Toward a Diffusion-Based Generalist for Dense Vision Tasks

Yue Fan[1,2]*, Yongqin Xian[2], Xiaohua Zhai[3], Alexander Kolesnikov[3], Muhammad Ferjad Naeem[4], Bernt Schiele[1], Federico Tombari[2,5]

[1]Max Planck Institute for Informatics, Saarland Informatics, [2]Google, [3]Google Deepmind, [4]ETH Zurich, [5]TU Munich

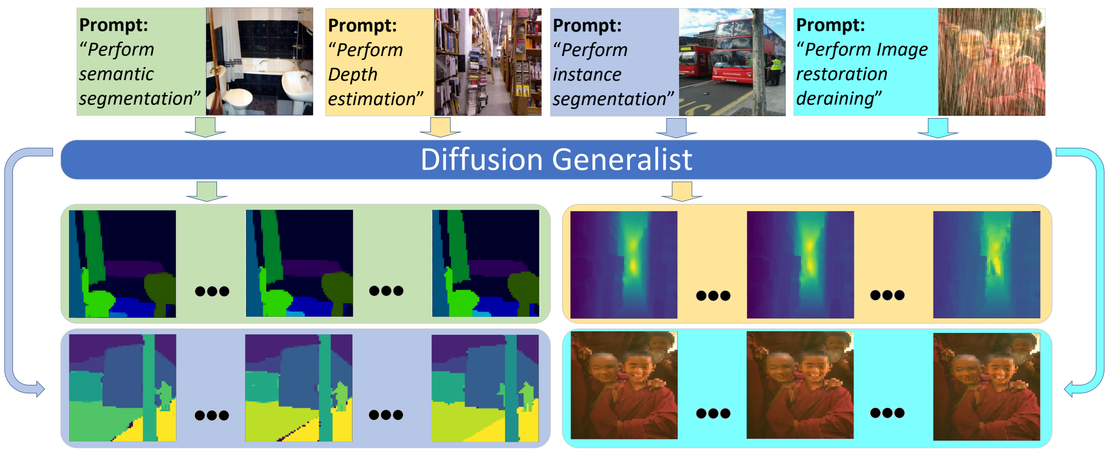


Figure 1. We present a diffusion-based vision generalist for dense vision tasks. Given an input image, the model performs the corresponding task following the text instruction. We showcase the effectiveness of our model on depth estimation, semantic segmentation, panoptic segmentation, and three types of image restoration tasks. The images are the actual output of our model.

Abstract

Building generalized models that can solve many computer vision tasks simultaneously is an intriguing direction. Recent works have shown image itself can be used as a natural interface for general-purpose visual perception and demonstrated inspiring results. In this paper, we explore diffusion-based vision generalists, where we unify different types of dense prediction tasks as conditional image generation and re-purpose pre-trained diffusion models for it. However, directly applying off-the-shelf latent diffusion models leads to a quantization issue. Thus, we propose to perform diffusion in pixel space and provide a recipe for finetuning pre-trained text-to-image diffusion models for dense vision tasks. In experiments, we evaluate our method on four different types of tasks and show competitive performance to the other vision generalists.

1. Introduction

The field of artificial intelligence has made significant progress in building generalized model frameworks. In particular, autoregressive transformers [27] have become a prominent unified approach in Natural Language Processing (NLP), effectively addressing a wide range of tasks with a singular model architecture [10, 19, 21, 25]. However, in computer vision (CV), building a unified framework remains challenging due to the inherent diversity of the tasks and output formats. Consequently, state-of-the-art computer vision models still have many complex task-specific designs [3, 8, 9, 15, 30], making it difficult for feature sharing across tasks and, thus, limiting knowledge transfer.

The stark contrast between NLP and CV has given rise to a growing interest in developing unified approaches for vision tasks [6, 7, 18, 28, 29, 34]. Recently, [28, 29] have shown image itself can be used as a robust interface for unifying different vision tasks and demonstrated good performance. In this paper, we propose a multi-task diffusion generalist for dense vision tasks by reformulating the dense prediction tasks as conditional image generation, and re-purpose pre-trained latent diffusion models for it. Fig. 1 visualizes the output of our model on semantic segmentation, panoptic segmentation, depth estimation, and image restoration. Based on text prompts, our model can perform different tasks with one set of parameters. However, directly finetuning the pre-trained latent diffusion models

*Intern at Google during the project.

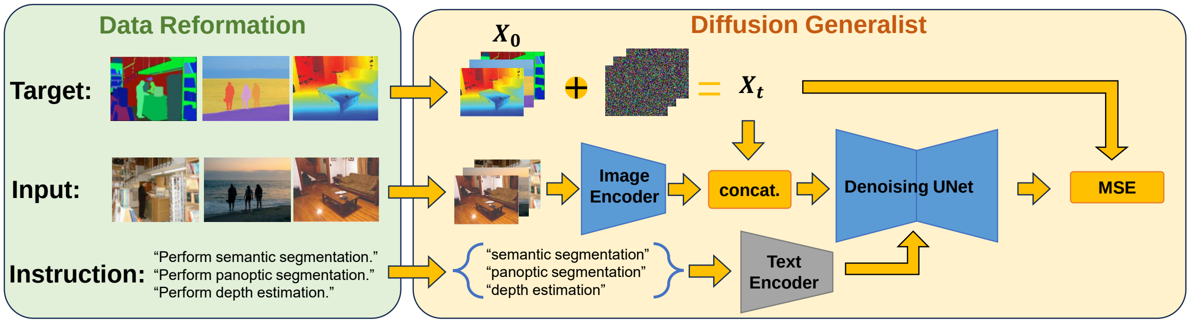


Figure 2. The training pipeline of the diffusion-based vision generalist consists of two parts: **Left:** Redefining the output space of different vision tasks as RGB images so that they can be unified under a conditional image generation framework. **Right:** We finetune a pre-trained diffusion model on the reformatted data from the first step. Diffusion is performed in the pixel space to mitigate the quantization error of the latent diffusion (see Table 3). The image and text conditionings are fed into the model via the corresponding encoders, where only the image encoder is tuned during the training.

(e.g. Stable Diffusion [22]) leads to quantization errors for segmentation tasks (see Table 3). To this end, we propose to do pixel-space diffusion which effectively improves the generation quality and does not suffer from quantization errors. Moreover, our exploration into training diffusion models as vision generalists reveals a list of interesting findings as follows:

- Diffusion-based generalists show superior performance over the non-diffusion-based generalists on tasks involving semantics or global understanding of the scene.
- We find conditioning on the image feature extracted from powerful pre-trained image encoders results in better performance than directly conditioning on the raw image.
- Pixel diffusion is better than latent diffusion as it does not have the quantization issue while upsampling.
- We observe that text-to-image generation pre-training stabilizes the training and leads to better performance.

In experiments, we demonstrate the model's versatility across six different dense prediction tasks on depth estimation, semantic segmentation, panoptic segmentation, image denoising, image draining, and light enhancement. Our method achieves competitive performance to the current state-of-the-art in many settings.

2. Related Work

Efforts have been made to unify various vision tasks with a single model, resulting in several vision generalists [6, 7, 18, 28, 29]. Inspired by the success of sequence-to-sequence modeling in Natural Language Processing (NLP), Pix2Seq [6] casts the output of object detection as sequences of discrete tokens and models them through next token prediction. The idea was further developed in Pix2SeqV2 [7] and Unified-IO [18], where dense prediction tasks such as segmentation, depth map, and image restoration are also unified either using features from a vector quantization variational auto-encoder (VQ-VAE) [26] or the coordinates of object polygons [4]. Painter [28] and SegGPT [29], on the other hand, reformulate vision tasks as an image inpainting problem, and perform in-context learning following [2]. Unlike the previous work, our method unifies different vision tasks under a conditional image generation framework and introduces a diffusion-based vision generalist for it.

3. Toward a Diffusion-Based Generalist

3.1. Unification with Conditional Image Generation

As the output of most vision tasks can be always visualized as images, we redefine the output space of different vision tasks as RGB images and unify them as conditional image generation to tackle the inherent difference of output formats of different vision tasks. Given a input image x and the corresponding ground-truth y, we first transform y into RGB images and then pair it with a task descriptor in text. By doing so, training sets of different tasks are combined into a holistic training set. And training the model jointly on it enables the knowledge transfer between tasks. At test time, given a new image, the model can perform different tasks following the text instructions (examples in Fig. 1).

In this paper, we consider four types of dense prediction tasks: depth estimation, semantic segmentation, panoptic segmentation, and image restoration.

Depth estimation outputs real number depth value for each pixel on x. Given the minimum and the maximum values, we map them into $[0, 255]$ linearly and discretize them into integers, which is then repeated and stacked along the channel to form the ground-truth RGB label.

Semantic segmentation predicts a class label for each pixel. We use a pre-defined injective class-to-color mapping to transform the segmentation mask into RGB images. Given a task with C categories, we define C colors which are evenly distributed in the 3-dimensional RGB space. Specifically, following [28], the class index is represented by a 3-digit number with b-base system, where $b = \lceil C^{\frac{1}{3}} \rceil$. Thus, the margin between two colors is defined as $\text{int}(\frac{256}{b})$. The color for the i-th class is then $[\text{int}(\frac{i}{b^2}) \times m,$

	Depth Estimation RMSE ↓ NYUv2	Semantic Seg. mIoU ↑ ADE-20K	Panoptic Seg. PQ ↑ COCO	Denoising SSIM ↑ SIDD	Deraining SSIM ↑ 5 datasets	Light Enhance. SSIM ↑ LoL
Ours	0.511	**48.0%**	**35.5%**	0.949	0.772	**0.704**
Non-diffusion	**0.443**	42.4%	19.8%	**0.951**	**0.773**	0.703
Train from scratch	0.528	46.6%	33.6%	0.948	0.764	**0.704**
Direct concat.	0.476	37.6%	27.1%	0.941	0.772	0.687

Table 1. We analyze the important design choices of our method and aim to provide a recipe for training diffusion-based generalists: 1. diffusion models greatly outperform non-diffusion models on panoptic segmentation; 2. text-to-image generation pre-training leads to an overall better performance; 3. conditioning on image features extracted from an encoder gives significant improvement over the raw image.

$\mathrm{int}(\frac{i}{b})\%b \times m, l\%b \times m]$. At test time, we find the nearest neighbor of the predicted color in the predefined class-to-color mapping and predict the corresponding category.

Panoptic segmentation is solved as a combination of semantic and instance segmentation. Semantic segmentation labels are constructed as stated above. For instance segmentation, we set N as the maximum number of instances a single training image can contain. Then, we define N colors which are evenly distributed in the 3-dimensional RGB space as in semantic segmentation. Finally, we assign colors to objects based on their spatial location to form the RGB ground-truth label. For example, the instance whose center is at the upper leftmost corner obtains the first color and the lower rightmost gets the last color. At test time, the model makes predictions twice with different text instructions and merge the results for panoptic segmentation.

Image restoration aims to predict the clean image from corrupted images. Thus, the output space is inherently RGB image and does not need further transformation to fit in the framework.

3.2. A Diffusion Multi-Task Generalist Framework

By reformating the output space of different vision tasks into images, it is natural to solve them together under a conditional image generation framework. To this end, we leverage the powerful diffusion models pre-trained for image generation and re-purpose them in our use case.

Fig. 2 shows the overall pipeline of the method, which is a conditional image generation framework with pixel-space diffusion. The input is the noised target image X_0 together with the image conditioning and text conditioning. Compared to the commonly used diffusion pipeline for conditional image generation, there are two notable differences:

- We propose to directly perform diffusion in the pixel space. As shown in Table 3, when mapping from the latent space to the pixel space, visually uniform regions actually have pixels of many different RGB values. This variance can lead to inaccurate class mappings, and consequently, suboptimal performance for semantic and panoptic segmentation.
- The image conditioning is provided via a feature extractor (we use ConvNeXt [17]) and is concatenated to the target image X_0. Compared to the widely adopted method of directly concatenating the raw image as the condition, this brings significant performance improvement, especially for semantic and panoptic segmentation (see Table 1 for ablation).

4. Experimental Results

Here, we first explain experimental settings in Section 4.1. Then, we highlight important design choices in diffusion-based multi-task generalists in Section 4.2 before comparing our method with previous approaches in Section 4.3.

4.1. Datasets and Implementation Details

Datasets: We evaluate our method on six different dense prediction tasks with various output formats. For depth estimation, we use NYUv2 [24] and report the Root Mean Square Error (RMSE). For semantic segmentation, we evaluate on ADE20K [33] and adopt the widely used metric of mean IoU (mIoU). For panoptic segmentation, we use MS-COCO [16] and report panoptic quality as the measure. During inference, the model is forwarded twice for each validation image with different instructions to obtain the results of semantic and instance segmentation respectively. The outputs are then merged together into the panoptic segmentation. Image restoration tasks are evaluated on several popular benchmarks, including SIDD [1] for image denoising, LoL [31] for low-light image enhancement, and 5 merged datasets [32] for deraining.

Implementation details. As mentioned above, we take the Stable Diffusion v1.4 [22] checkpoint and finetune it jointly on six tasks. The image feature extractor is an ImageNet-21K [23] pre-trained ConvNeXt-Large [17]. The text encoder is Open-CLIP [20], which is used in Stable Diffusion [22]. We adopt uniform sampling for each tasks except panoptic segmentation, whose weight is twice as much as the other tasks (as it is a combination of semantic and instance segmentation). Following [5], we also adjust the input scaling factor by a constant factor b in the forward noising processing of diffusion. We use AdamW optimizer [13] with constant learning rate of 0.0001, linearly warmed up in the first 20,000 iterations. The target image resolution is 128×128 while the conditioning image resolution is 512×512. We train our model for 180,000 steps in total with a batch size of 1024.

	Target image resolution	**Depth Estimation** RMSE ↓ NYUv2	**Semantic Seg.** mIoU ↑ ADE-20K	**Panoptic Seg.** PQ ↑ COCO	**Denoising** SSIM ↑ SIDD	**Deraining** SSIM ↑ 5 datasets	**Light Enhance.** SSIM ↑ LoL
Generalist framework, task-specific models							
UViM [14]	512×512	0.467	-	45.8%	-	-	-
Generalist models							
Unified-IO [18]	256×256	0.385	25.7%	-	-	-	-
InstructCV [11]	256×256	0.297	47.2%	-	-	-	-
Painter [28]	448×448	**0.288**	**49.9%**	**43.4%**	**0.954**	**0.868**	**0.872**
Painter [28]	128×128	0.435†	28.4%†	22.6%†	0.922†	0.626†	0.773†
Ours	128×128	0.448	48.7%	40.3%	**0.954**	0.815	0.758

Table 2. Our method achieves competitive performance in most of the tasks while trained at a much smaller target resolution of 128×128. When compared at the same resolution, our method shows superior performance over the previous best method (Painter [28]), especially on semantic segmentation and panoptic segmentation. The best number is in bold and the second best number is underscored. †indicates numbers from our reproduction.

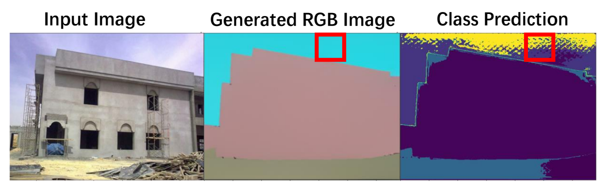
4.2. Recipes for Diffusion-Based Generalists

In this section, we analyze the design choices of our method and show their importance through ablation experiments. Specifically, we show the importance of diffusion by training the same model as in Fig. 2 to directly generate target images without using diffusion (non-diffusion). We study the significance of image generation pre-training and image encoder by training models without them (train from scratch and direct concat.). If not specified, we train all models at a target resolution of 64×64 for 50,000 steps.

We attribute the success of our method to four aspects. (1) While having similar results on image restoration tasks, diffusion-based generalist achieves better performance than non-diffusion models on segmentation tasks which requires a global understanding of the scene and the semantics. For example, the diffusion model reaches 35.5% PQ for panoptic segmentation while the non-diffusion model has only 19.8% (Table 1 ours v.s. non-diffusion). (2) Image generation pre-training on large scale dataset transfers useful knowledge to the many downstream tasks. The model fine-tuned from Stable Diffusion v1.4 [22] achieves better results than the one trained from scratch across the tasks (Table 1 ours vs train from scratch). (3) The image conditioning can take advantage of powerful pre-trained image encoders by conditioning on the image features rather than the raw image, which is in contrast to the standard practice for image generation tasks. On semantic segmentation and panoptic segmentation, extracting features gives 10.4% and 8.4% performance improvement, respectively (Table 1 ours v.s. direct concat.). (4) Pixel diffusion is better than latent diffusion as it does not suffer from the quantization issue while upsampling (see Table 3 for an example).

4.3. Comparisons with Prior Art

We compare our model with recent vision generalists in Table 2. With a much smaller target image resolution at 128×128, our method achieves competitive performance across the tasks. In particular, when compared with the previous best model Painter [28] at the same target resolution, our method has a significant margin over them, which highlights the potential of our method at a higher resolution.



	Semantic Seg. mIoU ↑ ADE-20K	**Panoptic Seg.** PQ ↑ COCO
Latent Diffusion	17.1%	11.7%
Pixel Diffusion	48.0%	35.5%

Table 3. **Upper:** Semantic segmentation output of the latent diffusion model. The perceptually same colored regions have different pixel values and, therefore, are mapped to different class labels, leading to bad final performance. While the red box contains only one ground-truth class sky in generated RGB image, the final class prediction has four classes after the quantization. **Lower:** Latent diffusion suffers from the quantization issue while pixel diffusion achieves good performance.

5. Conclusion and Limitations

In this work, we explore a diffusion-based vision generalist, where different dense prediction tasks are unified as conditional image generation and we re-purpose pre-trained diffusion models for it. Furthermore, we analyze different design choices of diffusion-based generalists and provide a recipe for training such a model. In experiments, we demonstrate the model's versatility across six different dense prediction tasks and achieve competitive performance to the current state-of-the-art. This work, however, is also subject to limitations. For example, full fine-tuning of the pre-trained diffusion model at a larger target image resolution is memory intensive due to the pixel space diffusion. Thus, exploring parameter-efficient tuning for such a model would be an interesting future direction.

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *CVPR*, 2018. 3

[2] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *NeurIPS*, 2022. 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1

[4] Lluis Castrejon, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Annotating object instances with a polygon-rnn. In *CVPR*, 2017. 2

[5] Ting Chen. On the importance of noise scheduling for diffusion models. *arXiv preprint*, 2023. 3, 1

[6] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv preprint*, 2021. 1, 2

[7] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey E Hinton. A unified sequence interface for vision tasks. *NeurIPS*, 2022. 1, 2

[8] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *NeurIPS*, 2021. 1

[9] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 1

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, 2018. 1

[11] Yulu Gan, Sungwoo Park, Alexander Schubert, Anthony Philippakis, and Ahmed M Alaa. Instructcv: Instruction-tuned text-to-image diffusion models as vision generalists. *arXiv preprint*, 2023. 4

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 1

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint*, 2014. 3

[14] Alexander Kolesnikov, André Susano Pinto, Lucas Beyer, Xiaohua Zhai, Jeremiah Harmsen, and Neil Houlsby. Uvim: A unified modeling approach for vision with learned guiding codes. *NeurIPS*, 2022. 4

[15] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint*, 2022. 1

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3

[17] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 3

[18] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. In *ICLR*, 2022. 1, 2, 4

[19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019. 1

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICLR*, 2021. 3

[21] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020. 1

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 4

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 3

[24] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 3

[25] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint*, 2023. 1

[26] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 2017. 2

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 1

[28] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *CVPR*, 2023. 1, 2, 4

[29] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint*, 2023. 1, 2

[30] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, 2022. 1

[31] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint*, 2018. 3

[32] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for fast image restoration and enhancement. *PAMI*, 2022. 3

[33] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 3

[34] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *CVPR*, 2022. 1

Toward a Diffusion-Based Generalist for Dense Vision Tasks
Supplementary Material

A. Ablation Study

In this section, we analyze the effect of other important hyper-parameters of our method, such as batch size, target image resolution, and noise-signal ratio. Similar to Section 4.2, we train all models at a target resolution of 64×64 for 50,000 steps by default.

Effect of batch size. Here, we discuss the effect of different batch sizes for our method. As shown in Table 4, the performance of most of the tasks improves with the increase of the batch size. In particular, panoptic segmentation greatly benefits from the large batch size.

	Depth RMSE ↓ NYUv2	**Sem. Seg.** mIoU ↑ ADE-20K	**Pan. Seg.** PQ ↑ COCO	**Denoise** SSIM ↑ SIDD	**Detrain** SSIM ↑ 5 datasets	**Enhance.** SSIM ↑ LoL
128	0.548	35.5%	26.2%	0.941	0.754	0.701
256	0.495	44.3%	30.0%	0.945	0.766	0.703
512	**0.491**	47.1%	33.5%	0.948	0.770	0.702
1024	0.511	**48.0%**	**35.5%**	**0.949**	**0.772**	**0.704**

Table 4. Large batch size improves the performance for all the tasks except depth estimation.

Effect of target resolution. Table 5 studies the effect of different target image resolutions. Since our method performs diffusion in the pixel space, increasing the target image resolution is important for good performance. Despite the increased memory cost, our method achieves its best performance at the resolution of 128×128 and can be further improved with even larger target images.

	Depth RMSE ↓ NYUv2	**Sem. Seg.** mIoU ↑ ADE-20K	**Pan. Seg.** PQ ↑ COCO	**Denoise** SSIM ↑ SIDD	**Detrain** SSIM ↑ 5 datasets	**Enhance.** SSIM ↑ LoL
32x32	0.514	44.4%	32.1%	0.940	0.743	0.653
64x64	0.511	48.0%	35.5%	0.949	0.772	0.704
128x128	**0.467**	**49.2%**	**36.7%**	**0.953**	**0.810**	**0.762**

Table 5. Effect of output resolution. Increasing the target image resolution significantly improves the performance across tasks.

Importance of noise-signal ratio. In DDPM [12], the forward diffusion process is defined as $x_t = \sqrt{\gamma_t} x_0 + \sqrt{1-\gamma_t}\epsilon$, where x_0 is the input image, ϵ is a Gaussian noise, and t is the number of diffusion step. As shown in [5], the denoising task at the same noise level (i.e. the same t) becomes simpler with the increase in the image size. In order to compensate for this, [5] proposed to scale the input with a constant b to explicitly control the noise-signal ratio, which results in the forward diffusion process as $x_t = \sqrt{\gamma_t} b x_0 + \sqrt{1-\gamma_t}\epsilon$. As we reduce b, it increases the noise levels. Table 6 shows the effect of the noise-signal ratio b where $b = 0.5$ gives the best performance.

	Depth RMSE ↓ NYUv2	**Sem. Seg.** mIoU ↑ ADE-20K	**Pan. Seg.** PQ ↑ COCO	**Denoise** SSIM ↑ SIDD	**Detrain** SSIM ↑ 5 datasets	**Enhance.** SSIM ↑ LoL
0.1	**0.497**	46.9%	33.1%	0.948	0.770	0.702
0.3	0.511	48.0%	35.5%	**0.949**	0.772	0.704
0.5	0.514	**49.3%**	**35.9%**	**0.949**	**0.774**	**0.708**
0.7	0.533	48.2%	34.4%	**0.949**	0.773	0.707
1.0	0.572	40.3%	31.1%	0.948	0.770	0.706

Table 6. Importance of noise-signal ratio b in the forward diffusion process $x_t = \sqrt{\gamma_t} b x_0 + \sqrt{1-\gamma_t}\epsilon$.

B. Qualitative Results

In this section, we visualize the output of our method on six different tasks in figs. 3 to 8. We use DDIM at inference time with 50 steps. Each figure shows the output of the denoising process at the 0-th, 25-th, and 50-th steps.

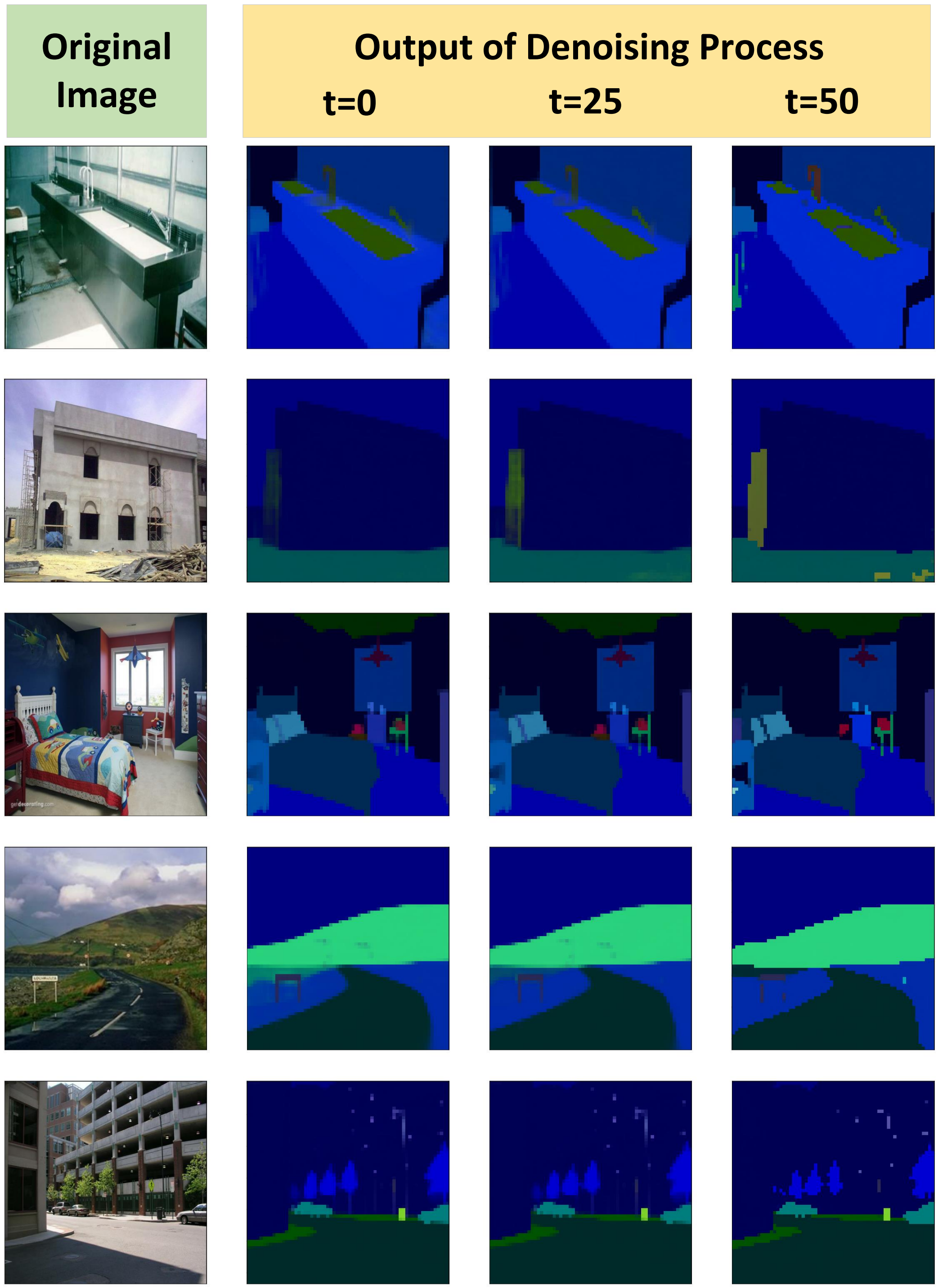


Figure 3. Qualitative results on images from ADE20K validation set. The text prompt is ”Performance semantic segmentation”. The images are not cherry-picked.

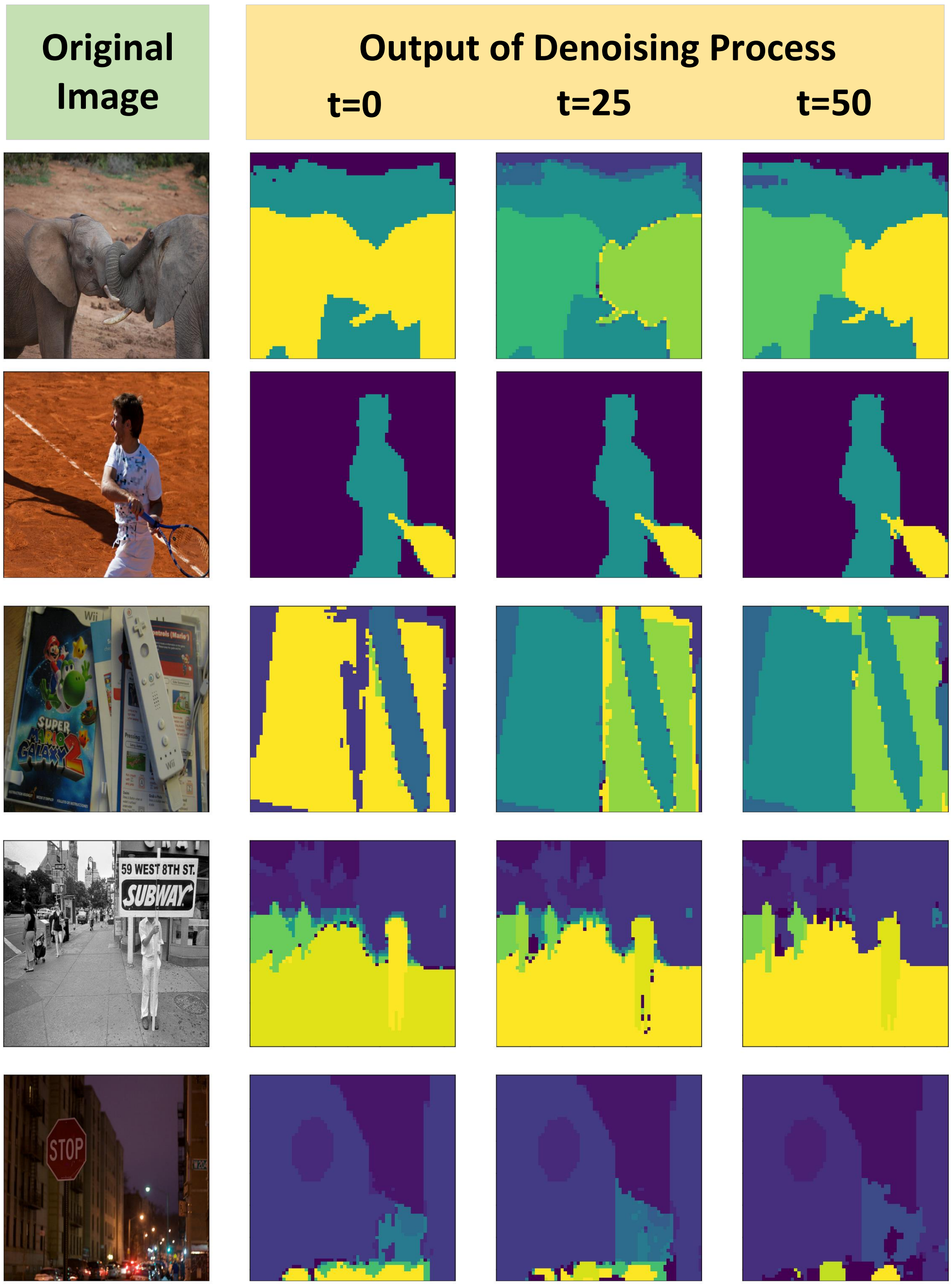


Figure 4. Qualitative results on images from MS-COCO validation set. The text prompt is ”Performance instance segmentation”. The images are not cherry-picked.

Figure 5. Qualitative results on images from NYU-V2 validation set. The text prompt is ”Performance depth estimation”. The images are not cherry-picked.

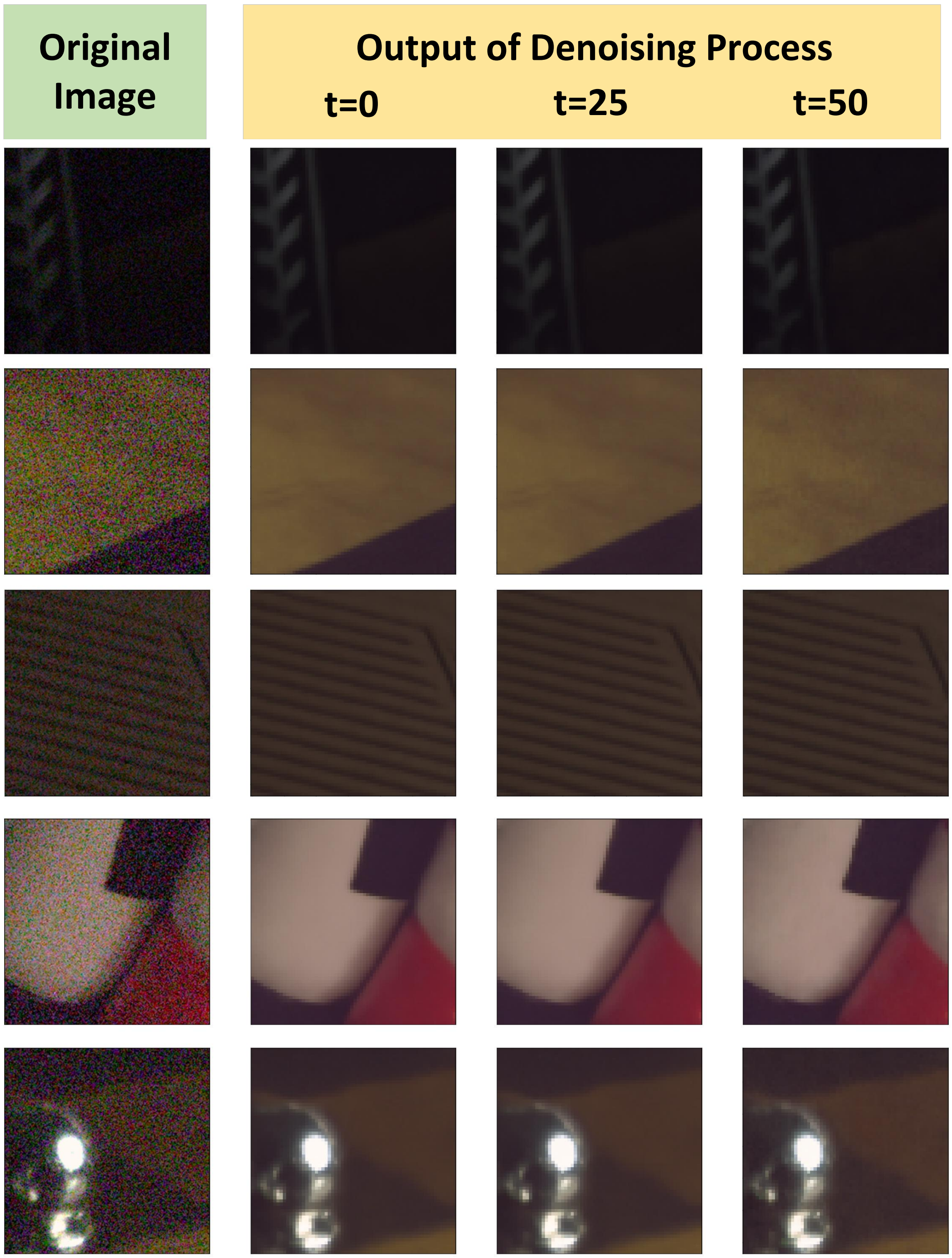


Figure 6. Qualitative results on images from SIDD validation set. The text prompt is ”Performance image restoration denoising”. The images are not cherry-picked.

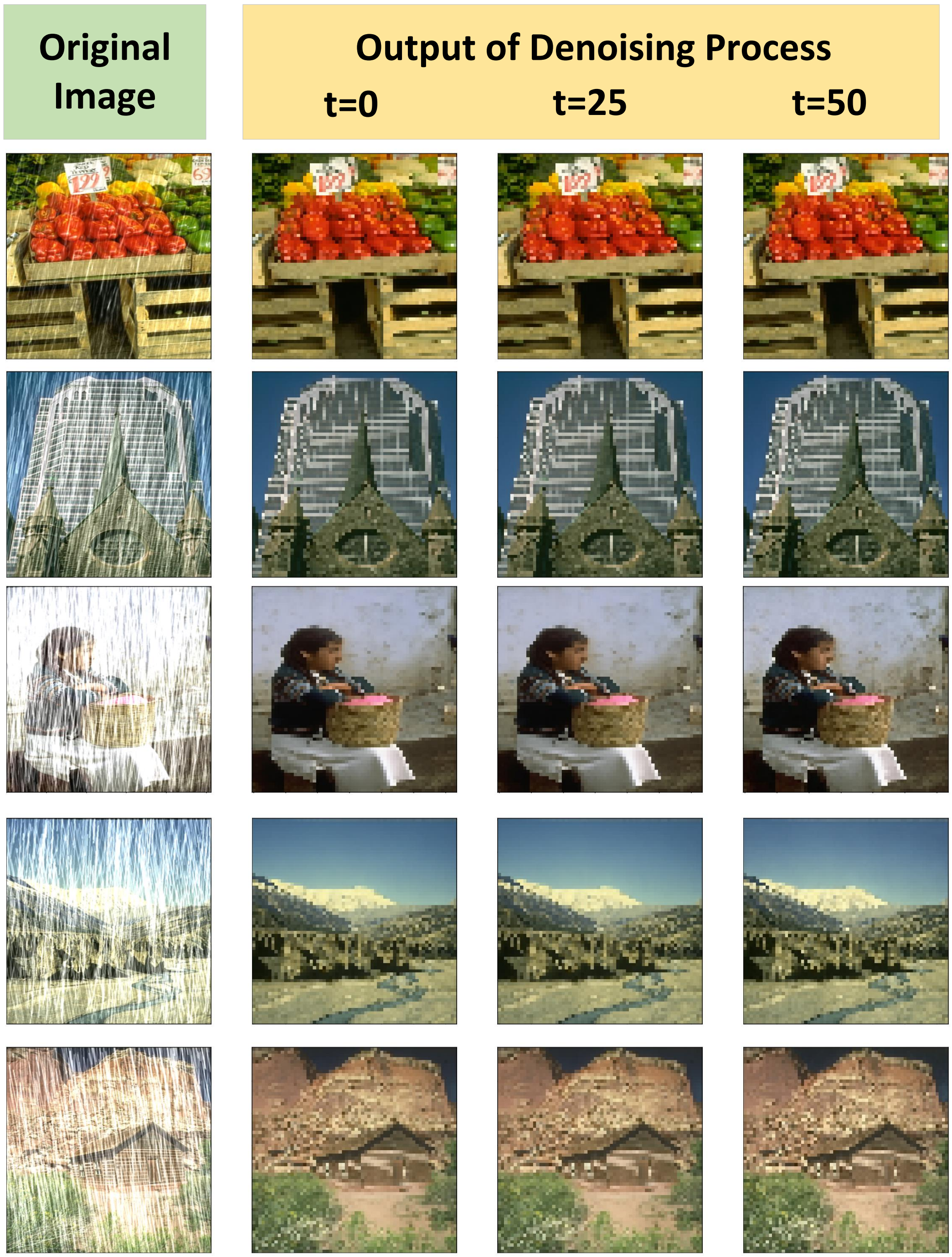


Figure 7. Qualitative results on images from Deraining datasets' validation sets. The text prompt is ”Performance image restoration deraining”. The images are not cherry-picked.

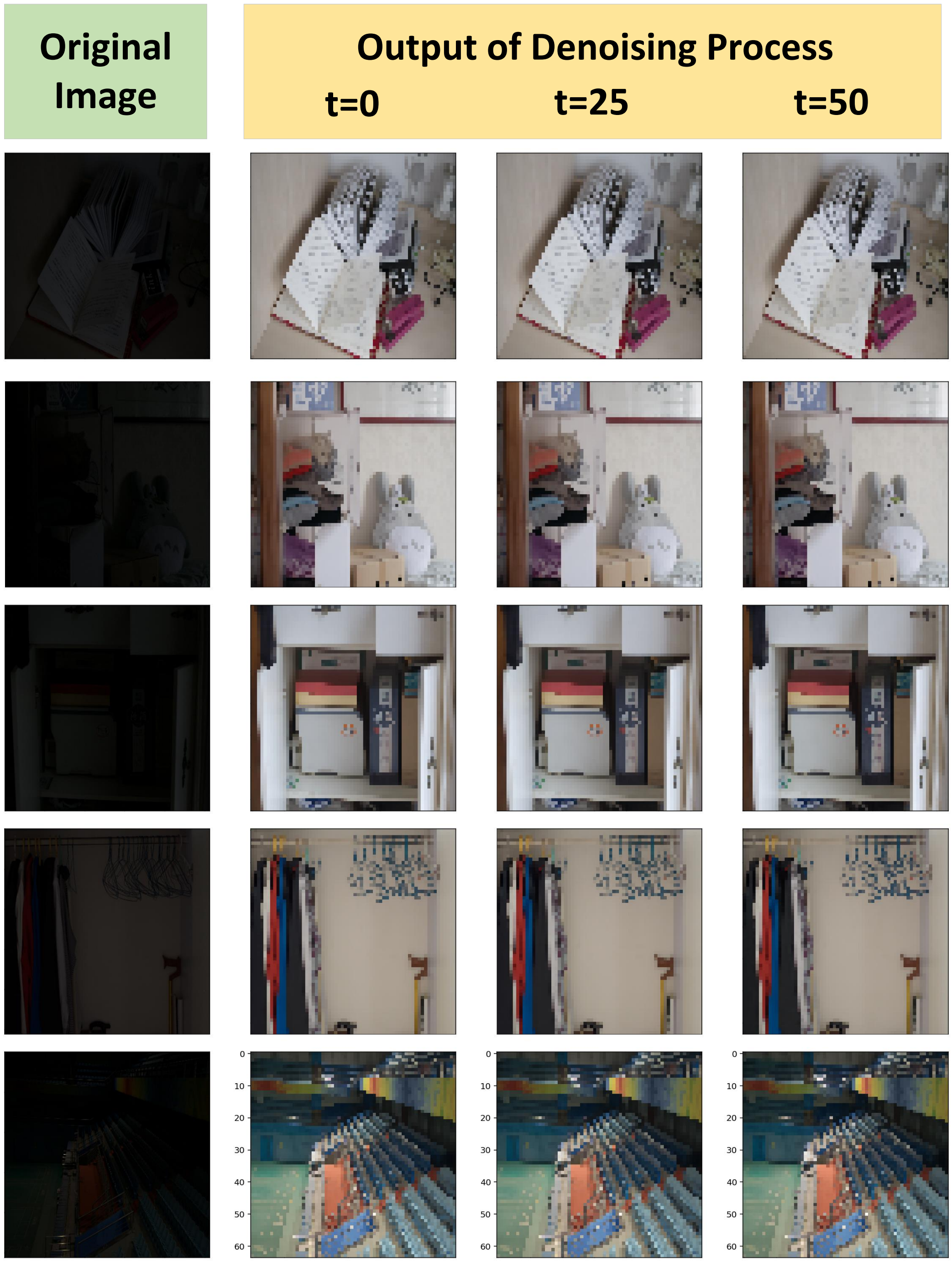


Figure 8. Qualitative results on images from LOL validation set. The text prompt is ”Performance image restoration light enhancement”. The images are not cherry-picked.