block represents the training process of DC using knowledge distillation from pre-trained teacher networks. Symbol © represents concatenation of inputs X_{JPEG}, X_{GBlur}, and X_{AWGN}. Grey/blue and orange/blue blocks represent pre-trained teacher and student networks.

els can improve the robustness of pre-trained classifiers against unseen adversarial attacks. Similarly, Gao *et al*. [11] has shown that diffusion models can classify diverse corrupted images with high severity without training by adapting the corrupted images back to the source or clean image domain with a classifier trained on clean images. However, the performance could be higher since their approach DDA [11] assumes that the adapted images from the diffusion models perfectly match the source domain. However, due to an unknown degradation model, adapting the images perfectly to the source domain is challenging. Our proposed method improves the robustness of the classifier, presuming that the adapted images will undergo imperfect adaptation from the diffusion model.

3. Method

3.1. Notations and Preliminaries

We categorize three types of training images, *i.e*., clean, degraded, and adapted, denoted as X_{clean}, X_{deg}, and X_{adapt}, respectively. Clean images are natural images without degradation; degraded images undergo synthesis using a specific degradation model, and the adapted images are sampled by applying DDPM on degraded images. Furthermore, there are two types of classifier in our study, *i.e*., simple classifier and distilled classifier, denoted as C and DC. C and DC are trained using image and label pairs $\{X_k, Y\}$, where $k \in \{clean, deg, adapt\}$ represents clean, degraded, and adapted images respectively. We represent classifiers trained with clean, adapted, and degraded images as C_{clean}, C_{adapt}, and C_{deg}, respectively. Likewise, distilled classifiers trained using X_{clean}, X_{adapt}, and X_{deg} images are represented as DC_{clean}, DC_{adapt}, and DC_{deg} respectively. Besides, there are two other symbols utilized in our study, *i.e*., Δ describes the DDPM process for adaptation such as the one described in DDA [11] and E denotes the ensemble, which comprises a set of distinct classifiers defined as $E(,)$.

3.2. Background and Motivation

The performance of typical classifiers significantly drops due to unknown degradation. Hence, we employ DDPM to adjust degraded images towards the domain of clean images. We inherently assume that the adapted images domain is better than directly using unknown degraded images for classification. Indeed, previous studies like DDA [11] have shown that DDPM helps improve the performance of classifying unknown degraded images. The DDA method applies an ensemble of classifiers trained on clean images to the input of degraded and adapted images to resolve imperfect adapted images' limitations. Contrastingly, we train two separate classifiers on adapted and degraded images that substantially improve classification performance for both adapted and degraded images. In particular, a classifier trained on adapted images with a limited set of known degradations anticipates imperfections in the image, thereby contributing to the robustness of our proposed

method. Similarly, a classifier trained on degraded images of a few dissimilar known degradations helps our proposed method handle the degraded images directly. Moreover, our ensemble of specialized classifiers trained on adapted and degraded images is more natural than the ensemble of C_{clean} in the case of DDA [11].

3.3. Known Degradations Selection

Our choice of degradations for training, *i.e.*, additive white Gaussian noise (AWGN), Gaussian blur (GBlur), and JPEG compression (JPEG), boosts classifier robustness given their prevalence in typical image or camera processing pipelines. All three degradations are disparate from each other, *i.e.*, JPEG compression is a type of digital compression, while AWGN and GBlur are dissimilar types of noise and blurring mechanisms. Furthermore, diversifying the range of possible degradations for training assists classifiers in learning diverse characteristics. While we aim for the model to generalize across various unknown degradations, training on a limited set is a pragmatic choice due to numerous potential degradations.

3.4. Proposed Method: DiffAUD

We propose DiffAUD, *i.e.*, diffusion-based adaptation for unknown degraded images as described in Figure 2, where the top block shows the overall process for the classification of degraded images, which constitutes applying a diffusion model and an ensemble of distilled classifiers DC_{adapt} and DC_{deg} to get the final classification prediction. Furthermore, to apply ensemble, we take the sum of logits from the two classifiers before the softmax function and apply $argmax$ to predict the input image class similar to DDA [11]. Our proposed method is invariant to any DDPM; we currently show a rough reverse diffusion process from DDA [11] since we apply their DDPM adaptation process.

To summarize, the following steps outline the process flow of our proposed method:

1. Apply DDPM on the degraded images X_{deg} to yield adapted images X_{adapt}.
2. Feed adapted images X_{adapt} to a distilled classifier trained on adapted images from known degradations, *i.e.*, DC_{adapt} and in parallel, we input degraded images X_{deg} directly to a distilled classifier trained on known degradation images, *i.e.*, DC_{deg}.
3. Apply ensemble on the outputs of two distilled classifiers to output Y_{deg}.

3.5. Distilled Classifier Training

The bottom block in Figure 2 shows the detailed training process of distilled classifier DC where the training process is the same irrespective of adapted or degraded images. Our distillation process takes inspiration from FusionDistill [4]. The loss function of FusionDistill can be expressed as

$$L = L_{CE} + \alpha L_{CON}\,, \tag{1}$$

where L_{CE}, L_{CON}, and α represent the cross-entropy loss, the consistency loss for the knowledge distillation, and the loss weight, respectively. Here we use three types of synthetically prepared degraded or corresponding adapted images shown as X_{JPEG}, X_{GBlur}, and X_{AWGN} for each relevant degradation. We also pre-trained three teacher networks with those three different datasets: T_{JPEG}, T_{GBlur}, and T_{AWGN}. Then, the consistency loss L_{CON} is defined by

$$L_{CON} = \sum_{i} \text{COS}(\phi_{T_i}, \phi_S)\,, \quad i \in \{\text{JPEG}, \text{GBlur}, \text{AWGN}\}\,, \tag{2}$$

where ϕ_{T_i} represents the feature vector of the teacher network T_i, ϕ_S represents the feature vector of the student network S, and COS represents the function of cosine similarity. We evaluate the consistency loss L_{CON} between teacher and student networks intermediate features. Specifically, intermediate features are compared before the feed-forward layers of the respective same-depth/backbone networks using the consistency loss function. Moreover, we utilize cross-entropy loss L_{CE} for supervision between ground truth labels and prediction Y_{pred}. Eq. (1) shows the total loss equation, which incorporates supervision and consistency loss. We primarily have only one hyperparameter for distilled classifier tuning, *i.e.*, α, the weight of the L_{CON} loss, which shows the easy applicability of our distilled classifier method. Furthermore, we perform fine-tuning using the ResNet-50 backbone to find optimal α values for degraded and adapted images. In Section 4.1.1, we provide more detailed explanations about the training images, and supplementary material includes information about hyperparameter α tuning.

4. Experiments

4.1. Dataset Preparation

4.1.1 Training Dataset: Known Degradations

We synthetically apply three known degradations on clean images, *i.e.*, JPEG, AWGN, and GBlur resulting in degraded images. Explicitly, each degradation is uniformly sampled over a wide range of degradation levels so that the classifier can learn a wide distribution of severities for each known degradation. Notably, for JPEG compression, quality factors range $\in [1, 101]$ where 101 represents clean images; for AWGN, the standard deviation of Gaussian kernel $\in [0, 50]$; and for GBlur standard deviation of Gaussian kernel $\in [0, 5]$. We apply all three degradations on CIFAR-10 [23] and Imagenet-1k [38] datasets to evaluate respective

Table 1. Details for sequential degradations evaluation dataset where degradation sequence order is GBlur → AWGN → JPEG. GBlur, AWGN, and JPEG magnitudes represent the standard deviation of the Gaussian kernel with respect to the percentage of the image size, the standard deviation of the Gaussian kernel, and the quality factor for JPEG compression, respectively. Quality metrics show degradation quality in comparison to the clean images.

Severity levels	Degradations			Quality			
				Imagenet		CIFAR-10	
	GBlur	AWGN	JPEG	PSNR	SSIM	PSNR	SSIM
Weak	2	10	80	28.44	0.78	28.54	0.94
Medium	4	20	60	23.27	0.52	24.61	0.85
Strong	8	30	40	20.81	0.35	22.13	0.76

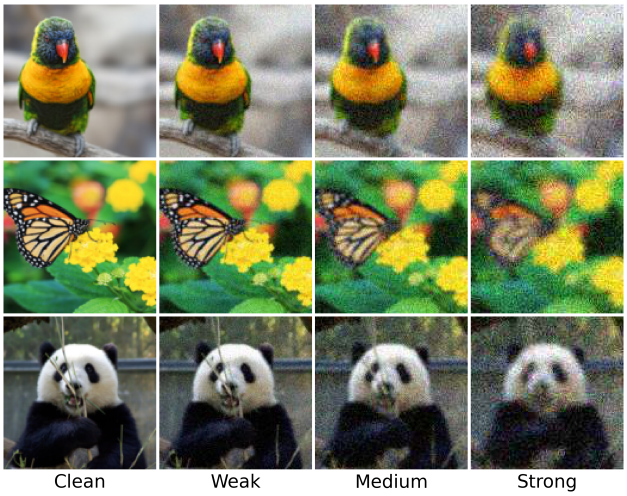


Figure 3. Sample images for sequential evaluation dataset with three severity levels corresponding to the Table 1.

datasets. It results in 150k images for training and 30k for validation for the CIFAR-10 dataset. However, due to the limited computational capacity, we use 60k images from Imagenet-1k for each known degradation, resulting in 150k images for training and 30k images for validation.

4.1.2 Evaluation dataset: Unknown Degradations

To assess the effectiveness of our proposed method, we conduct an extensive evaluation on two types of datasets, *i.e.*, single degradation and sequence of degradations datasets. First, we utilize widely used corruption datasets from Hendrycks *et al*. [17], *i.e.*, CIFAR-10-C, and Imagenet-C to measure the robustness of image classification models for single degradation. Specifically, we utilize the whole dataset for the CIFAR-10-C dataset. Due to the limited computational capacity, we utilize a randomly subsampled version of the Imagenet-C dataset containing five images for each class, in total, 5k per corruption per severity. All 15 corruptions and five severities were part of the evaluation from both CIFAR-10-C and Imagenet-C datasets, totaling 75 distinct corruptions. Moreover, to further strengthen our contributions, we prepare a dataset with a sequence of degradations denoted as SEQ-C similar to Moriyasu *et al*. [31]. Specifically, we introduce three levels of severity, *i.e.*, weak, medium, and strong, as described in Table 1 and sample images are shown in Figure 3. We derive the degradation sequence from a typical image processing pipeline while capturing an image in a camera, where blur can occur first due to a shifting camera. Next, noise can arise due to sensor-related components such as high ISO or environmental factors such as low light. Lastly, the image is stored using compression techniques such as JPEG.

4.2. Experiment Settings

4.2.1 Diffusion Models and Backbones

We utilize pre-trained unconditional diffusion models provided by OpenAI [7] trained on CIFAR-10 and Imagenet-1k dataset (256×256 resolution) to evaluate respective degraded images. Diffusion model hyperparameters to generate adapted images are the same as DDA [11] for both datasets, *i.e.*, diffusion ranges $N = 50$, scaling factor $D = 4$, and refinement range $w = 6$. Since the sampling process for image adaptation is computationally expensive and DDA is the closest work to our study, we use only the DDA method from a diffusion model perspective as a baseline method. However, to show the effectiveness of our proposed approach, we thoroughly perform experiments on several combinations of classifiers with disparate backbones and comprehensive evaluation on two datasets with a wide variety of degradations and severities.

We utilize frequently used ResNet-50 [14] for CIFAR-10 dataset evaluation. For Imagenet dataset evaluation, we follow the identical backbones as DDA, *i.e.*, ResNet-50 [14], ConvNeXt-Tiny [29], and Swin Transformer (Swin-Tiny) [28]. Furthermore, all the clean image classifiers are pre-trained.

4.2.2 Methods Comparison

Experiments are split into three categories in the experimental results shown in Table 2 and earlier visualized in Figure 1. The first category refers to "classifier only" methods, where there are three combinations, *i.e.*, C_{clean}, and C_{deg}, and DC_{deg} represent classifier trained on clean images, a classifier trained on degraded images, and distilled classifier trained on degraded images. Next, the "Δ + classifier" category refers to diffusion models with a classifier where the output adapted image from the diffusion model is input to the classifier. We compare three variations; the first is $\Delta + C_{clean}$ method, *i.e.*, DDA without ensemble [11], next $\Delta + C_{adapt}$ method where we apply the same diffusion process as DDA [11] and input the adapted images

Table 2. All methods comparison discussed in our work from different parameters perspective. Δ represents the DDPM process for adaptation, and the "Training images" column represents images utilized to train the respective classifiers. Next, "Distill" shows whether we utilize knowledge distillation [4] for training the network, and "Ensemble E" shows whether we utilize an ensemble of classifiers.

Method	DDPM Δ	Training images	Distill	Ensemble E
(*classifier only*)				
C_{clean}		clean		
C_{deg}		deg		
DC_{deg}		deg	✓	
(Δ + *classifier*)				
$\Delta + C_{clean}$†	✓	clean		
$\Delta + C_{adapt}$	✓	adapt		
$\Delta + DC_{adapt}$	✓	adapt	✓	
(Δ + *ensemble*)				
$\Delta + E(C_{clean}, C_{clean})$‡	✓	clean		✓
$\Delta + E(C_{adapt}, C_{deg})$	✓	adapt,deg		✓
$\Delta + E(DC_{adapt}, DC_{deg})$§	✓	adapt,deg	✓	✓

†, ‡, and § represents DDA without ensemble [11], DDA [11], and our proposed method DiffAUD respectively.

to C_{adapt} classifier. Only difference between $\Delta + C_{adapt}$ and $\Delta + DC_{adapt}$ is that $\Delta + DC_{adapt}$ contains distilled classifier DC_{adapt} instead of typical classifier. Afterwards, $\Delta + ensemble$ category where the first method $\Delta + E(C_{clean}, C_{clean})$ is our baseline method DDA [11], next is $\Delta + E(C_{adapt}, C_{deg})$ method where adapted images from Δ are input to the C_{adapt} and original degraded images are input to C_{deg}. At last, $\Delta + E(DC_{adapt}, DC_{deg})$ is our proposed method DiffAUD as described in Sec. 3.

4.3. Experimental Results

4.3.1 Single Degradation

For single degradation, first, we conduct the experiments on the CIFAR-10-C dataset with ResNet-50 backbone to evaluate the effectiveness of our proposed method in comparison with various combinations of classifiers and Δ as shown in Table 3. Subsequently, we carry out existing methods comparison on the Imagenet-C dataset with ResNet-50, Swin-Tiny, and ConvNeXt-Tiny backbones as shown in Tables 4a to 4c.

In Table 3, C_{clean} method attains good performance on a few natural degradations like brightness, contrast, fog, and snow. Nevertheless, it lags drastically on other degradations. Next, models trained on adapted images work well with degradations like Gaussian noise, glass blur, impulse noise, pixelate, shot noise, and zoom blur compared to those trained solely on clean images. Similarly, models trained on degraded images perform well with other degradation types. Combining the strengths of both adapted and degraded image models through ensemble technique enhances DiffAUD performance. While adaptation alone may not yield perfect results, ensemble integration boosts the overall effectiveness of our proposed method with an overall accuracy of 87.74%. On the other hand, DiffAUD and other methods trained based on known degradations include Gaussian blur for training; we hypothesize that this leads to improvement in the performance of four types of blur degradations included in the corruption dataset, *i.e.*, defocus blur, glass blur, motion blur, and zoom blur especially compared to the baseline C_{clean} method.

In Table 4a, on average, our proposed method performs 5.5% better than the baseline DDA [11] method. Additionally, our prepared method utilizing distillation and known degradations, *i.e.*, DC_{deg}, perform significantly better than baseline, *i.e.*, C_{clean} on ResNet-50. Next, In Table 4b and Table 4c, DC_{deg} achieves around 3%-4% improvement as compared to C_{clean} for both backbones, which is lower than ResNet-50 backbone performance improvement of 7.76% since Swin-Tiny and ConvNeXt-Tiny pre-trained clean image classifiers are comparatively more robust towards corruptions such as contrast, frost, and snow. While the distilled classifier (DC_{deg}) excels on a few degradations, such as brightness, contrast, and fog, our proposed method falls slightly short in these aspects. It highlights an opportunity

Table 3. Classification accuracy for CIFAR-10-C dataset with different corruptions averaged over all severities on ResNet-50 backbone where Δ represents DDPM and E represents the ensemble of particular classifiers. Italicized corruption columns represent corruptions utilized during classifier training.

Method	bright	contrast	defocus	elastic	fog	frost	*gauss*	glass	impulse	*jpeg*	motion	pixel	shot	snow	zoom	mean
(classifier only)																
C_{clean}	**93.89**	80.22	81.55	84.41	**88.19**	78.40	46.45	54.40	56.83	81.34	76.92	74.01	59.15	83.39	76.21	74.36
C_{deg}	91.87	**82.21**	90.95	86.74	87.56	**87.79**	88.65	78.39	86.71	89.13	85.63	87.00	90.03	86.13	89.91	87.25
DC_{deg}	92.06	81.85	**91.68**	86.86	87.93	87.51	88.70	76.52	86.68	88.93	85.91	85.62	90.12	86.32	**90.44**	87.14
(Δ + classifier)																
$\Delta + C_{clean}$†	78.06	49.68	79.43	76.37	53.81	71.30	80.59	78.15	79.49	79.88	77.27	80.29	80.74	74.78	78.85	74.58
$\Delta + C_{adapt}$	78.34	54.45	79.96	77.03	56.94	72.16	80.49	78.63	79.83	79.99	78.05	80.38	80.89	75.05	79.30	75.43
$\Delta + DC_{adapt}$	77.97	53.50	79.58	76.73	56.36	71.62	80.20	78.29	79.58	79.76	77.67	80.05	80.58	74.78	79.01	75.04
(Δ + ensemble)																
$\Delta + E(C_{clean}, C_{clean})$‡	89.23	68.56	85.79	84.01	76.43	79.74	80.44	78.47	81.09	84.41	82.44	84.17	82.51	82.66	83.35	81.55
$\Delta + E(C_{adapt}, C_{deg})$	90.87	80.42	90.49	87.48	85.87	86.77	89.04	**83.53**	87.73	89.29	86.87	**88.78**	90.11	86.15	89.60	87.53
$\Delta + E(DC_{adapt}, DC_{deg})$§	91.20	80.18	91.27	**87.84**	86.61	86.89	**89.31**	82.74	**87.95**	**89.32**	**87.30**	88.31	**90.30**	**86.54**	90.34	**87.74**

†, ‡, and § represents DDA without ensemble [11], DDA [11], and our proposed method DiffAUD respectively.

Table 4. Classification accuracy for Imagenet-C dataset with different corruptions averaged over all severities on several backbones. Italicized corruption columns represent corruptions utilized during classifier training. We compute the average of all training experiments using three different random seeds.

(a) ResNet-50

Method	bright	contrast	defocus	elastic	fog	frost	*gauss*	glass	impulse	*jpeg*	motion	pixel	shot	snow	zoom	mean
C_{clean}	65.94	35.73	35.30	43.72	41.94	37.24	33.92	26.34	29.37	56.71	35.67	51.76	32.12	31.36	35.55	39.51
C_{deg}	66.96	**38.04**	50.22	45.90	45.16	38.02	56.99	28.34	50.85	64.44	38.88	58.91	55.75	31.06	37.89	47.16
DC_{deg}	**67.25**	37.98	50.13	46.29	**45.51**	38.89	56.57	28.93	49.00	63.81	40.26	58.35	54.96	32.02	39.14	47.27
DDA [11]	63.19	32.54	34.46	50.26	37.57	42.61	57.80	37.12	56.47	59.66	36.47	59.24	57.64	36.17	37.16	46.56
Ours	66.54	35.11	**50.79**	**54.61**	41.53	**45.01**	**62.16**	**46.42**	**60.90**	**65.65**	**43.97**	**66.40**	**62.02**	**36.24**	**44.20**	**52.10**

(b) Swin-Tiny

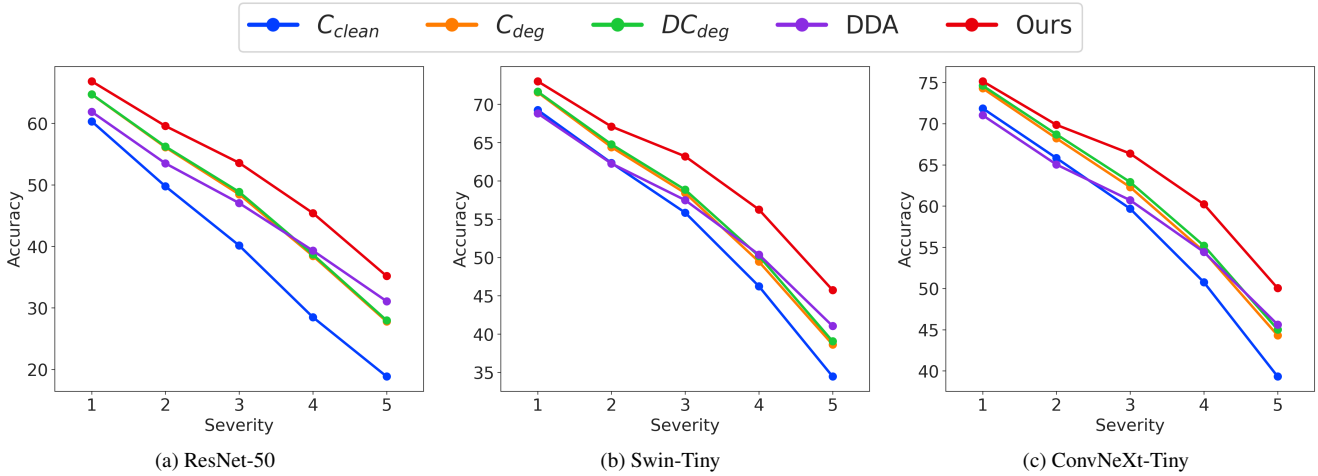
Method	bright	contrast	defocus	elastic	fog	frost	*gauss*	glass	impulse	*jpeg*	motion	pixel	shot	snow	zoom	mean
C_{clean}	73.43	61.26	44.70	51.90	61.27	**57.47**	55.56	32.30	52.38	62.52	49.73	56.12	53.07	**50.44**	42.30	53.63
C_{deg}	74.46	57.02	58.11	54.17	61.47	51.08	61.01	33.25	59.77	70.67	50.40	65.89	59.16	47.74	43.90	56.54
DC_{deg}	**74.53**	**57.98**	57.56	54.42	**61.81**	51.68	61.48	33.80	59.99	70.16	51.39	64.85	59.74	48.83	44.15	56.82
DDA [11]	70.75	57.18	43.50	56.87	55.44	57.04	63.54	43.25	62.02	63.41	48.18	62.77	62.85	49.86	43.27	56.00
Ours	73.72	56.36	**58.27**	**61.68**	58.71	56.76	**67.10**	**52.64**	**66.84**	**70.85**	**54.29**	**71.08**	**67.05**	49.83	**49.70**	**60.99**

(c) ConvNeXt-Tiny

Method	bright	contrast	defocus	elastic	fog	frost	*gauss*	glass	impulse	*jpeg*	motion	pixel	shot	snow	zoom	mean
C_{clean}	75.24	66.19	48.62	54.20	62.42	59.23	60.32	35.10	58.14	66.96	55.32	61.13	57.97	54.84	46.63	57.49
C_{deg}	76.83	66.23	58.59	55.05	66.87	57.40	67.36	34.89	65.54	**73.03**	54.92	67.07	66.44	53.57	46.86	60.71
DC_{deg}	**77.32**	**67.13**	58.78	55.72	**67.59**	58.57	67.61	35.66	66.35	72.80	55.89	67.56	66.85	**54.77**	46.98	61.31
DDA [11]	72.85	61.01	47.09	59.12	53.24	59.84	67.04	47.70	66.20	67.56	53.75	66.79	66.51	53.94	47.83	59.36
Ours	75.99	63.82	**60.67**	**62.72**	64.44	**60.82**	**70.22**	**54.89**	**69.84**	73.02	**58.12**	**73.32**	**69.84**	54.47	**52.78**	**64.33**

to enhance our ensemble approach and potentially prevent deterioration across various types of degradation. Nevertheless, while training distilled classifiers on known degradations, color-based augmentations can potentially resolve the low performance of brightness and contrast degradations. On the other hand, our proposed method can still outperform DDA [11] on both backbones with about the same performance gap of 5%.

To provide an in-depth view of the Imagenet-C dataset, we show accuracy over different severity levels averaged over all respective corruptions in Figure 4 with different backbones. With an increase in severity levels from 1 to 5, naturally, performance drops for all methods. On the ResNet-50 backbone, the performance of the DDA method becomes closer to the C_{clean} method towards low severity levels. Similarly, on Swin-Tiny and ConvNeXt-Tiny backbones, we can see similar patterns where C_{clean} performs almost similarly or, in fact, better on lower degradation levels than DDA. While DDA performs decently compared to C_{clean} on higher severity levels, severity levels are often unknown in real-world images, making the DDA method much more prone to performance reduction on lower severity levels. Next, performance of C_{deg} and DC_{deg} is very close to each other as previously shown in Tables 3 and 4a



(a) ResNet-50 (b) Swin-Tiny (c) ConvNeXt-Tiny

Figure 4. Performance with several backbones on Imagenet-C dataset with different severity levels averaged over all corruptions.

Table 5. Classification accuracy for Imagenet-SEQ-C sequential degradation dataset with different severity levels on ResNet-50, Swin-Tiny, and ConvNeXt-Tiny backbones. We compute the average of all training experiments using three different random seeds.

Method	ResNet-50				Swin-Tiny				ConvNeXt-Tiny			
	weak	medium	strong	mean	weak	medium	strong	mean	weak	medium	strong	mean
C_{clean}	46.61	19.28	68.09	44.66	53.51	31.84	69.91	51.75	61.28	41.10	75.33	59.24
C_{deg}	53.83	25.85	70.81	50.16	58.22	40.11	73.79	57.37	67.54	49.90	77.74	65.06
DC_{deg}	54.22	26.53	71.15	50.63	58.77	40.91	74.02	57.90	67.05	48.34	77.93	64.44
DDA [11]	58.78	45.26	68.51	57.52	60.33	49.08	69.97	59.79	66.46	55.70	74.80	65.65
Ours	**65.50**	**57.37**	**72.10**	**64.99**	**69.03**	**62.42**	**75.16**	**68.87**	**72.09**	**64.93**	**78.10**	**71.70**

to 4c; however, DC_{deg} performing slightly better on Swin-Tiny and ConvNeXt-Tiny backbones; showing the effectiveness of our distillation strategy for training classifiers.

On the other hand, our proposed method DiffAUD consistently performs drastically better on all severity levels as compared to C_{clean} as well as DDA methods, which shows that DiffAUD is invariant of lower adaptation quality from the diffusion models following the same diffusion process as DDA. It makes our work more significant toward achieving higher robustness and generalization, which can work with different diffusion models and adaptation processes.

4.3.2 Sequential Degradations

We evaluate the performance of sequential degradation images on ResNet-50, Swin-Tiny, and ConvNeXt-Tiny with the prepared sequential degradation dataset (Imagenet-SEQ-C) as shown in Table 5. In the classifier-only category of existing methods, similar to corruption datasets such as CIFAR-10-C and Imagenet-C, the performance of methods trained on known degradations is consistently better than clean images trained classifiers. It's worth noting that the distilled classifier method, *i.e.*, DC_{deg}, performs quite comparable to C_{deg}, suggesting that in sequences of degradations, the distillation method closely matches its non-distillation counterpart. However, that was not the case with single degradations, where distillation-based methods were slightly better than the non-distillation methods. We hypothesize that the resulting distribution from a sequence of degradations is discretely different from single degradations and that lead features learned from single degradations during distillation might not be greatly helpful in the case of sequential degradations.

Another critical insight to observe here is that DDA performs better than classifier-only methods on an average in the case of all backbones, specifically on ResNet-50 by a large gap and in the case of Swin-Tiny and ConvNeXt-Tiny backbone by a small gap given the Swin-Tiny and ConvNeXt-Tiny backbones are more robust towards corrupted images. It proves the effectiveness of the diffusion-based process introduced by DDA, which improves the performance in more realistic sequential degradations than classifier-only methods. On the other hand, that was not distinctly apparent in the case of a single corruption dataset such as Imagenet-C in the Table 4. Besides, our proposed method performs consistently well on all backbones compared to the other methods demonstrating the sturdy robustness of DiffAUD against sequential degradations with different severities.

5. Limitations

In our experiments, baseline classifiers use specific backbones, whereas our proposed method combines diffusion models with these backbones. Therefore, baseline classifier methods are much smaller in model complexity and parameters, as diffusion models typically have high latency and additional parameters. While our approach achieves better robustness compared to the other methods, it comes at the expense of increased model complexity. Nonetheless, recent works to enhance the efficiency of diffusion models [15, 43–45] suggest that a lighter diffusion model, combined with our proposed approach, could achieve improved performance without significant complexity overhead.

6. Conclusion

In this work, we propose DiffAUD, a novel diffusion-based adaptation for unknown degraded images using known degradations and distillation that complements diffusion models to classify unknown degraded images. We systematically perform comprehensive experiments on single degradation datasets such as Imagenet-C and CIFAR-10-C, and our own synthetically prepared sequential degradations dataset Imagenet-SEQ-C with various levels of severities, which shows that DiffAUD performs consistently well over existing methods. Moreover, we demonstrate that our proposed method based on known degradations and distillation drastically improves the performance for the classification of unknown degraded images.

Given the imperfections of diffusion-based adaptation in our proposed method, there is an opportunity to explore how image restoration techniques for unknown degradations complement our proposed approach in future research. Additionally, further exploration is needed to determine the most effective degradation methods utilized for training.

References

[1] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to end learning for self-driving cars, 2016. 1

[2] Tim Brooks, Bill Peebles, Connor Homes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Wing Yin Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 2

[3] Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated mri. *Medical Image Analysis*, 80: 102479, 2022. 2

[4] Dinesh Daultani and Hugo Larochelle. Consolidating separate degradations model via weights fusion and distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 440–449, 2024. 2, 4, 6

[5] Dinesh Daultani, Masayuki Tanaka, Masatoshi Okutomi, and Kazuki Endo. Iliac: Efficient classification of degraded images using knowledge distillation with cutout data augmentation. *Electronic Imaging*, 35(9):296–1, 2023. 2

[6] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2

[7] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *Neural Information Processing Systems*, 2021. 1, 5

[8] Samuel Dodge and Lina Karam. Understanding how image quality affects deep neural networks. In *Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2016. 1

[9] Kazuki Endo, Masayuki Tanaka, and Masatoshi Okutomi. Cnn-based classification of degraded images with awareness of degradation levels. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):4046–4057, 2021. 1, 2

[10] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark A. DePristo, Katherine Chou, Claire Cui, Greg S. Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature Medicine*, 25: 24 – 29, 2019. 1

[11] Jin Gao, Jialing Zhang, Xihui Liu, Trevor Darrell, Evan Shelhamer, and Dequan Wang. Back to the source: Diffusion-driven adaptation to test-time corruption. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11786–11796, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[12] Christopher Goodin, Daniel Carruth, Matthew Doude, and Christopher Hudson. Predicting the influence of rain on lidar in adas. *Electronics*, 8(1), 2019. 2

[13] Sorin Mihai Grigorescu, Bogdan Trasnea, Tiberiu T. Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37:362 – 386, 2019. 1

[14] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 5

[15] Yefei He, Luping Liu, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. Ptqd: Accurate post-training quantization for diffusion models, 2023. 8

[16] Robin Heinzler, Philipp Schindler, Jürgen Seekircher, Werner Ritter, and Wilhelm Stork. Weather influence and classification with automotive lidar sensors. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 1527–1534, 2019. 2

[17] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019. 5

[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[19] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, and Sergey Levine. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *ArXiv*, abs/1806.10293, 2018. 1

[20] Andreas Kamilaris and Francesc X. Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147:70–90, 2018. 1

[21] Samil Karahan, Merve Kilinc Yildirum, Kadir Kirtac, Ferhat Sukru Rende, Gultekin Butun, and Hazim Kemal Ekenel. How image degradations affect deep cnn-based face recognition? In *International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, 2016. 1

[22] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Neural Information Processing Systems*, 2022. 2

[23] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. 4

[24] Alfred Laugros, Alice Caplier, and Matthieu Ospici. Are adversarial robustness and common perturbation robustness independant attributes ? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019. 2

[25] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xiaocui Peng. All-in-one image restoration for unknown corruption. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17431–17441, 2022. 1

[26] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 2

[27] Jingyun Liang, Jie Cao, Guolei Sun, K. Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1833–1844, 2021. 1

[28] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, Los Alamitos, CA, USA, 2021. IEEE Computer Society. 1, 5

[29] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 5

[30] Daniel Maturana and Sebastian A. Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 922–928, 2015. 1

[31] Yu Moriyasu, Takashi Shibata, Masayuki Tanaka, and Masatoshi Okutomi. Top-k confidence map aggregation for robust semantic segmentation against unexpected degradation. In *IEEE International Conference on Consumer Electronics (ICCE)*, pages 1–6, 2023. 5

[32] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 1

[33] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning*, 2022. 2

[34] Yanting Pei, Yaping Huang, Qi Zou, Xingyuan Zhang, and Song Wang. Effects of image degradation and degradation removal to cnn-based image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4): 1239–1253, 2021. 1

[35] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[37] Prasun Roy, Subhankar Ghosh, Saumik Bhattacharya, and Umapada Pal. Effects of degradations on deep neural network architectures. *arXiv preprint arXiv:1807.10108*, 2018. 1, 2

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3): 211–252, 2015. 4

[39] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. *Neural Information Processing Systems*, 30, 2017. 2

[40] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2023. 2

[41] Tenta Sasaya, Takashi Watanabe, Takashi Ida, and Toshiyuki Ono. Simple self-distillation learning for noisy image classification. In *IEEE International Conference on Image Processing (ICIP)*, pages 795–799, 2023. 2

[42] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. 1

[43] Anwaar Ulhaq and Naveed Akhtar. Efficient diffusion models for vision: A survey, 2024. 8

[44] Zhendong Wang, Yifan Jiang, Huangjie Zheng, Peihao Wang, Pengcheng He, Zhangyang Wang, Weizhu Chen, and Mingyuan Zhou. Patch diffusion: Faster and more data-efficient training of diffusion models. In *Neural Information Processing Systems*, 2023.

[45] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, and L. Van Gool. Diffir: Efficient diffusion model for image restoration. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13049–13059. IEEE Computer Society, 2023. 8

[46] Yutong Xie and Quanzheng Li. Measurement-conditioned denoising diffusion probabilistic model for under-sampled medical image reconstruction. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 655–664, Cham, 2022. Springer Nature Switzerland. 2

[47] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Comput. Surv.*, 56(4), 2023. 2