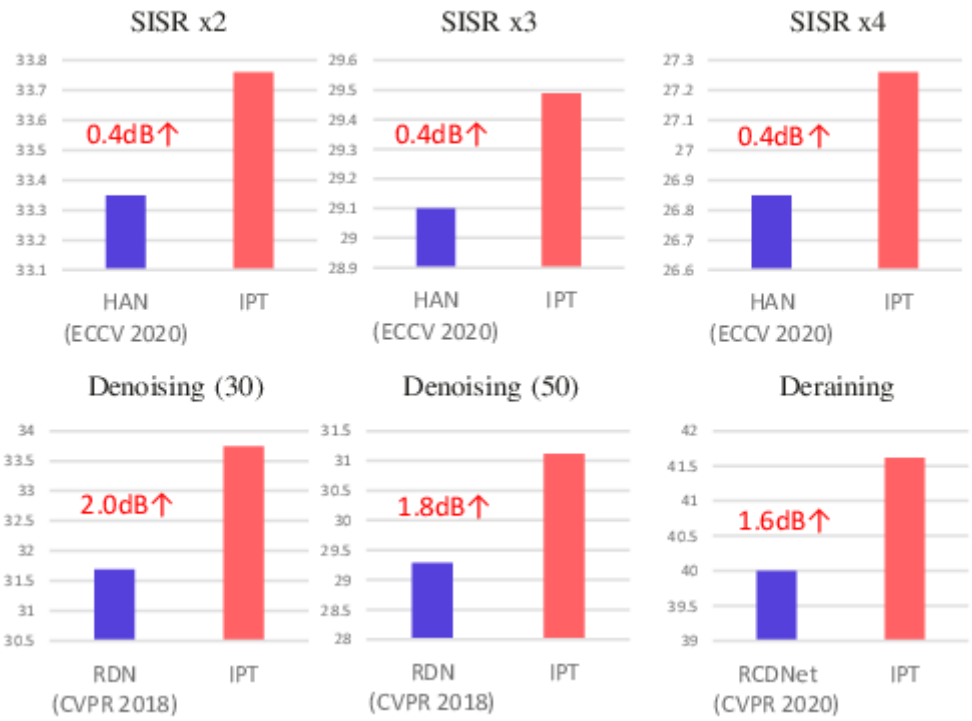
预训练图像处理Transformer

Abstract

随着现代硬件的计算能力的日渐增长，预训练深度学习模型（例如 BERT, GPT-3）在大规模数据集上的学习表现出了超过传统方式的效果。这个巨大进步主要提高了 transformer 及其变体架构的表现。本文研究了底层的计算机视觉任务（如去噪、超分辨率和去雨）并提出了一种新的预训练模型，即 image processing transformer（IPT）。为了最大限度地挖掘 transformer 的性能，我们提出利用知名的 ImageNet 生成大量损坏的图像对。IPT 模型使用多头注意力和多尾注意力在这些图片上进行训练。此外，我们引入了对比学习来更好的适应不同的图像处理任务。因此，预训练模型可以在经过微调后有效地用于所需任务。由于只有一个预训练模型，IPT 在各类底层基准上优于现有最先进的方法。

1. Introduction

图像处理是全局图像分析或计算机视觉系统的底层组成部分。图像处理的结果对后续高阶识别和理解图像数据有很大影响。近来，深度学习广泛应用于解决底层视觉任务，如图像超分辨率、修复、去雨和着色。因为很多图像处理任务是相关的，自然期望一个基于数据集预训练的模型能对另一个数据集有作用。但是很少有关于图像处理任务的泛化预训练的研究。



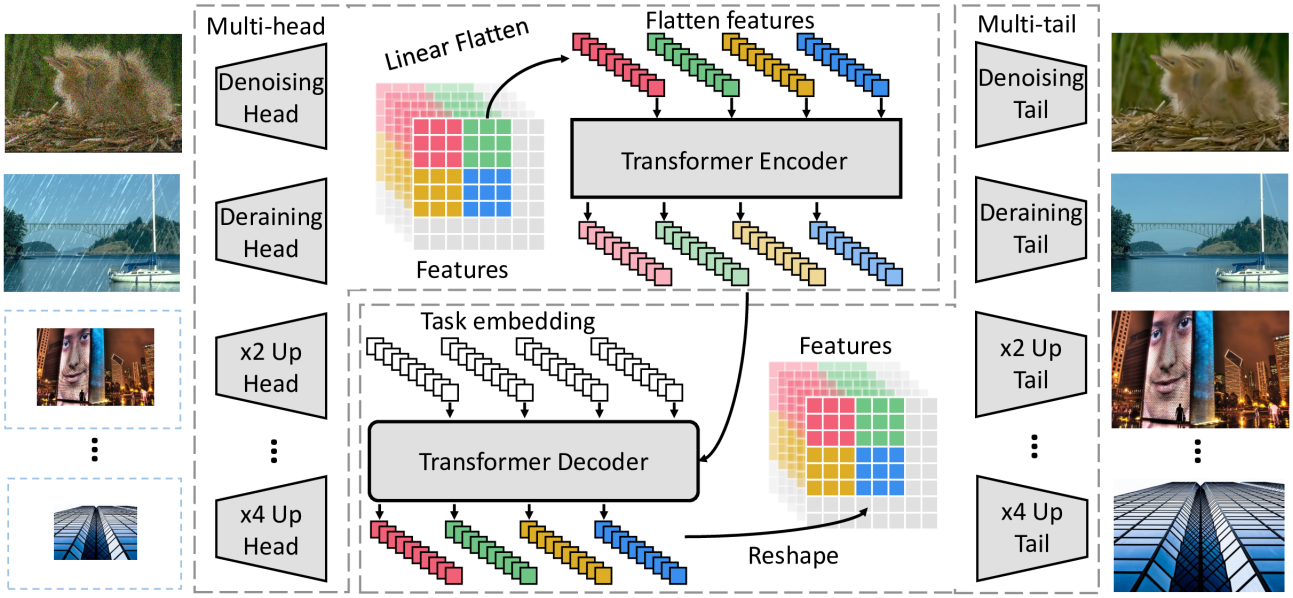
通过解决以下两个挑战，预训练有可能为图像处理任务提供一个有吸引力的解决方案。一、特定任务的数据是受限制的。这个问题在涉及付费数据或数据隐私，如医疗图像和卫星图像更为严峻。各种不一致的因素（如相机参数、光照、和天气）会进一步干扰用于训练的捕获数据的分布。二、在给出测试图像前，不知道要求的是何种图像处理任务。因此我们不得不准备好一系列的图像处理模块。它们有不同的目标，但是一些基础操作是可以共享的。

在计算机视觉和自然语言处理中使用预训练在现在是很常见的。例如，目标检测模型的主干部分通常在ImageNet分类上进行预训练。许多训练好的网络现在可以容易地在互联网上获取，包括 AlexNet、VGGNet 和 ResNet。Transformers 这个开创性的成果被广泛应用于许多自然语言处理任务，如翻译和问答。它成功的秘诀就是在大型语料库预训练基于 transformer 的模型，并在特定任务数据集上进行微调。Transformers 的变体，如 BERT 和 GPT-3，进一步扩充了训练数据并提高了预训练能力。这里有一些扩展 Transformers 的成功到计算机视觉领域的有趣尝试。例如，应用基于自注意力的模型来获取图像全局信息。Carion等人建议DERT使用 transformers 架构来进行 end-to-end 对象检测。最近，Dosovitskiy 等人引入视觉 Transformer 将输入图像视为 16*16 的 words 并取得了很好的图像识别效果。

前文提到的计算机视觉和自然语言的预训练主要研究先验分类任务，但是图像处理任务输入和输出都是图像。对这些现有的预训练策略的直接应用可能并不可行。而且，如何在预训练阶段有效解决不同目标图像处理任务仍是个难题。仍需要注意的是，图像处理模型预训练具有基于原始真实图像的自生成训练实例的便利。综合处理后的图像用于训练，是对原始图像本身的重构。

在这篇文章中，我们提出了一种使用 transformer 架构的图像处理的预训练模型，即 IPT。由于预训练模型需要兼容包括超分辨率、降噪、去雨等不同的图像处理任务，整个网络由多对匹配的不同任务的头尾和一个共享体构成。由于需要用大规模数据来挖掘 transformer 的潜力，我们需要准备大量差异化的图片来训练 IPT 模型。为此，我们选择了 ImageNet benchmark，包括 1000 种高分辨率。对于在 ImageNet 中的每一张图片，我们使用一些精心设计的操作生成许多损坏副本。例如，超分辨率任务的训练样本通过降采样原始图像来得到。我们用于训练 IPT 的完整数据集包括超过约 1000w 张图片。

然后，按下文步骤在大数据集上训练 transformer 架构。训练图像被输入到特定头，生成的特征被分割为块（即“词”）随后扁平化为序列。transformer 体用于处理扁平化特征，分别对 encoder 和 decoder 进行位置和任务嵌入。此外，为了很好适应不同的图像处理任务，引用了不同输入批的关系对比损失。本文提出的图像处理 transformer 通过端到端的方式学习。在数个 benchmarks 上的实验结果表明 IPT 预训练模型在微调后能胜过大多数现存的模型。



2. Related Works

2.1. Image Processing

图像处理包括超分辨率、降噪、去雾、去雨去模糊等对图像的操作。人们提出多种深度学习基础方法去处理一种或多种图像处理任务。对超分辨率任务来说，Dong 等人提出 SRCNN 算法被认为是引入端到端模型重建 HR（High Resolution，高分辨率）图片到 LR（Low Resolution，低分辨率）副本的开创性算法。Kim 等人通过更深层次的卷积网络进一步探索了深度神经网络的能力。Ahn 等人和Lim 等人提出引入残差块到 SR（超分辨率）任务中。Zhang 等人和Anwar 和Barnes 利用注意力机制提升了 SR 任务的性能。针对其他的任务也提出了很多优秀的算

法，如去噪、去雾、去雨和去模糊。不同以上方法，我们同时挖掘了大模型和大数据的性能。然后介绍了一种能处理多种图像任务的预训练模型。

2.2. Transformer

Transforme r及其变体在自然语言处理任务中强大的无监督学习和自监督学习预训练框架证明了它的成功。例如，GPTs 通过自回归方法预训练来在大文本数据集上预测下一个词。BERT 没有明确监督地从数据中学习，并依据上下文预测词。Colin 等人提出一种针对数个下游任务的通用预训练框架。Yinhan 等人提出一种原生 BERT 的稳固变体。

由于基于 Transformer 的模型的在 NLP 领域的成功，有很多人尝试探索 Transformer 在 CV 领域的优点。这些尝试可以粗略的分为两类。第一种是引入自注意力到传统的 CNN 网络。Yuan 等人引入用于图像分割的空间注意力。Fu 等人提出 DANET，结合空间和通道的注意力来利用上下文信息。Wang 等人、Chen 等人和Zhang 等人还通过自注意力增强特征来在几个高级视觉任务上加强模型性能。另一种是用自注意力块来替代 CNN。举个例子，Kolesnikov 等人和Dosovitskiy 等人利用 transformer 块进行图像分类。Carion 等人和Zhu 等人在检测中实现基于 transformer 的模型。Chen 等人提出了一种预训练的 GPT 模型用于生成和分类任务。Wu 等和Zhao 等提出了用于图像识别任务的基于 Transformer 的模型的预训练方法。然而，少有相关工作聚焦于底层视觉任务。在这篇文章里，我们探索了一种针对图像处理任务的通用预训练方法。

3. Image Processing Transformer

为了挖掘 transformer 在图像处理任务的使用潜力来实现一个更好的效果，我们在这里提出了一种在大规模数据集预训练的图像处理 transformer。

3.1. IPT architecture

我们的 IPT 的完整架构包括四个部分：从输入的损坏图像提取的特征头部（如，噪声图像和低分辨率图像），一对为了恢复输入数据中缺失信息的编码器-解码器 transformer，用于将特征恢复到重建的图像的 tails。这里我们简单的介绍一下 IPT 的架构，细节可以在补充材料中找到。

Heads. 为了调整不同的图像处理任务，我们使用了一个多头部架构来分别处理每个任务，每个头部由三个卷积层组成。表示输入图像为 $x \in \mathbb{R}^{3\times H\times W}$ （3表示R, G，和B），头部生成一个特征映射 $f_H \in \mathbb{R}^{C\times H\times W}$，具有 C 个通道，相同的高度和宽度（通常我们使用 C = 64）。计算公式为 $f_H = H^i(x)$，其中 $(i = \{1, \ldots, N_t\})$ 表示第 i 个任务的头，N_t 表示任务的数量。

Transformer encoder. 在将特征输入到 transformer 本体之前，我们将给定的特征分割成小块，每个小块被视为一个“词”。将 $f_H \in \mathbb{R}^{C\times H\times W}$ 的特征重塑为一组 patches，即 $f_{p_i} \in \mathbb{R}^{P^2\times C}, i = \{1, \ldots, N\}$，其中 $N = \frac{HW}{P^2}$ 为 patch 的个数(即序列长度)，P 为 patch 的大小。为了保持每个 patch 的位置信息，我们接下来向每个对于特征 f_{p_i} 批添加了可学习的位置编码 $E_{p_i} \in \mathbb{R}^{P^2\times C}$，以及 $E_{p_i} + f_{p_i}$ 将直接输入到 transformer 编码器中。编码器层的架构遵循原有的一个多头自注意力模块和一个前馈网络的结构。编码器对每个 patch 的输出 $f_{E_i} \in \mathbb{R}^{P^2\times C}$ 有相同大小的输入 f_{p_i}。计算可以用公式表示为：

$$
\begin{aligned}
&y_0 = \left[E_{p_1} + f_{p_1}, E_{p_2} + f_{p_2}, \ldots, E_{p_N} + f_{p_N}\right], \\
&q_i = k_i = v_i = \mathrm{LN}(y_{i-1}), \\
&y'_i = \mathrm{MSA}(q_i, k_i, v_i) + y_{i-1}, \\
&y_i = \mathrm{FFN}(\mathrm{LN}(y'_i)) + y'_i, \qquad\qquad i = 1, \ldots, l \\
&[f_{E_1}, f_{E_2}, \ldots, f_{E_N}] = y_l,
\end{aligned}
\tag{1}
$$

其中 l 为编码器的层数，MSA 为传统 transformer 模型中的多头自注意模块，FFN 为前馈网络，其中包含两个完全连通的层。

Transformer decoder. 解码器也采用相同的结构，并将解码器的输出作为 transformer 体的输入， transformer 体由两个多头自注意（MSA）层和一个前馈网络（FFN）组成。与原生 transformer 不同的是我们利用特定任务的嵌入作为解码器的额外输入。这些特定任务的嵌入 $E_t^i \in \mathbb{R}^{P^2 \times C}, i = \{1, \ldots, N_t\}$ 学习了从不同任务中解码特征。解码器的计算公式为：

$$
\begin{aligned}
&z_0 = [f_{E_1}, f_{E_2}, \ldots, f_{E_N}], \\
&q_i = k_i = \mathrm{LN}(z_{i-1}) + E_t, v_i = \mathrm{LN}(z_{i-1}), \\
&z_i' = \mathrm{MSA}(q_i, k_i, v_i) + z_{i-1}, \\
&q_i' = \mathrm{LN}(z_i') + E_t, k_i' = v_i' = \mathrm{LN}(z_0), \\
&z_i'' = \mathrm{MSA}(q_i', k_i', v_i') + z_i', \\
&z_i = \mathrm{FFN}(\mathrm{LN}(z_i'')) + z_i'', \qquad\qquad i = 1, \ldots, l \\
&[f_{D_1}, f_{D_2}, \ldots, f_{D_N}] = y_l,
\end{aligned}
\tag{2}
$$

其中 $f_{D_i} \in \mathbb{R}^{P^2 \times C}$ 是解码器的输出。被解码的 N 个大小为 $P^2 \times C$ 的特征被重塑为大小为 $C \times H \times W$的特征 f_D 。

Tails. 尾和头的性质相同，我们使用多个尾部来处理不同的任务。计算式为 $f_T = T^i(f_D)$，其中 T^i $(i = \{1, \ldots, N_t\})$ 表示第 i 个任务的头部， N_t 表示任务的数量。输出 f_T 是大小为 $3 \times H' \times W'$ 的由具体任务决定的输出图像。举个例子，对于一个两倍的超分辨率任务，$H' = 2H, W = 2W$。

3.2. Pre-training on ImageNet

除了 transformer 本身的结构之外，成功训练出一个优秀的 transformer 的关键因素之一是对大规模数据集的良好运用。与图像分类相比，用于图像处理任务的可用数据数量相对较少(例如，用于图像超分辨率任务的 DIV2K 数据集上只有 2000 张图像)，我们建议利用众所周知的 ImageNet 作为基线数据集，对我们的 IPT 模型进行预训练，然后我们为几个任务（如，超分辨率和去噪）生成完整的数据集。因为 ImageNet 集中的图像具有很大的的差异，包含了 1000 个类别的超过一百万张自然图片。这些图像具有丰富的纹理和颜色信息。我们首先去除语义标签，然后针对不同的任务使用不同的退化模型从这些未标记的图像中人工合成各种损坏的图像。需要注意的是，这些图像处理任务通常也使用合成数据集，我们使用的退行性变方法与 [27,1] 中建议的相同。例如，超分辨率任务通常采用双三次退化来生成低分辨率图像，去噪任务在不同噪声级别的干净图像中加入高斯噪声，生成有噪图像。这些合成的图像可以显著提高学习的深度网络的性能，包括 CNN 和 transformer 架构，这将在实验部分展示。基本上，将损坏的图像合成为:

$$
I_{corrupted} = \boldsymbol{f}(I_{clean}), \tag{3}
$$

其中 $\boldsymbol{f}$ 为退化变换函数，与特定任务有关。超分辨率任务中， $\boldsymbol{f}_{sr}$ 就是双三次插值；图像去噪任务中，$\boldsymbol{f}_{noise}(I) = I + \eta$ ，这里的 η 是加性高斯噪声；在去雨任务中， $\boldsymbol{f}_{rain}(I) = I + r$ ，其中 r 为人工雨条纹。我们的IPT以监督方式学习的损失函数可以表述为:

$$
\mathcal{L}_{supervised} = \sum_{i=1}^{N_t} L_1(\mathrm{IPT}(I_{corrupted}^i), I_{clean}), \tag{4}
$$

其中，L_1 为重建所需图像的常规 L1损失， $I_{corrupted}^i$ 为任务 i 的损坏图像。此外 式(4) 显示，我们提出的框架是由多个图像处理任务同时训练的。对于每个 batch，我们从 N_t 监督任务中随机选择一个任务进行训练，每个任务同时使用相应头部、尾部和任务嵌入进行处理。在对IPT模型预训练之后，会获取大量种类图片地固有属性和变化，可以进行进一步的微调来将新提供的数据应用于所需的任务。此外，还会去掉多余的头部和尾部来节约计算成本，并根据反向传播更新余下的头部、尾部和 transformer 体的参数。

然而，由于退化模型的多样性，我们无法对所有的图像处理任务合成图像。例如，在训练中会有可能存在大范围的噪音等级。因此，IPT 的泛化能力有待提高。和预训练的自然语言处理模型相似，图像 patches 之间的关系也提供了信息。图像版中的 patch 可以看作是NLP中的一个词。例如，从同一特征图中提取的 patches 更有可能同时出现，这些应该被嵌入到相似的位置。因此，我们引入了对比学习 [[11][],[29][]]，用于学习通用特征，是预训练 IPT 模型可以用于未知任务。在实践中，将 IPT 解码器为给定输入 s_j 生成的输出 patch 特征表示为 $f_{D_i}^j \in \mathbb{R}^{P^2 \times C}, i = \{1, \ldots, N\}$，其中 x_j 是从一批训练图像 $X = \{x_1, x_2, \ldots, x_B\}$ 中选择的。我们的目标是最小化相同图像的 patch 特征，最大化不同图像之间的 patch 特征。对比学习的损失函数公式化表示为

$$l(f^j_{D_{i_1}}, f^j_{D_{i_2}}) = -\log \frac{\exp(d(f^j_{D_{i_1}}, f^j_{D_{i_2}}))}{\sum_{k=1}^{B} \mathbb{I}_{k \neq j} \exp(d(f^j_{D_{i_1}}, f^k_{D_{i_2}}))},$$
$$\mathcal{L}_{constrastive} = \frac{1}{BN^2} \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \sum_{j=1}^{B} l(f^j_{D_{i_1}}, f^j_{D_{i_2}}), \tag{5}$$

其中 $d(a,b) = \frac{a^T b}{\|a\|\|b\|}$ 为余弦相似度。此外，为了充分利用监督信息和自我监督信息，我们将损失函数重新表述为:

$$\mathcal{L}_{IPT} = \lambda \cdot \mathcal{L}_{contrastive} + \mathcal{L}_{supervised}. \tag{6}$$

其中，我们将λ平衡的对比损失和监督损失结合起来作为IPT的最终目标函数。因此，利用 式(6) 训练的变压器网络可以有效地用于现有的各种图像处理任务。

4. Experiments

在本节中，我们评估提出的IPT在各种图像处理任务上的性能，包括超分辨率和图像去噪。我们证明预先训练的IPT模型可以在这些任务上达到最先进的性能。此外，大量的剥离实验表明，在使用大规模数据集解决图像处理问题时，基于 transformer 的模型比卷积神经网络表现更好。

Datasets. 为了获得更好的 IPT 模型的预训练结果，我们使用了知名的 ImageNet 数据集，该数据集包含超过一百万的高多样性彩色图像。训练图像被分割成 48 × 48 个 patches，有 3 个通道进行训练，即 IPT 模型有超过 10M 的 patches 进行训练。然后分别采用 2×、3×、4× 双三次插值、30、50 噪声级高斯噪声和添加雨水等 6 种退化方式生成损坏图像。我们通过下面的方法生成连续添加雨水的图像。在测试过程中，我们将测试集中的图像裁剪为 48 × 48 块，重叠 10 个像素。需要注意的是，对于基于 CNN 的模型，我们也采用了相同的测试策略，以便进行公平的比较，得到的 CNN 模型的 PSNR 值与它们的基线相同。

Training & Fine-tuning. 我们在 32 张英伟达 Tesla V100 上使用传统的 Adam 优化器（$\beta_1 = 0.9, \beta_2 = 0.999$），在修改后的 ImageNet 数据集上训练我们的 IPT 模型 300 个 epoch。初始学习率设为 $5e^{-5}$ ，在 256 个 batch 大小的情况下 200 epoch 后衰减到 $2e^{-5}$。训练集由不同的任务组成，由于昂贵的内存成本，我们不能在单个批处理中输入所有的任务。因此，我们在每次迭代时从随机选择的任务中提取一批图像进行处理。在整个合成数据集上进行预训练后，我们根据期望的任务（例如，×3 单图像超分辨率）对 IPT 模型进行学习速率为 $2e^5$ 的 30 个 epochs 的微调。注意，SRCNN 也发现使用 ImageNet 训练可以提高超分辨率任务的性能，而我们提出了一个适合一般底层视觉任务的模型。

4.1. Super-resolution

我们将我们的模型与几种最先进的基于 CNN 的 SR 方法进行比较。如 表1 所示，我们的预训练 IPT 优于所有其他方法，在 ×2，×3，×4 比例情况下在所有数据集上都取得了最好的性能。值得强调的是，我们的模型在×2 Urban100数据集达到 33.76 dB PSNR，超过其他方法有超过约 0.4dB，而以前的 SOTA 方法与其他方式相比只能实现一个小于 0.2 dB 的提升，这表明该模型进行大规模训练的优越性。

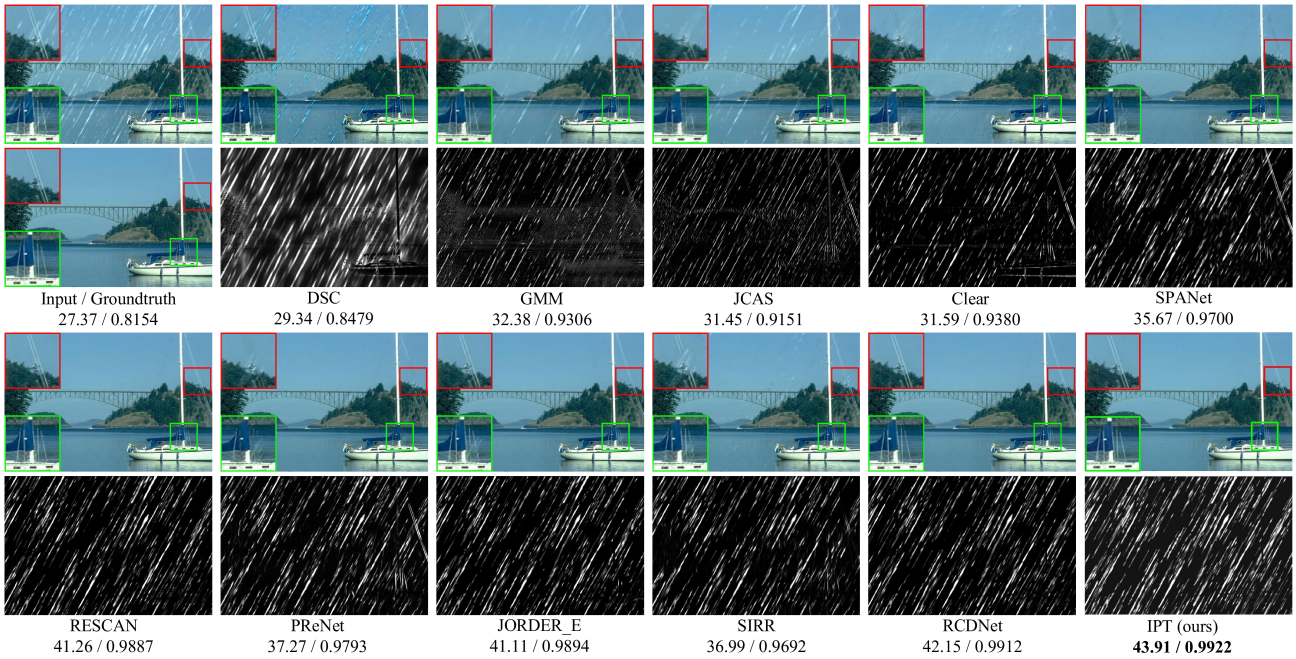
我们进一步展示了我们的模型在 4× 比例、Urban100 数据集上的可视化结果。如 图3 所示，高比例系数会导致大量信息丢失，原有的高分辨率图像很难恢复。以往的方法生成的图像是模糊的，而我们的模型生成的超分辨率图像可以很好地从低分辨率图像中恢复细节。

4.2. Denoising

因为我们预先训练的模型可以很好地适应许多任务，所以我们在去噪任务上评估我们的模型的表现。训练和测试数据是通过在干净图像上添加 σ = 30,50 的高斯噪声生成的。

为了验证所提方法的有效性，我们将我们的结果与各种最先进的模型进行比较。表2 展示了在 BSD68 和 Urban100 数据集上的彩色图像去噪结果。结果表明，不同高斯噪声等级下，我们的 IPT 在所有的去噪方法中的去噪效果最好。此外，我们惊奇地发现我们的模型在 Urban100 数据集上提高了约2dB的最高性能，这证明了预训练的有效性和基于transformer的模型的优越性。

图4 显示了结果可视化图像。如图所示，有噪声的图像很难被识别，很难恢复干净的图像。因此，现有的方法无法重现足够的细节，产生异常像素。结果显示，我们的预训练模型可以很好地恢复这只猫毛发中的几个细节，我们的视觉质量明显优于之前的所有模型。



4.3. Deraining

对于图像去噪任务，我们在合成的 Rain100L 数据集上评估我们的模型，该数据集由 100 张雨图像组成。定量结果见 表3。与最先进的方法相比，我们能达到最高 41.62dB 的最高性能，提高了 1.62dB。

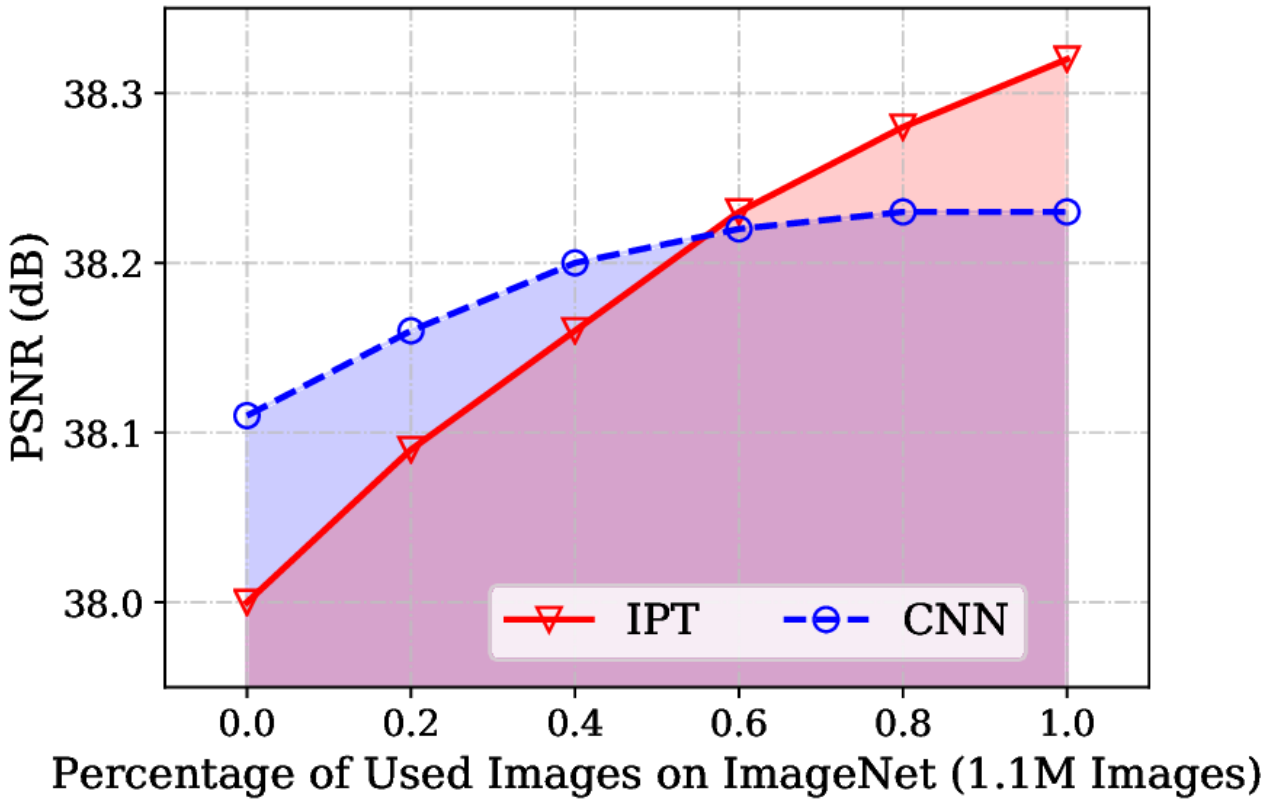
图5显示了可视化结果。由于缺乏图像的先验性，以往的方法无法重构出原始的干净图像。因此，我们的 IPT 模型可以呈现与 ground truth 完全相同的图像，并且在视觉质量上超过了以往所有算法。这一结果证明了所提出的模型的通用性。

4.4. Generalization Ability

虽然我们可以生成各种损坏的图像，但是自然图像的复杂度很高，我们不能合成所有可能的图像来对transformer模型进行预训练。然而，一个良好的预训练模型应该能够很好地适应自然语言处理领域的其他任务。为此，我们进行了几个实验来验证我们模型的泛化能力。在实践中，我们测试了我们合成的ImageNet数据集中没有包含的损坏图像，即分别对噪声级别为10和70的图像进行去噪。我们将图像去噪任务的头部和尾部作为预训练模型。

详细的结果在 表4 中展示。我们比较了使用预训练的 IPT 模型和最先进的图像去噪方法的性能。显然，IPT 模型优于其他的传统方法，这表明预训练的模型能够从大规模数据集中捕捉到更多有用的信息和特征。

4.5. Ablation Study

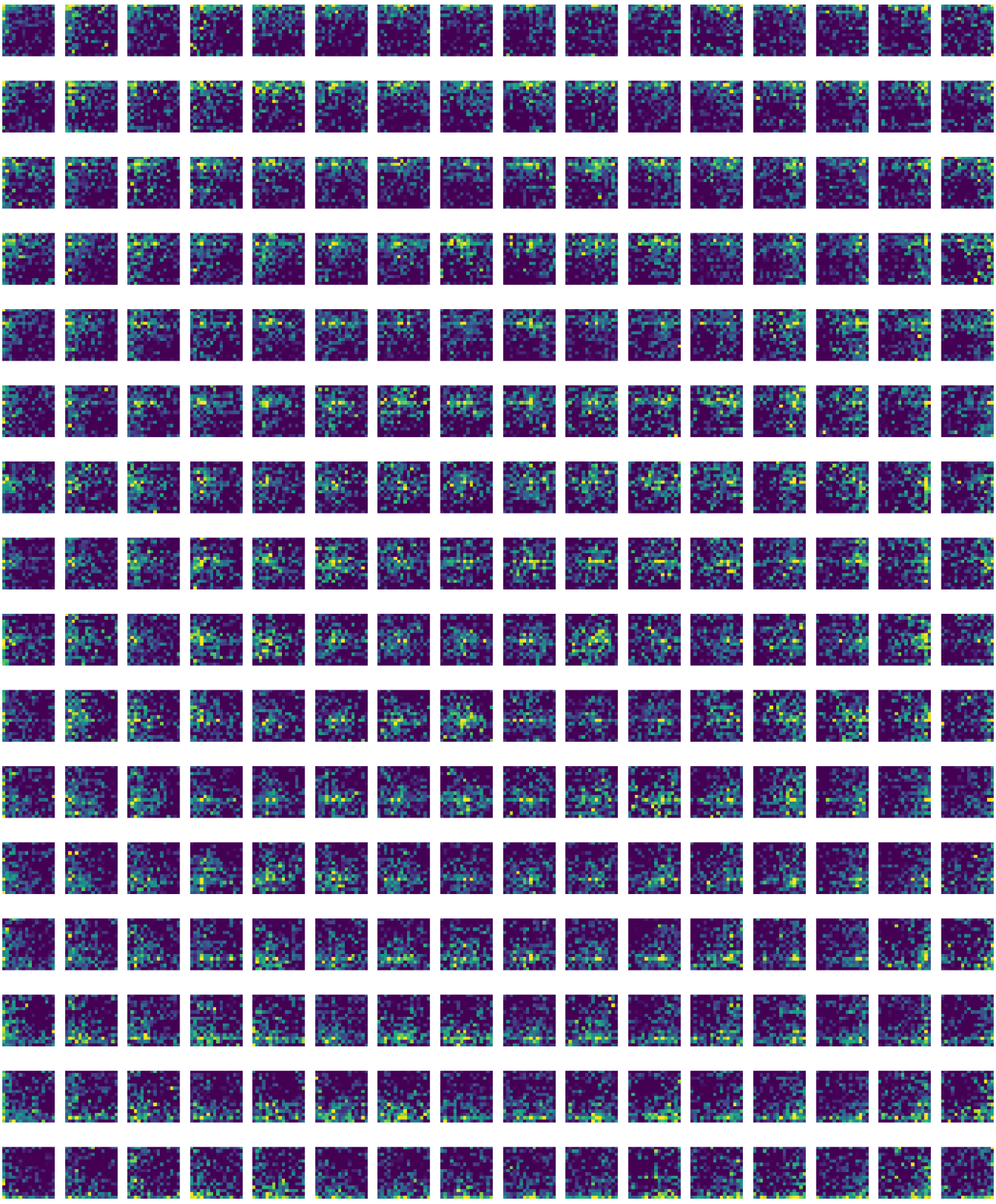


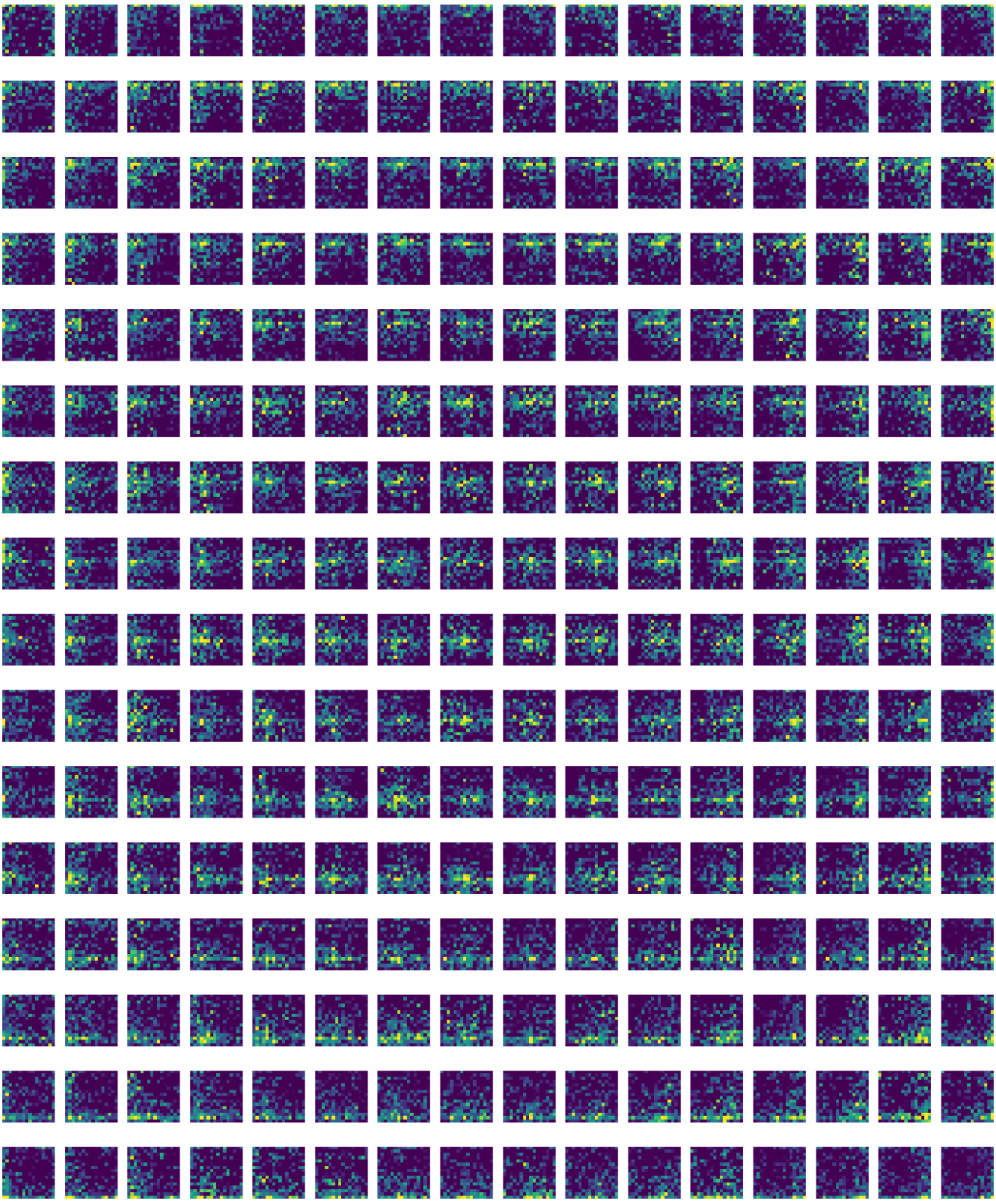
Impact of data percentage. 为了评估 transformer 架构的有效性，我们进行了实验，分析了基于 CNN 的模型和基于 transformer 模型的预训练的改进。我们利用已知的 EDSR 模型作为 CNN 基线，并在合成的 ImageNet 数据集上对其和IPT模型进行了预训练。我们使用合成ImageNet数据集的 20%，40%，60%，80%和100% 来分析使用的数据数量对结果性能的影响。图6 显示了不同预训练模型的结果。当没有对模型进行预训练或对整个数据集进行少量（< 60%）的预训练时，CNN模型相较于 IPT 达到了更好的性能。相比之下，当使用大规模数据时，基于 transformer 的模型压倒性胜过了 CNN 模型，这证明了我们预训练的 IPT 模型的有效性。

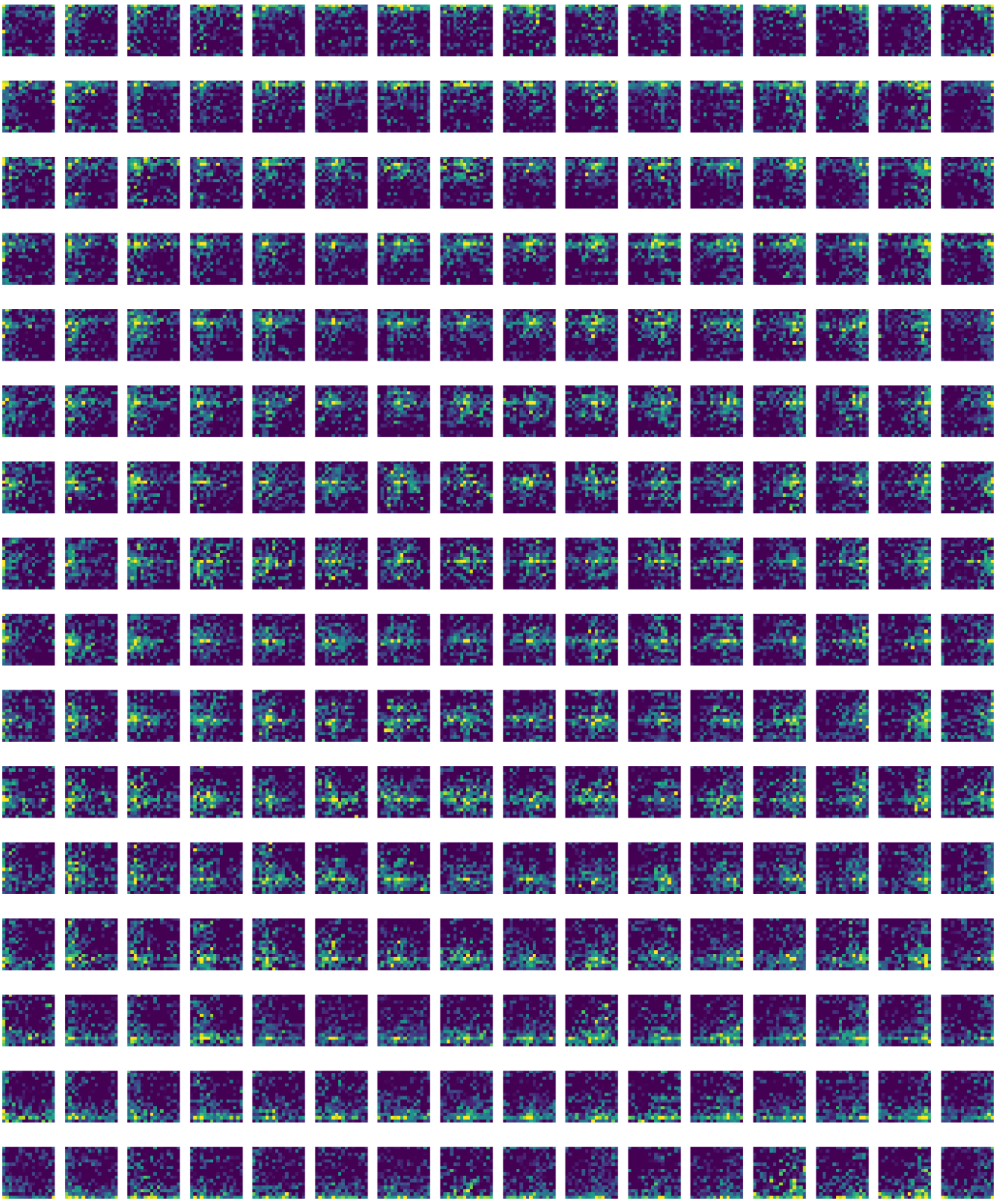
Impact of contrastive learning. 如上所述，为了提高我们预训练模型的表示能力，我们将对比学习损失（式(6)）加入到训练过程中。然后，我们使用 Set4 数据集来评估其在 ×2 scale 超分辨率任务中的效果。表5 显示了超参数 λ 对平衡 式(6) 中的两项的影响。当 λ=0 时，仅使用监督学习方法训练 IPT 模型，得到的 PSNR 值为 38.27dB。采用对比损失进行自监督学习时，模型的 PSNR 值可以达到 38.37dB （λ = 0.1），比用 λ = 0 训练的模型高出约 0.1dB。这些结果进一步证明了对比学习对于学习更好的预训练 IPT 模型的有效性。

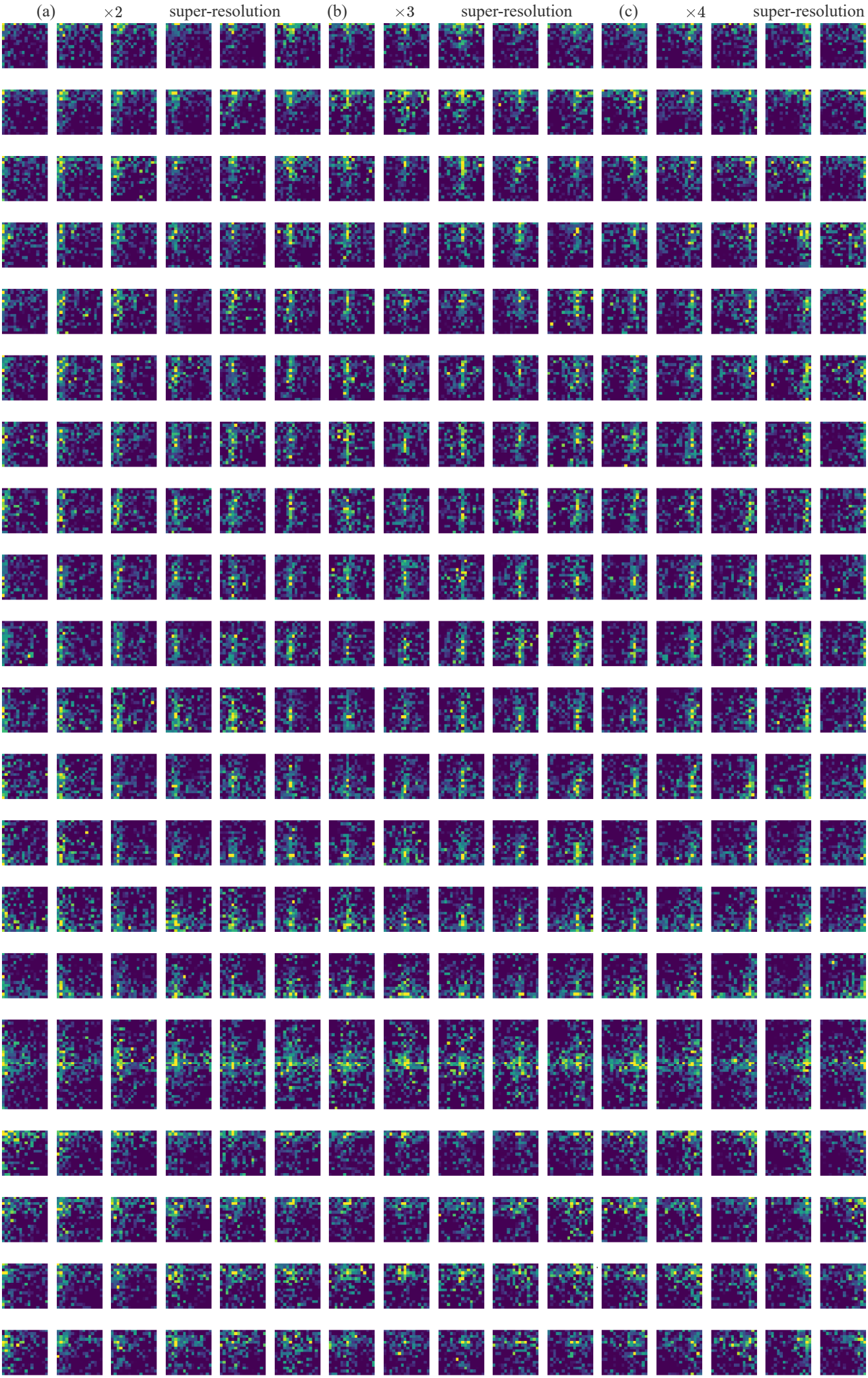
5. Conclusions and Discussions

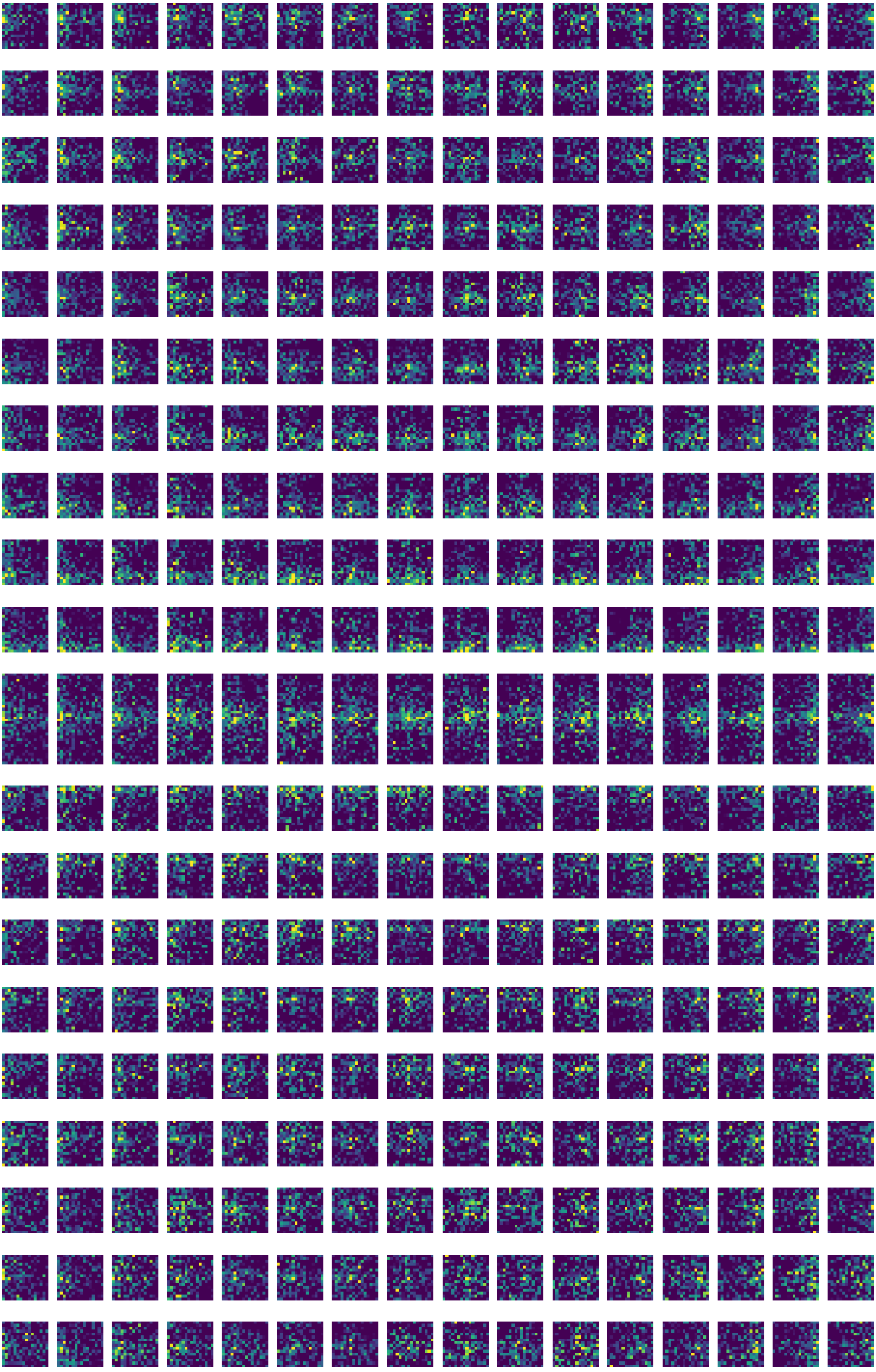
本文旨在利用预先训练的 transformer 模型（IPT）来解决图像处理问题。为满足图像超分辨率和去噪等不同的图像处理任务，设计了多头、多尾的共享 transformer 体。为了最大限度地挖掘 transformer 架构在各种任务上的性能，我们探求了一个合成的 ImageNet 数据集。在其中将每幅原始图像退化为一系列对应的图像，作为配对的训练数据。然后，使用监督和自我监督的方法训练 IPT 模型，模型显示了强大的捕捉内在特征，以进行低级图像处理的能力。实验结果表明，仅使用一个经过快速微调后的预训练的模型，我们的 IPT 可以超过目前最先进的方法。在未来的工作中，我们将扩展我们的 IPT 模型到更多的任务，如去模糊，去雾等。

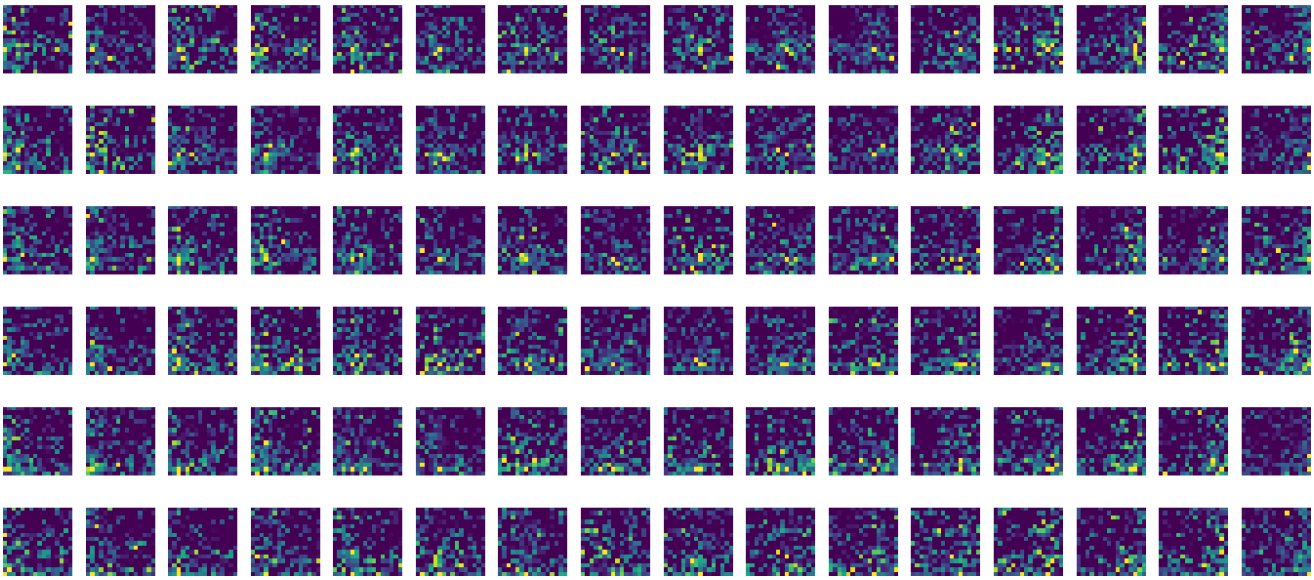




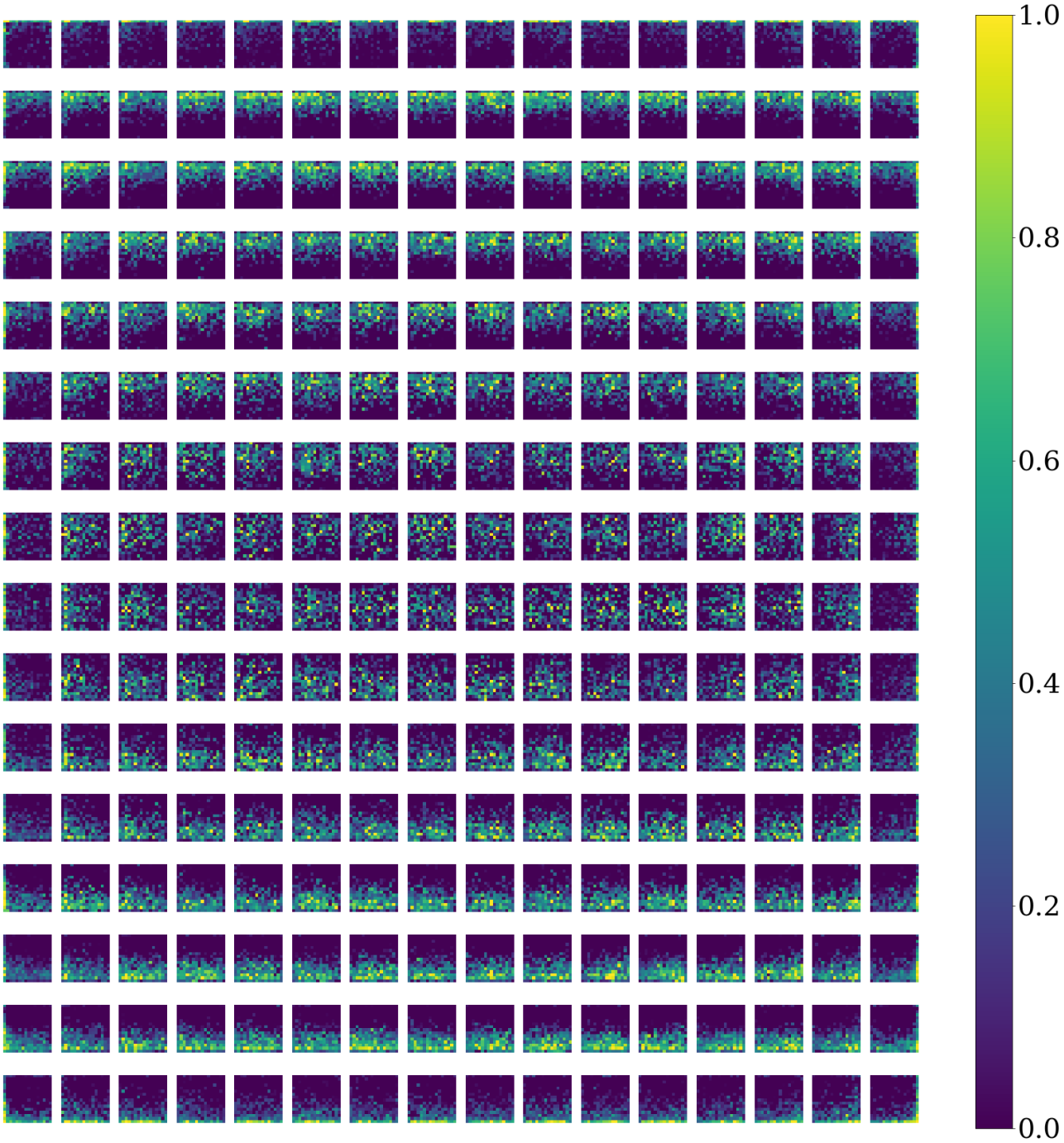



(a) ×2 super-resolution
(b) ×3 super-resolution
(c) ×4 super-resolution





(d) deraining (e) denoising with 30 noise level (f) denoising with 50 noise level


1.0
0.8
0.6
0.4
0.2
0.0