深度特征作为感知度量的惊人有效性

The Unreasonable Effectiveness of Deep Features as a Perceptual Metric

Richard Zhang[1] Phillip Isola[12] Alexei A. Efros[1] Eli Shechtman[3] Oliver Wang[3]
[1]UC Berkeley [2]OpenAI [3]Adobe Research
{rich.zhang, isola, efros}@eecs.berkeley.edu {elishe,owang}@adobe.com

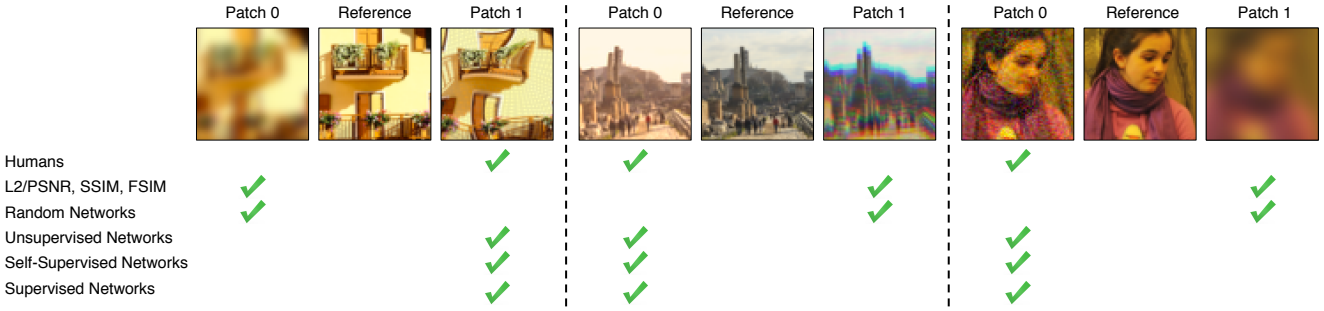


图 1: **在这些例子里，哪个图块（左边还是右边）与中间的图块“更接近”?** 每一组中，传统度量（L2/PSNR、SSIM、FSIM）都与人类判断相左。然而深度网络——即便横跨不同架构（SqueezeNet [20]、AlexNet [27]、VGG [52]）与不同监督类型（监督 [47]、自监督 [13, 40, 43, 64]，乃至无监督 [26]）——却都能给出一种涌现的嵌入，与人类判断出奇地吻合。我们进一步在一个大规模感知判断数据库上对已有的深度嵌入加以校准；模型与数据参见 `https://www.github.com/richzhang/PerceptualSimilarity`。

摘要

While it is nearly effortless for humans to quickly assess the perceptual similarity between two images, the underlying processes are thought to be quite complex. Despite this, the most widely used perceptual metrics today, such as PSNR and SSIM, are simple, shallow functions, and fail to account for many nuances of human perception. Recently, the deep learning community has found that features of the VGG network trained on ImageNet classification has been remarkably useful as a training loss for image synthesis. But how perceptual are these so-called "perceptual losses"? What elements are critical for their success? To answer these questions, we introduce a new dataset of human perceptual similarity judgments. We systematically evaluate deep features across different architectures and tasks and compare them with classic metrics. We find that deep features outperform all previous metrics by large margins on our dataset. More surprisingly, this result is not restricted to ImageNet-trained VGG features, but holds across different deep architectures and levels of supervision (supervised, self-supervised, or even unsupervised). Our results suggest that perceptual similarity is an emergent property shared across deep visual representations.

对人类而言，迅速判断两幅图像之间的感知相似性几乎毫不费力，但其背后的加工过程据信相当复杂。尽管如此，当今最广为使用的感知度量——例如 *PSNR* 与 *SSIM*——却都是简单、浅层的函数，无法刻画人类感知的诸多微妙之处。近来，深度学习界发现，在 *ImageNet* 分类任务上训练的 *VGG* 网络，其特征作为图像合成的训练损失异常有用。但这些所谓的“感知损失”究竟有多“感知”? 它们成功的关键要素又是什么? 为回答这些问题，我们构建了一个新的人类感知相似性判断数据集。我们系统地评测了不同架构、不

本文为 AI 辅助翻译版本，仅供学习交流；引用请以英文原文为准：arXiv:1801.03924。

同任务下的深度特征，并将其与经典度量作比较。我们发现，在本数据集上，深度特征以很大优势超越了以往所有度量。更出人意料的是，这一结果并不限于 *ImageNet* 训练的 *VGG* 特征，而是在不同的深度架构与不同监督程度（监督、自监督乃至无监督）下都成立。这些结果表明，感知相似性是各类深度视觉表示所共有的一种涌现属性。

1. 动机 Motivation

The ability to compare data items is perhaps the most fundamental operation underlying all of computing. In many areas of computer science it does not pose much difficulty: one can use Hamming distance to compare binary patterns, edit distance to compare text files, Euclidean distance to compare vectors, etc. The unique challenge of computer vision is that even this seemingly simple task of comparing visual patterns remains a wide-open problem. Not only are visual patterns very high-dimensional and highly correlated, but, the very notion of visual similarity is often subjective, aiming to mimic human visual perception. For instance, in image compression, the goal is for the compressed image to be indistinguishable from the original by a human observer, irrespective of the fact that their pixel representations might be very different.

比较数据项的能力，或许是支撑一切计算的最基本操作。在计算机科学的许多领域，这并不构成太大困难：可以用汉明距离比较二进制模式，用编辑距离比较文本文件，用欧氏距离比较向量，等等。计算机视觉的独特挑战在于，即便是比较视觉模式这样看似简单的任务，至今仍是一个悬而未决的问题。视觉模式不仅维度极高、彼此高度相关，而且视觉相似性这一概念本身往往是主观的，其目标是模拟人类的视觉感知。例如在图像压缩中，目标是让压缩后的图像在人类观察者看来与原图无从区分，而不论二者的像素表示可能相差甚大。

Classic per-pixel measures, such as ℓ_2 Euclidean distance, commonly used for regression problems, or the related Peak Signal-to-Noise Ratio (PSNR), are insufficient for assessing structured outputs such as images, as they assume pixel-wise independence. A well-known example is that blurring causes large perceptual but small ℓ_2 change.

经典的逐像素度量——如回归问题中常用的 ℓ_2 欧氏距离，或与之相关的峰值信噪比（PSNR）——不足以评价图像这类结构化输出，因为它们假设各像素相互独立。一个众所周知的例子是：模糊会带来很大的感知变化，ℓ_2 上的变化却很小。

What we would really like is a "perceptual distance," which measures how similar are two images in a way that coincides with human judgment. This problem has been a longstanding goal, and there have been numerous perceptually motivated distance metrics proposed, such as SSIM [58], MSSIM [60], FSIM [62], and HDR-VDP [34].

我们真正想要的是一种 "感知距离"，它以契合人类判断的方式来衡量两幅图像有多相似。这一问题是长期以来的目标，人们已提出了众多受感知启发的距离度量，如 SSIM [58]、MSSIM [60]、FSIM [62] 与 HDR-VDP [34]。

However, constructing a perceptual metric is challenging, because human judgments of similarity (1) depend on high-order image structure [58], (2) are context-dependent [19, 36, 35], and (3) may not actually constitute a distance metric [56]. The crux of (2) is that there are many different "senses of similarity" that we can simultaneously hold in mind: is a red circle more similar to a red square or to a blue circle? Directly fitting a function to human judgments may be intractable due the the context-dependent and pairwise nature of the judgments (which compare the similarity between *two* images). Indeed, we show in this paper a negative result where this approach fails to generalize, even when trained on a large-scale dataset containing many distortion types.

然而，构建感知度量并不容易，因为人类的相似性判断 (1) 依赖高阶图像结构 [58]，(2) 依赖上下文 [19, 36, 35]，且 (3) 未必真的构成一个距离度量 [56]。第 (2) 点的关键在于，我们心中可以同时持有许多不同的 "相似之感"：一个红色圆更像一个红色方，还是更像一个蓝色圆？由于判断既依赖上下文又是成对的（比较两幅图像之间的相似性），直接对人类判断拟合一个函数可能难以处理。事实上，本文给出了一个反面结

(a) **传统** (b) **基于 CNN**

图 2: **失真示例**。展示了分别用我们的 (a) 传统方法与 (b) 基于 CNN 的方法所生成的失真示例。

果：即便在包含多种失真类型的大规模数据集上训练，这种做法仍无法泛化。

Instead, might there be a way to learn a notion of perceptual similarity without directly training for it? The computer vision community has discovered that internal activations of deep convolutional networks, though trained on a high-level image classification task, are often surprisingly useful as a representational space for a much wider variety of tasks. For example, features from the VGG architecture [52] have been used on tasks such as neural style transfer [17], image superresolution [23], and conditional image synthesis [14, 8]. These methods measure distance in VGG feature space as a "perceptual loss" for image regression problems [23, 14].

换个思路：是否存在一种无需直接为之训练、却能学到感知相似性概念的途径？计算机视觉界已经发现，深度卷积网络的内部激活尽管是在高层图像分类任务上训练得到的，却常常出人意料地可用作范围广得多的各类任务的表示空间。例如，VGG 架构 [52]的特征已被用于神经风格迁移 [17]、图像超分辨率 [23] 以及条件图像合成 [14, 8] 等任务。这些方法把 VGG 特征空间中的距离用作图像回归问题的 "感知损失" [23, 14]。

But how well do these "perceptual losses" actually correspond to human visual perception? How do they compare to traditional perceptual image evaluation metrics? Does the network architecture matter? Does it have to be trained on the ImageNet classification task, or would other tasks work just as well? Do the networks need to be trained at all?

但这些 "感知损失" 究竟在多大程度上契合人类的视觉感知？与传统的感知图像评价度量相比又如何？网络架构是否重要？是否非得在 ImageNet 分类任务上训练，抑或换成别的任务同样奏效？网络到底需不需要经过训练？

In this paper, we evaluate these questions on a new large-scale database of human judgments, and arrive at several surprising conclusions. We find that internal activations of networks trained for high-level classification tasks, even across network architectures [20, 28, 52] and no further calibration, do indeed correspond to human perceptual judgments. In fact, they correspond far better than the commonly used metrics like SSIM and FSIM [58, 62], which were not designed to handle situations where spatial ambiguities are a factor [49]. Furthermore, the best performing self-supervised networks, including BiGANs [13], cross-channel prediction [64], and puzzle solving [40] perform just as well at this task, even without the benefit of human-labeled training data. Even a simple unsupervised network initialization with stacked k-means [26] beats the classic metrics by a large margin! This illustrates an *emergent property* shared across networks, even across architectures and training signals. Importantly, however, having *some* training signal appears crucial – a randomly initialized network achieves much lower performance.

本文在一个新的大规模人类判断数据库上考察这些问题，并得出若干令人惊讶的结论。我们发现，为高层分类任务训练的网络，其内部激活即便跨越不同网络架构 [20, 28, 52]、且不作任何进一步校准，也确实与人类的感知判断相符。事实上，它们的吻合程度远胜

数据集	输入图像/图块数	输入类型	失真数目	失真类型数	强度级数	失真图像/图块数	判断数目	判断类型
LIVE [51]	29	图像	5	传统	连续	.8k	25k	MOS
CSIQ [29]	30	图像	6	传统	5	.8k	25k	MOS
TID2008 [46]	25	图像	17	传统	4	2.7k	250k	MOS
TID2013 [45]	25	图像	24	传统	5	3.0k	500k	MOS
BAPPS (2AFC–Distort)	160.8k	64×64 图块	425	传统 +CNN	连续	321.6k	349.8k	2AFC
BAPPS (2AFC–Real alg)	26.9k	64×64 图块	–	算法输出	–	53.8k	134.5k	2AFC
BAPPS (JND–Distort)	9.6k	64×64 图块	425	传统 +CNN	连续	9.6k	28.8k	相同/不同

表 1: **数据集对比**。我们提出的 Berkeley-Adobe 感知图块相似性（BAPPS）数据集与以往工作的一个主要区别，在于失真类型的规模。我们用取自 [7, 10] 的未压缩图像，在失真集合上提供人类感知判断。以往数据集使用少量失真、且强度取离散级别；我们则使用大量失真（由若干原子失真按顺序组合而成）并连续采样。对每个输入图块，我们用两种失真对其进行破坏，并就每对征集少量人类判断（训练集 2 条、测试集 5 条），这使我们得以在大量图块上获取判断。以往数据库把判断汇总为一个平均意见得分（MOS）；我们则直接报告成对判断（二选一强迫选择）。此外，我们还提供对真实算法输出的判断，以及一个相同/不同的恰可察觉差异（JND）感知测试。

于 SSIM、FSIM [58, 62] 等常用度量——后者本就不是为应对空间歧义成为主要因素的情形而设计的 [49]。此外，表现最好的自监督网络——包括 BiGAN [13]、跨通道预测 [64] 与拼图求解 [40]——在这项任务上同样出色，哪怕没有人工标注的训练数据可用。就连用堆叠 k-means [26] 做简单无监督初始化的网络，也以很大优势击败了经典度量！这体现出一种在各网络之间、乃至跨架构与跨训练信号所共有的涌现属性。但重要的是，具备某种训练信号似乎不可或缺——随机初始化的网络性能要低得多。

Our study is based on a newly collected perceptual similarity dataset, using a large set of distortions and real algorithm outputs. It contains both traditional distortions, such as contrast and saturation adjustments, noise patterns, filtering, and spatial warping operations, and CNN-based algorithm outputs, such as autoencoding, denoising, and colorization, produced by a variety of architectures and losses. Our dataset is richer and more varied than previous datasets of this kind [45]. We also collect judgments on outputs from real algorithms for the tasks of superresolution, frame interpolation, and image deblurring, which is especially important as these are the real-world use cases for a perceptual metric. We show that our data can be used to "calibrate" existing networks, by learning a simple linear scaling of layer activations, to better match low-level human judgments.

我们的研究基于一个新采集的感知相似性数据集，使用了大量失真以及真实算法的输出。它既包含传统失真，如对比度与饱和度调整、噪声模式、滤波以及空间形变操作，也包含由多种架构与损失产生的、基于 CNN 的算法输出，如自编码、去噪与着色。相比以往的同类数据集 [45]，我们的数据集更丰富、更多样。我们还针对超分辨率、帧插值与图像去模糊等任务，采集了对真实算法输出的判断——这一点尤为重要，因为它们正是感知度量在现实中的用武之地。我们表明，通过学习对各层激活作一个简单的线性缩放，我们的数据可用来 "校准" 已有网络，使之更贴合低层的人类判断。

Our results are consistent with the hypothesis that perceptual similarity is not a special function all of its own, but rather a *consequence* of visual representations tuned to be predictive about important structure in the world. Representations that are effective at semantic prediction tasks are also representations in which Euclidean distance is highly predictive of perceptual similarity judgments.

我们的结果与如下假设相符：感知相似性并非一项自成一体的特殊功能，而是视觉表示被调校得能对世界中的重要结构作出预测之后的一种产物。凡是在语义预测任务上有效的表示，也正是欧氏距离能高度预测感知相似性判断的表示。

Our contributions are as follows:

我们的贡献如下：

- We introduce a large-scale, highly varied, perceptual similarity dataset, containing 484k human judgments. Our dataset not only includes parameterized distortions, but also real algorithm outputs. We also collect judgments on a different perceptual test, just noticeable differences (JND).

  我们提出了一个大规模、高度多样的感知相似性数据集，包含 48.4 万条人类判断。该数据集不仅含有参数化失真，还含有真实算法的输出。我们还就另一种感知测试——恰可察觉差异（JND）——采集了判断。

- We show that deep features, trained on supervised, self-supervised, and unsupervised objectives alike, model low-level perceptual similarity surprisingly well, outperforming previous, widely-used metrics.

  我们表明，无论以监督、自监督还是无监督目标训练，深度特征都能出人意料地很好地刻画低层感知相似性，超越以往广泛使用的度量。

- We demonstrate that network architecture alone does not account for the performance: untrained nets achieve much lower performance.

  我们证明，仅凭网络架构并不足以解释这一性能：未经训练的网络性能要低得多。

- With our data, we can improve performance by "calibrating" feature responses from a pre-trained network.

  借助我们的数据，可以通过 "校准" 预训练网络的特征响应来提升性能。

数据集方面的已有工作 Prior work on datasets In order to evaluate existing similarity measures, a number of datasets have been proposed. Some of the most popular are the LIVE [51], TID2008 [46], CSIQ [29], and TID2013 [45] datasets. These datasets are referred to Full-Reference Image Quality Assessment (FR-IQA) datasets and have served as the de-facto baselines for development and evaluation of similarity metrics. A related line of work is on No-Reference Image Quality Assessment (NR-IQA), such as AVA [38] and LIVE In the Wild [18]. These datasets investigate the "quality" of individual images by themselves, without a reference image. We collect a new dataset that is complementary to these: it contains a substantially larger number of distortions, including some from newer, deep network based outputs, as well as geometric distortions. Our dataset is focused on perceptual similarity, rather than quality assessment. Additionally, it is collected on patches as opposed to full images, in the wild, with a different experimental design (more details in Sec 2).

为评价已有的相似性度量，人们提出了若干数据集，其中最流行的包括 LIVE [51]、TID2008 [46]、CSIQ [29] 与 TID2013 [45]。这些被称为全参考图像质量评价（FR-IQA）数据集，事实上已成为相似性度量研发与评价的基准。与之相关的另一条研究线是无参考图像质量评价（NR-IQA），如 AVA [38] 与 LIVE In the Wild [18]。这类数据集在没有参考图像的情况下，单独考察每幅图像自身的 "质量"。我们采集的新数据集与它们互补：它包含的失真数量大得多，其中既有来自较新的、基于深度网络的输出，也有几何失真。我们的数据集聚焦于感知相似性，而非质量评价；此外，它采集的是图块而非整幅图像，在自然环境下进行，并采用了不同的实验设计（详见第 2 节）。

深度网络与人类判断方面的已有工作 Prior work on deep networks and human judgments Recently, advances in DNNs have motivated investigation of applications in the context of visual similarity and image quality assessment. Kim and Lee [25] use a CNN to predict visual similarity by training on low-level differences. Concurrent work by Talebi and Milanfar [54, 55] train a deep network in the context of NR-IQA for image aesthetics. Gao et al. [16] and Amirshahi et al. [3] propose techniques involving leveraging internal activations of deep networks (VGG and AlexNet, respectively) along with additional multiscale post-processing. In this work, we conduct a more in-depth study across different architectures, training signals, on a new, large scale, highly-varied dataset.

近来，深度神经网络（DNN）的进展推动了在视觉相似性与图像质量评价场景下的应用研究。Kim 与 Lee [25] 用一个 CNN，通过在低层差异上训练来预测视觉相似性。Talebi 与 Milanfar [54, 55] 的同期工作，则在 NR-IQA 的背景下训练深度网络以评估图像美学。

Gao 等 [16] 与 Amirshahi 等 [3] 提出的方法，分别借助深度网络（VGG 与 AlexNet）的内部激活，再辅以额外的多尺度后处理。在本工作中，我们在一个全新的、大规模且高度多样的数据集上，跨越不同架构与不同训练信号开展了更深入的研究。

Recently, Berardino et al. [6] train networks on perceptual similarity, and importantly, assess the ability of deep networks to make predictions on a *separate* task – predicting most and least perceptually-noticeable directions of distortion. Similarly, we not only assess image patch similarity on parameterized distortions, but also test generalization to real algorithms, as well as generalization to a separate perceptual task – just noticeable differences.

近来，Berardino 等 [6] 在感知相似性上训练网络，并且重要的是，评估了深度网络在一项另设任务上作预测的能力——预测失真中感知上最显著与最不显著的方向。类似地，我们不仅在参数化失真上评估图像块相似性，还测试了对真实算法的泛化，以及对另一项感知任务——恰可察觉差异——的泛化。

2. Berkeley-Adobe 感知图块相似性（BAPPS）数据集 Berkeley-Adobe Perceptual Patch Similarity (BAPPS) Dataset

To evaluate the performance of different perceptual metrics, we collect a large-scale highly diverse dataset of perceptual judgments using two approaches. Our main data collection employs a two alternative forced choice (2AFC) test, that asks which of two distortions is more similar to a reference. This is validated by a second experiment where we perform a just noticeable difference (JND) test, which asks whether two patches – one reference and one distorted – are the same or different. These judgments are collected over a wide space of distortions and real algorithm outputs.

为评价不同感知度量的性能，我们用两种方式采集了一个大规模、高度多样的感知判断数据集。主体数据采用二选一强迫选择（2AFC）测试，询问两个失真中哪一个与参考更相似。我们再用第二个实验加以验证，即恰可察觉差异（JND）测试，询问两个图块——一个是参考、一个是失真——是相同还是不同。这些判断是在一个广阔的失真与真实算法输出空间上采集的。

2.1. 失真 Distortions

传统失真 Traditional distortions We create a set of "traditional" distortions consisting of common operations performed on the input patches, listed in Table 2 (left). In general, we use photometric distortions, random noise, blurring, spatial shifts and corruptions, and compression artifacts. We show qualitative examples of our traditional distortions in Figure 2. The severity of each perturbation is parameterized - for example, for Gaussian blur, the kernel width determines the amount of corruption applied to the input image. We also compose pairs of distortions sequentially to increase the overall space of possible distortions. In total, we have 20 distortions and 308 sequentially composed distortions.

我们构造了一组 "传统" 失真，由施加在输入图块上的常见操作组成，列于表 2（左）。总体上，我们使用光度失真、随机噪声、模糊、空间平移与破坏，以及压缩伪影。图 2 给出了传统失真的定性示例。每种扰动的强度都被参数化——例如对于高斯模糊，核宽度决定了施加于输入图像的破坏量。我们还把成对的失真按顺序组合，以扩大可能失真的整体空间。总计我们有 20 种失真与 308 种顺序组合而成的失真。

基于 CNN 的失真 CNN-based distortions To more closely simulate the space of artifacts that can arise from deep-learning based methods, we create a set of distortions created by neural networks. We simulate possible algorithm outputs by exploring a variety of tasks, architectures, and losses, as shown in Table 2 (right). Such tasks include autoencoding, denoising, colorization, and superresolution. All of these tasks can be achieved by applying the appropriate corruption to the input. In total, we generated 96 "denoising autoencoders" and use these as CNN-based distortion functions. We train each of these networks on the 1.3M ImageNet dataset [47] for 1 epoch. The goal of each network is not to solve the task per se, but rather to explore common artifacts that plague the outputs of deep learning based methods.

为更贴近地模拟基于深度学习的方法可能产生的

子类型	失真类型
光度	亮度偏移、颜色偏移、对比度、饱和度
噪声	均匀白噪声、高斯白噪声、粉噪声、蓝噪声、高斯有色（介于紫与棕之间）噪声、棋盘伪影
模糊	高斯、双边滤波
空间	平移、仿射形变、单应、线性形变、三次形变、重影、色差
压缩	jpeg

参数类型	参数
输入破坏	null、粉噪声、白噪声、去色、下采样
生成器网络架构	层数、跳连数、带 dropout 的层数、在最高层强制跳连、上采样方法、归一化方法、首层步幅、第 1 层通道数、最大通道数
判别器	层数
损失/学习	逐像素（ℓ_1）、VGG、判别器损失的权重，学习率

表 2: **我们的失真**。我们的传统失真（左）由基本的低层图像编辑操作实现，我们还将它们按顺序组合，以更好地探索该空间。基于 CNN 的失真（右）则通过随机改变任务、网络架构与学习参数等生成。这些失真的目标，是模拟真实算法输出中可能出现的失真。

伪影空间，我们构造了一组由神经网络生成的失真。如表 2（右）所示，我们通过探索多种任务、架构与损失来模拟可能的算法输出。这些任务包括自编码、去噪、着色与超分辨率，且都可以通过对输入施加适当的破坏来实现。我们总共生成了 96 个 “去噪自编码器”，并把它们用作基于 CNN 的失真函数。每个网络都在 130 万张的 ImageNet 数据集 [47] 上训练 1 个 epoch。每个网络的目标并不在于解决任务本身，而是要探索那些困扰深度学习方法输出的常见伪影。

来自真实算法的失真图块 Distorted image patches from real algorithms The true test of an image assessment algorithm is on real problems and real algorithms. We gather perceptual judgments using such outputs. Data on real algorithms is more limited, as each application will have their own unique properties. For example, different colorization methods will not show much structural variation, but will be prone to effects such as color bleeding and color variation. On the other hand, superresolution will not have color ambiguity, but may see larger structural changes from algorithm to algorithm.

对图像评价算法的真正考验，在于真实问题与真实算法。我们利用这类输出来收集感知判断。真实算法的数据更为有限，因为每种应用都有其独特性质。例如，不同的着色方法在结构上不会有太大变化，却容易出现颜色渗溢、颜色偏差等效应；而超分辨率不存在颜色歧义，却可能因算法不同而出现更大的结构变化。

超分辨率 Superresolution We evaluate results from the 2017 NTIRE workshop [2]. We use 3 tracks from the workshop – $\times 2$, $\times 3$, $\times 4$ upsampling rates using “unknown” downsampling to create the input images. Each track had approximately 20 algorithm submissions. We also evaluate several additional methods, including bicubic upsampling, and four of the top performing deep superresolution methods [24, 59, 31, 48]. A common qualitative way of presenting superresolution results is zooming into specific patches and comparing differences. As such, we sample random 64×64 triplets from random locations of images in the Div2K [2] dataset – the ground truth high-resolution image, along with two algorithm outputs.

我们评测 2017 年 NTIRE workshop [2] 的结果。我们采用其中 3 个赛道——分别用 “未知” 下采样生成输入图像的 $\times 2$、$\times 3$、$\times 4$ 上采样倍率，每个赛道约有 20 份算法提交。我们还评测了若干额外方法，包括双三次上采样，以及四种表现最好的深度超分辨率方法 [24, 59, 31, 48]。呈现超分辨率结果的一种常见定性方式，是放大特定图块并比较差异。为此，我们从 Div2K [2] 数据集图像的随机位置采样随机的 64×64

三元组——真值高分辨率图像，连同两份算法输出。

帧插值 Frame interpolation We sample patches from different frame interpolation algorithms, including three variants of flow-based interpolation [33], CNN-based interpolation [39], and phase-based interpolation [37] on the Davis Middleburry dataset [50]. Because artifacts arising from frame interpolation may occur at different scales, we randomly rescale the image before sampling a patch triplet.

我们从不同的帧插值算法中采样图块，包括三种基于光流的插值 [33] 变体、基于 CNN 的插值 [39] 与基于相位的插值 [37]，数据取自 Davis Middleburry 数据集 [50]。由于帧插值产生的伪影可能出现在不同尺度上，我们在采样图块三元组前先对图像作随机缩放。

视频去模糊 Video deblurring We sample from the video deblurring dataset [53], along with deblurring outputs from Photoshop Shake Reduction, Weighted Fourier Aggregation [11], and three variants of a deep video deblurring method [53].

我们从视频去模糊数据集 [53] 中采样，同时纳入来自 Photoshop 抖动消减、加权傅里叶聚合 [11] 以及某深度视频去模糊方法 [53] 三个变体的去模糊输出。

着色 Colorization We sample patches using random scales on the colorization task, on images from the ImageNet dataset [47]. The algorithms are from pix2pix [22], Larsson et al. [30], and variants from Zhang et al. [63].

针对着色任务，我们在 ImageNet 数据集 [47] 的图像上以随机尺度采样图块。所用算法来自 pix2pix [22]、Larsson 等 [30]，以及 Zhang 等 [63] 的若干变体。

2.2. 心理物理相似性测量 Psychophysical Similarity Measurements

2AFC 相似性判断 2AFC similarity judgments We randomly select an image patch x and apply two distortions to produce patches x_0, x_1. We then ask a human which is closer to the original patch x, and record response $h \in \{0, 1\}$. On average, people spent approximately 3 seconds per judgment. Let $\mathcal{T}$ denote our dataset of patch triplets (x, x_0, x_1, h).

我们随机选取一个图像块 x，施加两种失真以产生图块 x_0, x_1，然后请人判断哪一个与原始图块 x 更接近，并记录回答 $h \in \{0, 1\}$。平均而言，人们每次判断约花 3 秒。令 $\mathcal{T}$ 表示我们由图块三元组 (x, x_0, x_1, h) 构成的数据集。

A comparison between our dataset and previous datasets is shown in Table 1. Previous datasets have focused on collecting large numbers of human judgments for a few images and distortion types. For example, the largest dataset, TID2013 [45], has 500k judgments on 3000 distortions (from 25 input images with 24 distortions types, each sampled at 5 levels). We provide a complementary dataset that focuses instead on a large number of distortions types. In, addition, we collect judgments on a large number of 64×64 patches rather than a small number of images. There are three reasons for this. First, the space of full images is extremely large, which makes it much harder to cover a reasonable portion of the domain with judgments (even 64×64 color patches represent an intractable 12k-dimensional space). Second, by choosing a smaller patch size, we focus on lower-level aspects of similarity, to mitigate the effect of differing "respects of similarity" that may be influenced by high-level semantics [36]. Finally, modern methods for image synthesis train deep networks with patch-based losses (implemented as convolutions) [8, 21]. Our dataset consists of over 161k patches, derived from the MIT-Adobe 5k dataset [7] (5000 uncompressed images) for training, and the RAISE1k dataset [10] for validation.

表 1 给出了我们的数据集与以往数据集的对比。以往数据集侧重于就少数图像与失真类型采集大量人类判断。例如最大的数据集 TID2013 [45] 对 3000 种失真（源自 25 幅输入图像、24 种失真类型，每种在 5 个强度上采样）收集了 50 万条判断。我们提供的则是一个互补的数据集，转而侧重于大量的失真类型；此外，我们在大量 64×64 图块（而非少量整幅图像）上采集判断。这样做有三个原因。其一，整幅图像的空间极其庞大，用判断覆盖该领域相当一部分要难得多（即便 64×64 的彩色图块也对应一个难以处理的 12k 维空间）。其二，选用较小的图块尺寸能让我们聚焦于相似性中更低层的方面，从而减轻可能受高层语义影响的 "相似的不同方面" 所带来的干扰 [36]。其三，现代

图像合成方法用基于图块的损失（以卷积实现）来训练深度网络 [8, 21]。我们的数据集包含超过 16.1 万个图块，训练部分取自 MIT-Adobe 5k 数据集 [7]（5000 幅未压缩图像），验证部分取自 RAISE1k 数据集 [10]。

To enable large-scale collection, our data is collected "in-the-wild" on Amazon Mechanical Turk, as opposed to a controlled lab setting. Crump et al. [9] show that AMT can be reliably used to replicate many psychophysics studies, despite the inability to control all environmental factors. We ask for 2 judgments per example in our "train" set and 5 judgments in our "val" sets. Asking for fewer judgments enables us to explore a larger set of image patches and distortions. We add sentinels which consist of pairs of patches with obvious deformations, e.g., a large amount of Gaussian noise vs a small amount of Gaussian noise. Approximately 90% of Turkers were able to correctly pass at least 93% of the sentinels (14 of 15), indicating that they understood the task and were paying attention. We choose to use a larger number of distortions than prior datasets.

为实现大规模采集，我们的数据是在亚马逊 Mechanical Turk 上 "自然环境" 式收集的，而非受控的实验室环境。Crump 等 [9] 表明，尽管无法控制所有环境因素，AMT 仍可可靠地用于复现许多心理物理学研究。我们在 "训练" 集上对每个样本征集 2 条判断，在 "验证" 集上征集 5 条。征集较少的判断使我们得以探索更大的图像块与失真集合。我们加入了哨兵样本，即包含明显形变的图块对，例如大量高斯噪声对少量高斯噪声。约 90% 的标注者能正确通过至少 93% 的哨兵样本（15 个中的 14 个），表明他们理解任务且保持专注。我们选择使用比以往数据集更多的失真种类。

恰可察觉差异（JND） Just noticeable differences (JND)

A potential shortcoming of the 2AFC task is that it is "cognitively penetrable," in the sense that participants can consciously choose which respects of similarity they will choose to focus on in completing the task [36], which introduces subjectivity into the judgments. To validate that the judgments actually reflected something objective and meaningful, we also collected user judgments of "just noticeable differences" (JNDs). We show a reference image, followed by a randomly distorted image, and ask a human if the images are the same or different. The two image patches are shown for 1 second each, with a 250 ms gap in between. Two images which look similar may be easily confused, and a good perceptual metric will be able to order pairs from most to least confusable. JND tests like this may be considered less subjective, since there is a single correct answer for each judgment, and participants are presumed to be aware of what correct behavior entails. We gather 3 JND observations for each of the 4.8k patches in our traditional and CNN-based validation sets. Each subject is shown 160 pairs, along with 40 sentinels (32 identical and 8 with large Gaussian noise distortion applied). We also provide a short training period of 10 pairs which contain 4 "same" pairs, 1 obviously different pair, and 5 "different" pairs generated by our distortions. We chose to do this in order to prime the users towards expecting approximately

数据集	数据来源	训练/验证	样本数	每样本判断数
传统	[7]	训练	56.6k	2
基于 CNN	[7]	训练	38.1k	2
混合	[7]	训练	56.6k	2
2AFC–失真 [训练]	–	训练	151.4k	2
传统	[10]	训练	4.7k	5
基于 CNN	[10]	训练	4.7k	5
2AFC–失真 [验证]	–	验证	9.4k	5
超分辨率	[32]	验证	10.9k	5
帧插值	[50]	验证	1.9	5
视频去模糊	[5]	验证	9.4	5
着色	[47]	验证	4.7	5
2AFC–真实算法 [验证]	–	验证	26.9k	5
传统	[10]	验证	4.8k	3
基于 CNN	[10]	验证	4.8k	3
JND–失真	–	验证	9.6k	3

表 3: **我们的数据集划分**。我们把 2AFC 数据集划分为三个主要部分：(1,2) 使用我们失真的训练集与测试集，其图块分别采样自 MIT5k [7] 与 RAISE1k [10] 数据集；(3) 一个包含真实算法输出的测试集，其图块来自多种应用。我们的 JND 数据取自传统与基于 CNN 的失真。

40% of the patch pairs to be identical. Indeed, 36.4% of the pairs were marked "same" (70.4% of sentinels and 27.9% of test pairs).

2AFC 任务的一个潜在缺点是它 "认知可渗透": 参与者在完成任务时可以有意识地选择关注相似性的哪些方面 [36]，这会给判断引入主观性。为验证这些判断确实反映了某种客观而有意义的东西，我们还采集了用户对 "恰可察觉差异"（JND）的判断。我们先展示一幅参考图像，再展示一幅随机失真的图像，并请人判断两幅图像相同还是不同。两个图块各展示 1 秒，中间间隔 250 毫秒。看起来相似的两幅图像容易被混淆，而好的感知度量应能把图像对按从最易混淆到最不易混淆排序。这样的 JND 测试可视为更少主观性，因为每次判断只有一个正确答案，且可假定参与者清楚何为正确行为。我们对传统与基于 CNN 的验证集中的 4800 个图块，每个采集 3 次 JND 观测。每位被试看到 160 对图像，外加 40 个哨兵样本（32 个完全相同，8 个施加了大量高斯噪声失真）。我们还提供了一段由 10 对图像组成的简短训练，其中含 4 对 "相同"、1 对明显不同，以及 5 对由我们的失真生成的 "不同" 图像。我们这样做，是为了让用户在预期上倾向于认为约 40% 的图块对是完全相同的。结果确有 36.4% 的图块对被标为 "相同"（哨兵样本中占 70.4%，测试图块对中占 27.9%）。

3. 深度特征空间 Deep Feature Spaces

We evaluate feature distances in different networks. For a given convolutional layer, we compute cosine distance (in the channel dimension) and average across spatial dimensions and layers of the network. We also discuss how to tune an existing network on our data.

我们评测不同网络中的特征距离。对给定的卷积层，我们（在通道维度上）计算余弦距离，并在网络的空间维度与各层上取平均。我们还讨论如何在我们的数据上微调一个已有网络。

网络架构 Network architectures We evaluate the SqueezeNet [20], AlexNet [28], and VGG [52] architectures. We use 5 `conv` layers from the VGG network, which has become the de facto standard for image generation tasks [17, 14, 8]. We also compare against the shallower AlexNet network, which may more closely match the architecture of the human visual cortex [61]. We use the `conv1-conv5` layers from [27]. Finally, the SqueezeNet architecture was designed to be extremely lightweight (2.8 MB) in size, with similar classification performance to AlexNet. We use the first `conv` layer and some subsequent "`fire`" modules.

我们评测 SqueezeNet [20]、AlexNet [28] 与 VGG [52] 三种架构。我们使用 VGG 网络的 5 个 `conv` 层——VGG 已成为图像生成任务事实上的标准 [17, 14, 8]。我们还与更浅的 AlexNet 网络作比较，后者或许更贴近人类视觉皮层的结构 [61]。我们取自 [27] 的 `conv1-conv5` 层。最后，SqueezeNet 架构被设计得极为轻量（2.8 MB），分类性能却与 AlexNet 相近。我们使用其第一个 `conv` 层以及随后的若干 "`fire`" 模块。

We additionally evaluate self-supervised methods, including puzzle-solving [40], cross-channel prediction [63, 64], learning from video [43], and generative modeling [13]. We use publicly available networks from these and other methods, which use variants of AlexNet [28].

我们还额外评测了若干自监督方法，包括拼图求解 [40]、跨通道预测 [63, 64]、从视频中学习 [43] 与生成式建模 [13]。我们使用这些方法及其他方法公开可得的网络，它们均采用 AlexNet [28] 的变体。

从网络激活到距离 Network activations to distance Figure 3 (left) and Equation 1 illustrate how we obtain the distance between reference and distorted patches x, x_0 with network $\mathcal{F}$. We extract feature stack from L layers and unit-normalize in the channel dimension, which we designate as $\hat{y}^l, \hat{y}^l_0 \in \mathbb{R}^{H_l \times W_l \times C_l}$ for layer l. We scale the activations channel-wise by vector $w^l \in \mathbb{R}^{C_l}$ and compute the ℓ_2 distance. Finally, we average spatially and sum channel-wise. Note that using $w_l = 1 \forall l$ is equivalent to computing cosine distance.

图 3（左）与式 1 说明了我们如何用网络 $\mathcal{F}$ 得到参考图块与失真图块 x, x_0 之间的距离。我们从 L 个层提取特征堆栈，并在通道维度上作单位归一化，对层 l

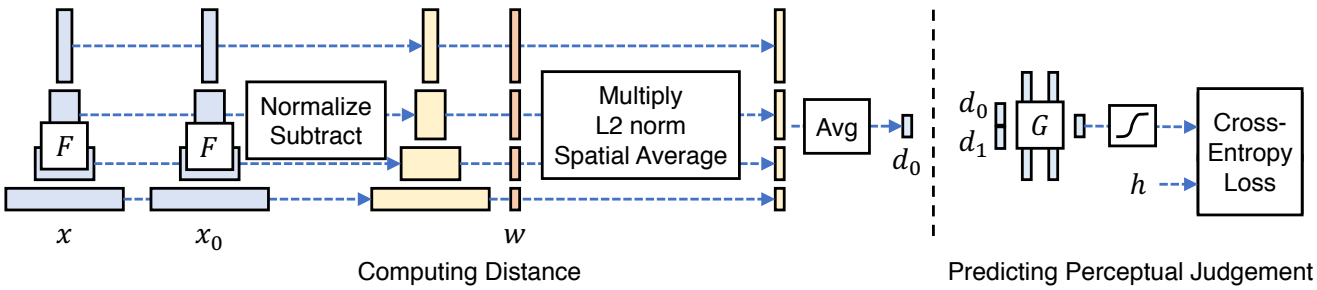


图 3: **用网络计算距离**。（左）给定网络 $\mathcal{F}$，为计算两个图块 x、x_0 之间的距离 d_0，我们先计算深度嵌入，在通道维度上对激活作归一化，再用向量 w 逐通道缩放，然后取 ℓ_2 距离。之后在空间维度以及所有层上取平均。（右）训练一个小网络 $\mathcal{G}$，从距离对 (d_0, d_1) 预测感知判断 h。

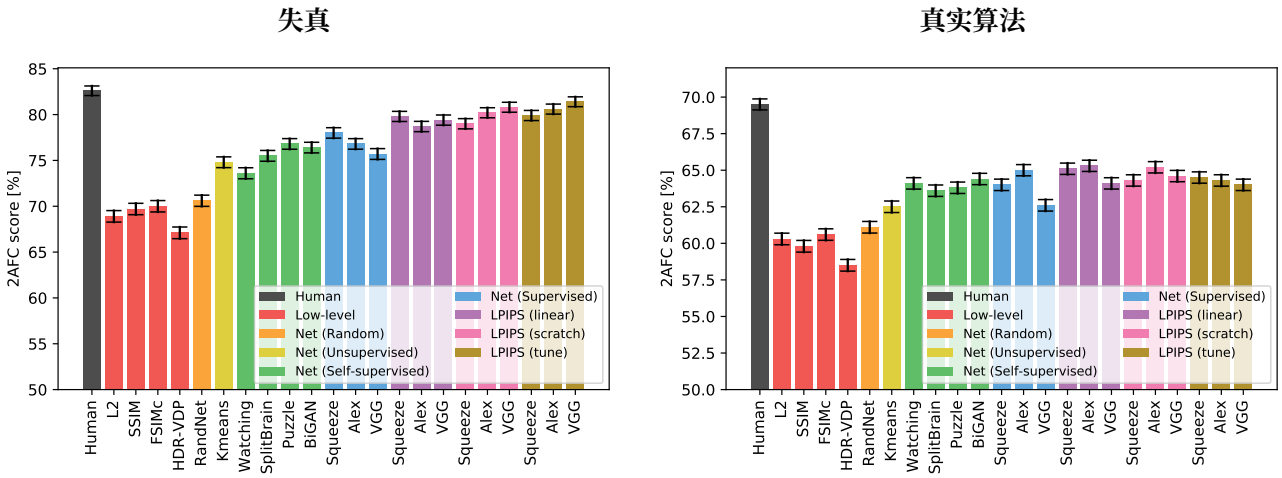


图 4: **定量比较**。我们在测试集上对各度量作定量比较。（左）在传统与基于 CNN 的失真上取平均的结果。（右）在 4 个真实算法集合上取平均的结果。

记为 $\hat{y}^l, \hat{y}^l_0 \in \mathbb{R}^{H_l \times W_l \times C_l}$。我们用向量 $w^l \in \mathbb{R}^{C_l}$ 逐通道缩放激活，并计算 ℓ_2 距离。最后在空间上取平均、在通道上求和。注意，取 $w_l = 1 \forall l$ 等价于计算余弦距离。

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} ||w_l \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw})||_2^2 \quad (1)$$

在我们的数据上训练 Training on our data We consider a few variants for training with our perceptual judgments: ***lin***, ***tune***, and ***scratch***. For the ***lin*** configuration, we keep pre-trained network weights $\mathcal{F}$ fixed, and learn linear weights w on top. This constitutes a "perceptual calibration" of a few parameters in an existing feature space. For example, for the VGG network, 1472 parameters are learned. For the ***tune*** configuration, we initialize from a pre-trained classification model, and allow all the weights for network $\mathcal{F}$ to be fine-tuned. Finally, for ***scratch***, we initialize the network from random Gaussian weights and train it entirely on our judgments. Overall, we refer to these as variants of our proposed **Learned Perceptual Image Patch Similarity (LPIPS)** metric. We illustrate the training loss function in Figure 3 (right) and describe it further in the appendix.

我们考虑用感知判断进行训练的几种变体：***lin***、***tune*** 与 ***scratch***。在 ***lin*** 配置下，我们保持预训练网络权重 $\mathcal{F}$ 固定，只在其上学习线性权重 w。这相当于对已有特征空间中的少量参数作一次 "感知校准"。例如对 VGG 网络，需学习 1472 个参数。在 ***tune*** 配置下，我们从预训练的分类模型初始化，允许网络 $\mathcal{F}$ 的所有权重被微调。最后在 ***scratch*** 配置下，我们用随机高斯权重初始化网络，完全在我们的判断上训练它。

总体上，我们把这些统称为所提出的**学习式感知图像块相似性（LPIPS）**度量的各变体。训练损失函数见图 3（右），并在附录中作进一步描述。

4. 实验 Experiments

Results on our validation sets are shown in Figure 4. We first evaluate how well our metrics and networks work. All validation sets contain 5 pairwise judgments for each triplet. Because this is an inherently noisy process, we compute agreement of an algorithm with *all* of the judgments. For example, if there are 4 preferences for x_0 and 1 for x_1, an algorithm which predicts the more popular choice x_0 would receive 80% credit. If a given example is scored with fraction p humans in one direction and $1-p$ in the other, a human would achieve score $p^2+(1-p)^2$ on expectation.

我们在验证集上的结果见图 4。我们首先评估各度量与网络的表现如何。所有验证集对每个三元组都含 5 条成对判断。由于这本质上是一个含噪过程，我们计算算法与全部判断的一致程度。例如，若有 4 票偏向 x_0、1 票偏向 x_1，则预测更受欢迎选项 x_0 的算法可得 80% 的分数。若某样本在一个方向上得到比例为 p 的人类判断、另一方向为 $1-p$，则一个人在期望上会得到 $p^2+(1-p)^2$ 的分数。

4.1. 评测 Evaluations

低层度量与分类网络表现如何? How well do low-level metrics and classification networks perform? Figure 4 shows the performance of various low-level metrics (in red), deep networks, and human ceiling (in black). The scores are averaged across the 2 distortion test sets (traditional+CNN-based) in Figure 4 (left), and 4 real algorithm benchmarks (superresolution, frame interpolation, video deblurring, colorization) in Figure 4 (right). All scores within each test set are shown in the appendix. Averaged across all 6 test sets, humans are 73.9% consistent. Interestingly, the supervised networks perform at about the same level to each other, at 68.6%, 68.9%, and 67.0%, even across variation in model sizes – SqueezeNet (2.8 MB), AlexNet (9.1 MB), and VGG (58.9 MB) (only convolutional layers are counted). They all perform better than traditional metrics ℓ_2, SSIM, and FSIM at 63.2%, 63.1%, 63.8%, respectively. Despite its common use, SSIM was not designed for situations where geometric distortion is a large factor [49].

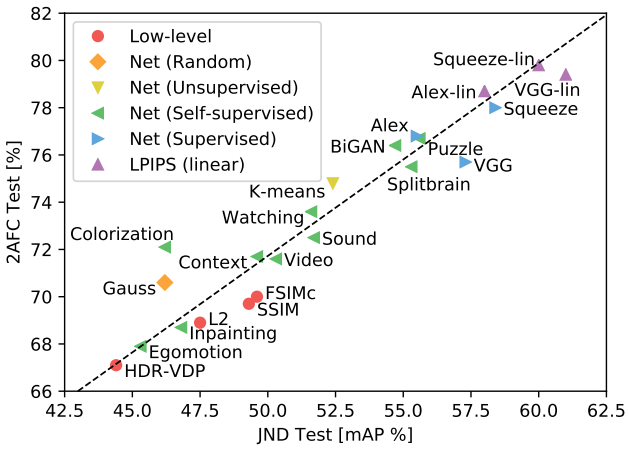


图 5: **关联各感知测试**。我们展示了各类方法的表现，包括无监督 [26]、自监督 [1, 44, 12, 57, 63, 41, 43, 40, 13, 64]、监督 [27, 52, 20]，以及我们经感知学习得到的度量（LPIPS）。分数取自我们的 2AFC 与 JND 测试，并在传统与基于 CNN 的失真上取平均。

图 4 展示了各种低层度量（红色）、深度网络以及人类上限（黑色）的表现。分数在图 4（左）中于 2 个失真测试集（传统 + 基于 CNN）上取平均，在图 4（右）中于 4 个真实算法基准（超分辨率、帧插值、视频去模糊、着色）上取平均。各测试集内的全部分数见附录。在全部 6 个测试集上取平均，人类的一致率为 73.9%。有意思的是，监督网络彼此表现相当，分别为 68.6%、68.9% 与 67.0%，即便模型大小差异明显——SqueezeNet（2.8 MB）、AlexNet（9.1 MB）与 VGG（58.9 MB）（仅计卷积层）。它们都优于传统度量 ℓ_2、SSIM 与 FSIM，后者分别为 63.2%、63.1% 与 63.8%。SSIM 虽被广泛使用，却并非为几何失真占主要因素的情形而设计 [49]。

网络是否必须在分类任务上训练? Does the network have to be trained on classification? In Figure 4, we show model performance across a variety of unsupervised and self-supervised tasks, shown in green – generative modeling with BiGANs [13], solving puzzles [40], cross-channel

		2AFC	JND	分类	检测	平均
感知	**2AFC**	–	.928	.640	.363	.644
感知	**JND**	.928	–	.612	.232	.591
PASCAL	**分类**	.640	.612	–	.429	.560
PASCAL	**检测**	.363	.232	.429	–	.341

表 4: **任务相关性**。我们跨方法计算低层感知测试与高层语义测试之间的分数相关性。感知分数在传统与基于 CNN 的失真集合上取平均。相关性分数针对类 AlexNet 架构计算。

prediction [64], and segmenting foreground objects from video [43]. These self-supervised tasks perform on par with classification networks. This indicates that tasks across a large spectrum can induce representations which transfer well to perceptual distances. Also, the performance of the stacked k-means method [26], shown in yellow, outperforms low-level metrics. Random networks, shown in orange, with weights drawn from a Gaussian, do not yield much improvement. This indicates that the combination of network structure, along with orienting filters in directions where data is more dense, can better correlate to perceptual judgments.

在图 4 中，我们展示了模型在多种无监督与自监督任务上的表现（绿色）——用 BiGAN 做生成式建模 [13]、拼图求解 [40]、跨通道预测 [64]，以及从视频中分割前景物体 [43]。这些自监督任务的表现与分类网络相当。这表明，跨越广泛谱系的任务都能诱导出可良好迁移到感知距离的表示。此外，堆叠 k-means 方法 [26]（黄色）的表现优于低层度量。而权重从高斯分布中抽取的随机网络（橙色）几乎没有带来提升。这表明，网络结构本身，加上把滤波器朝数据更稠密的方向对齐，能更好地与感知判断相关联。

In Table 5, we explore how well our perceptual task correlates to semantic tasks on the PASCAL dataset [15], using results summarized in [64], including additional self-supervised methods [1, 44, 12, 57, 63, 41]. We compute the correlation coefficient between each task (perceptual or semantic) across different methods. The correlation from our 2AFC distortion preference task to classification and detection is .640 and .363, respectively. Interestingly, this is similar to the correlation between the classification and detection tasks (.429), even though both are considered "high-level" semantic tasks, and our perceptual task is "low-level."

在表 5 中，我们借助 [64] 汇总的结果（并纳入若干额外的自监督方法 [1, 44, 12, 57, 63, 41]），考察我们的感知任务与 PASCAL 数据集 [15] 上的语义任务相关程度如何。我们跨不同方法计算各任务（感知或语义）之间的相关系数。从我们的 2AFC 失真偏好任务到分类与检测的相关系数分别为.640 与.363。有意思的是，这与分类和检测两个任务之间的相关系数（.429）相近，尽管后两者都被视作 "高层" 语义任务，而我们的感知任务是 "低层" 的。

各度量在不同感知任务间是否相关? Do metrics correlate across different perceptual tasks? We test if training for the 2AFC distortion preference test corresponds with another perceptual task, the JND test. We order patch pairs by ascending order by a given metric, and compute precision-recall on our CNN-based distortions – for a good metric, patches which are close together are more likely to be confused for being the same. We compute area under the curve, known as mAP [15]. The 2AFC distortion preference test has high correlation to JND: $\rho = .928$ when averaging the results across distortion types. Figure 5 shows how different methods perform under each perceptual test. This indicates that 2AFC generalizes to another perceptual test and is giving us signal regarding human judgments.

我们检验为 2AFC 失真偏好测试所作的训练，是否与另一项感知任务——JND 测试——相吻合。我们按给定度量对图块对升序排列，并在基于 CNN 的失真上计算精确率-召回率——对好的度量而言，彼此接近的图块更可能被混淆为相同。我们计算曲线下面积，即所谓的 mAP [15]。2AFC 失真偏好测试与 JND 高度相关：在各失真类型上取平均时 $\rho = .928$。图 5 展示了不同方法在各感知测试下的表现。这表明 2AFC 能泛化到另一项感知测试，并为我们提供了关于人类判断的信号。

我们能否在传统与基于 CNN 的失真上训练一个度量? Can we train a metric on traditional and CNN-based distortions? In Figure 4, we show performance using our ***lin***, ***scratch***, and ***tune*** configurations, shown in pur-

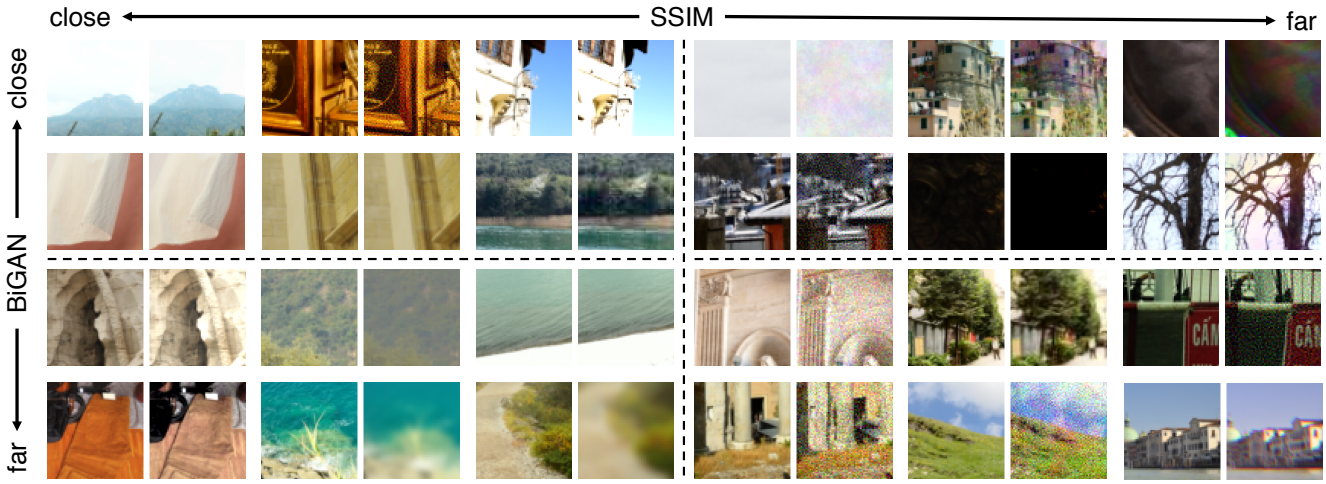


图 6: **失真上的定性比较**。我们用 SSIM [58] 度量与 BiGAN 网络 [13]，对传统失真作定性比较，展示了二者一致与不一致的例子。一个主要差异是，深度嵌入似乎对模糊更敏感。更多示例请见附录。

ple, pink, and brown, respectively. When validating on the traditional and CNN-based distortions (Figure 4(a)), we see improvements. Allowing the network to tune all the way through (brown) achieves higher performance than simply learning linear weights (purple) or training from scratch (pink). The higher capacity network VGG also performs better than the lower capacity SqueezeNet and AlexNet architectures. These results verify that networks can indeed learn from perceptual judgments.

在图 4 中，我们展示了 ***lin***、***scratch*** 与 ***tune*** 三种配置的表现，分别以紫色、粉色与棕色标示。在传统与基于 CNN 的失真上验证时（图 4(a)），我们看到了提升。允许网络端到端微调（棕色）比仅学习线性权重（紫色）或从头训练（粉色）取得更高性能。容量更大的 VGG 网络也优于容量更小的 SqueezeNet 与 AlexNet 架构。这些结果验证了网络确实能从感知判断中学习。

在传统与基于 CNN 的失真上训练能否迁移到真实场景? Does training on traditional and CNN-based distortions transfer to real-world scenarios? We are more interested in how performance generalizes to *real-world algorithms*, shown in Figure 4(b). The SqueezeNet, AlexNet, and VGG architectures start at 64.0%, 65.0%, and 62.6%, respectively. Learning a linear classifier (purple) improves performance for all networks. Across the 3 networks and 4 real-algorithm tasks, 11 of the 12 scores improved, indicating that "calibrating" activations on a pre-existing representation using our data is a safe way to achieve a small boost in performance (1.1%, 0.3%, and 1.5%, respectively). Training a network from scratch (pink) yields slightly lower performance for AlexNet, and slightly higher performance for VGG than linear calibration. However, these still outperform low-level metrics. This indicates that the distortions we have expressed do project onto our test-time tasks of judging real algorithms.

我们更关心性能如何泛化到真实世界的算法，见图 4(b)。SqueezeNet、AlexNet 与 VGG 架构的起点分别为 64.0%、65.0% 与 62.6%。学习一个线性分类器（紫色）使所有网络的性能都得到提升。在 3 个网络与 4 个真实算法任务的 12 个分数中，有 11 个得到提高，表明用我们的数据在既有表示上 "校准" 激活，是获得小幅性能提升的稳妥途径（三者分别提升 1.1%、0.3% 与 1.5%）。从头训练网络（粉色）对 AlexNet 的性能略低于线性校准，对 VGG 则略高；但它们仍优于低层度量。这表明，我们所刻画的失真确实能投射到测试时评判真实算法的任务上。

Interestingly, starting with a pre-trained network and tuning throughout *lowers* transfer performance. This is an interesting negative result, as training for a low-level perceptual task directly does not necessarily perform as well as transferring a representation trained for the high-level task.

有意思的是，从预训练网络出发并全程微调反而降低了迁移性能。这是一个耐人寻味的反面结果：直接为低层感知任务训练，未必比迁移一个为高层任务训练的表示表现更好。

深度度量与低层度量在何处产生分歧？ Where do deep metrics and low-level metrics disagree? In Figure 11, we show a qualitative comparison across our traditional distortions for a deep method, BiGANs [13], and a representation traditional perceptual method, SSIM [58]. Pairs which BiGAN perceives to be far but SSIM to be close generally contain some blur. BiGAN tends to perceive correlated noise patterns to be a smaller distortion than SSIM.

在图 11 中，我们就传统失真，对一种深度方法 BiGAN [13] 与一种有代表性的传统感知方法 SSIM [58] 作定性比较。那些被 BiGAN 判为远、却被 SSIM 判为近的图块对，通常都带有一定模糊。相比 SSIM，BiGAN 倾向于把相关噪声模式视为更小的失真。

5. 结论 Conclusions

Our results indicate that networks trained to solve challenging visual prediction and modeling tasks end up learning a representation of the world that correlates well with perceptual judgments. A similar story has recently emerged in the representation learning literature: networks trained on self-supervised and unsupervised objectives end up learning a representation that is also effective at semantic tasks [12]. Interestingly, recent findings in neuroscience make much the same point: representations trained on computer vision tasks also end up being effective models of neural activity in macaque visual cortex [61]. Moreover (and roughly speaking), the stronger the representation is at the computer vision task, the stronger it is as a model of cortical activity. Our paper makes a similar finding: the stronger a feature set is at classification and detection, the stronger it is as a model of perceptual similarity judgments, as suggested in Table 4. Together, these results suggest that a good feature is a good feature. Features that are good at semantic tasks are also good at self-supervised and unsupervised tasks, and also provide good models of both human perceptual behavior and macaque neural activity. This last point aligns with the "rational analysis" explanation of visual cognition [4], suggesting that the idiosyncrasies of biological perception arise as a consequence of a rational agent attempting to solve natural tasks. Further refining the degree to which this is true is an important question for future research.

我们的结果表明，为解决具有挑战性的视觉预测与建模任务而训练的网络，最终学到了一种与感知判断高度相关的世界表示。表示学习文献中近来出现了类似的现象：以自监督与无监督目标训练的网络，最终学到的表示在语义任务上同样有效 [12]。有意思的是，神经科学的近期发现也道出了大致相同的道理：为计算机视觉任务训练的表示，最终也成了猕猴视觉皮层神经活动的有效模型 [61]。而且（粗略地说），表示在计算机视觉任务上越强，作为皮层活动模型也越强。本文得出了类似的发现：如表 4 所示，一组特征在分类与检测上越强，作为感知相似性判断的模型也越强。综合来看，这些结果提示：好特征就是好特征。在语义任务上表现好的特征，在自监督与无监督任务上也表现好，同时还能为人类的感知行为与猕猴的神经活动提供良好的模型。最后这一点与视觉认知的 "理性分析" 解释 [4]相吻合，即生物感知的种种特性，是一个理性主体试图解决自然任务的产物。进一步厘清这一说法在多大程度上成立，是未来研究的一个重要问题。

致谢 Acknowledgements This research was supported, in part, by grants from Berkeley Deep Drive, NSF IIS-1633310, and hardware donations by NVIDIA. We thank members of the Berkeley AI Research Lab and Adobe Research for helpful discussions. We thank Alan Bovik for his insightful comments. We also thank Radu Timofte, Zhaowen Wang, Michael Waechter, Simon Niklaus, and Sergio Guadarrama for help preparing data. RZ is partially supported by an Adobe Research Fellowship and much of this work was done while RZ was an intern at Adobe Research.

本研究部分受 Berkeley Deep Drive、NSF IIS-1633310 项目资助，并得到 NVIDIA 的硬件捐赠。感谢 Berkeley AI Research Lab 与 Adobe Research 的同仁提供的有益讨论，感谢 Alan Bovik 富有洞见的意见，也感谢 Radu Timofte、Zhaowen Wang、Michael Waechter、Simon Niklaus 与 Sergio Guadarrama 在数据准备上的

帮助。RZ 部分受 Adobe Research Fellowship 资助，本文的大部分工作是在 RZ 于 Adobe Research 实习期间完成的。

参考文献

[1] P. Agrawal, J. Carreira, and J. Malik. Learning to see by moving. In *ICCV*, pages 37–45, 2015. 12, 13

[2] E. Agustsson and R. Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPR Workshops*, July 2017. 7

[3] S. Ali Amirshahi, M. Pedersen, and S. X. Yu. Image quality assessment by comparing cnn features between images. *Electronic Imaging*, 2017(12):42–51, 2017. 5, 6

[4] J. R. Anderson. *The adaptive character of thought*. Psychology Press, 1990. 15

[5] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *IJCV*, 2011. 9

[6] A. Berardino, V. Laparra, J. Ballé, and E. Simoncelli. Eigen-distortions of hierarchical representations. In *NIPS*, 2017. 6

[7] V. Bychkovsky, S. Paris, E. Chan, and F. Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *CVPR*, 2011. 4, 8, 9

[8] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. *ICCV*, 2017. 3, 8, 9, 10

[9] M. J. Crump, J. V. McDonnell, and T. M. Gureckis. Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS one*, 2013. 9

[10] D.-T. Dang-Nguyen, C. Pasquini, V. Conotter, and G. Boato. Raise: a raw images dataset for digital image forensics. In *Proceedings of the 6th ACM Multimedia Systems Conference*, pages 219–224. ACM, 2015. 4, 8, 9

[11] M. Delbracio and G. Sapiro. Hand-held video deblurring via efficient fourier aggregation. *IEEE Transactions on Computational Imaging*, 1(4):270–283, 2015. 8

[12] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015. 12, 13, 15

[13] J. Donahue, P. Krähenbühl, and T. Darrell. Adversarial feature learning. *ICLR*, 2017. 1, 3, 4, 10, 12, 13, 14, 15, 21, 23

[14] A. Dosovitskiy and T. Brox. Generating images with perceptual similarity metrics based on deep networks. In *HIPS*, pages 658–666, 2016. 3, 10

[15] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 13

[16] F. Gao, Y. Wang, P. Li, M. Tan, J. Yu, and Y. Zhu. Deepsim: Deep similarity for image quality assessment. *Neurocomputing*, 2017. 5, 6

[17] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423, 2016. 3, 10

[18] D. Ghadiyaram and A. C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *TIP*, 2016. 5

[19] N. Goodman. Seven strictures on similarity. *Problems and Projects*, 1972. 2

[20] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and $<$ 0.5 mb model size. *CVPR*, 2017. 1, 3, 10, 12, 21

[21] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 8, 9

[22] P. Isola, D. Zoran, D. Krishnan, and E. H. Adelson. Learning visual groups from co-occurrences in space and time. *ICCV Workshop*, 2016. 8

[23] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *ECCV*, 2016. 3

[24] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, pages 1646–1654, 2016. 7

[25] J. Kim and S. Lee. Deep learning of human visual sensitivity in image quality assessment framework. In *CVPR*, 2017. 5

[26] P. Krähenbühl, C. Doersch, J. Donahue, and T. Darrell. Data-dependent initializations of convolutional neural networks. *International Conference on Learning Representations*, 2016. 1, 3, 4, 12, 13, 21

[27] A. Krizhevsky. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*, 2014. 1, 10, 12, 21, 23

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 3, 10

[29] E. C. Larson and D. M. Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 2010. 4, 5

[30] G. Larsson, M. Maire, and G. Shakhnarovich. Learning representations for automatic colorization. *ECCV*, 2016. 8

[31] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *CVPR*, 2016. 7

[32] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. In *CVPR Workshops*, 2017. 9

[33] C. Liu et al. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Massachusetts Institute of Technology, 2009. 8

[34] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. In *ACM Transactions on Graphics (TOG)*, 2011. 2, 21

[35] A. B. Markman and D. Gentner. Nonintentional similarity processing. *The new unconscious*, pages 107–137, 2005. 2

[36] D. L. Medin, R. L. Goldstone, and D. Gentner. Respects for similarity. *Psychological review*, 100(2):254, 1993. 2, 8, 9, 10

[37] S. Meyer, O. Wang, H. Zimmer, M. Grosse, and A. Sorkine-Hornung. Phase-based frame interpolation for video. In *CVPR*, pages 1410–1418, 2015. 8

[38] N. Murray, L. Marchesotti, and F. Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012. 5

[39] S. Niklaus, L. Mai, and F. Liu. Video frame interpolation via adaptive separable convolution. In *ICCV*, 2017. 8

[40] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. *ECCV*, 2016. 1, 3, 4, 10, 12, 13, 21

[41] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman. Visually indicated sounds. *CVPR*, 2016. 12, 13

[42] A. Paszke, S. Chintala, R. Collobert, K. Kavukcuoglu, C. Farabet, S. Bengio, I. Melvin, J. Weston, and J. Mariethoz. Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration, may 2017. 21, 22

[43] D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan. Learning features by watching objects move. *CVPR*, 2017. 1, 10, 12, 13, 21

[44] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 12, 13

[45] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 2015. 4, 5, 8, 18, 23

[46] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. Tid2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 2009. 4, 5

[47] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 1, 6, 7, 8, 9

[48] M. S. Sajjadi, B. Schölkopf, and M. Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. *ICCV*, 2017. 7

[49] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey. Complex wavelet structural similarity: A new image similarity index. *TIP*, 2009. 3, 4, 12, 23, 24

[50] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 2002. 8, 9

[51] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *TIP*, 2006. 4, 5

[52] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014. 1, 3, 10, 12, 21

[53] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang. Deep video deblurring for hand-held cameras. In *CVPR*, 2017. 8

[54] H. Talebi and P. Milanfar. Learned perceptual image enhancement. *ICCP*, 2018. 5

[55] H. Talebi and P. Milanfar. Nima: Neural image assessment. *TIP*, 2018. 5

[56] A. Tversky. Features of similarity. *Psychological review*, 1977. 2

[57] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, 2015. 12, 13

[58] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 2, 3, 4, 14, 15, 21, 23

[59] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang. Deep networks for image super-resolution with sparse prior. In *ICCV*, 2015. 7

[60] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *Signals, Systems and Computers*. IEEE, 2004. 2

[61] D. L. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 2016. 10, 15

[62] L. Zhang, L. Zhang, X. Mou, and D. Zhang. Fsim: A feature similarity index for image quality assessment. *TIP*, 2011. 2, 3, 4, 21, 23

[63] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. *ECCV*, 2016. 8, 10, 12, 13

[64] R. Zhang, P. Isola, and A. A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *CVPR*, 2017. 1, 3, 4, 10, 12, 13, 21

附录 Appendix

We show full quantitative details in Appendix A. We also discuss training details in Appendix B. Finally, we show results on the TID2013 dataset [45] in Appendix C.

我们在附录 A 中给出完整的定量细节，在附录 B 中讨论训练细节，最后在附录 C 中展示在 TID2013 数据集 [45] 上的结果。

A. 定量结果 Quantitative Results

In Table 5, we show full quantitative results across all validation sets and considered metrics, including low-level metrics, along with random, unsupervised, self-supervised, supervised, and perceptually-learned networks.

在表 5 中，我们展示了所有验证集与所考虑度量的完整定量结果，包括低层度量，以及随机、无监督、自监督、监督与经感知学习的网络。

In Figures 7, 8, 9, we plot performance in individual validation sets. Figure 7 shows our traditional and CNN-based distortions, and Figures 8, 9 show results on real algorithm applications individually.

在图 7、8、9 中，我们分别绘制了各个验证集上的表现。图 7 展示传统与基于 CNN 的失真，图 8、9 则分别展示在真实算法应用上的结果。

人类表现 Human performance If humans chose patches $\{x_1,x_0\}$ with fraction $\{p,1-p\}$, the theoretical maximum for an oracle is $\max(p, 1-p)$. However, human performance is lower. If an agent chooses them with probability $\{q,1-q\}$, the agent will agree with $qp + (1-q)(1-p)$ humans on expectation. With a human agent, $q = p$, the expected human score is $p^2 + (1-p)^2$.

若人类以比例 $\{p,1-p\}$ 选择图块 $\{x_1,x_0\}$，则一个理想判断者（oracle）的理论上限为 $\max(p, 1-p)$。然而人类的表现低于此。若某主体以概率 $\{q,1-q\}$ 作选择，则该主体在期望上会与比例为 $qp + (1-q)(1-p)$

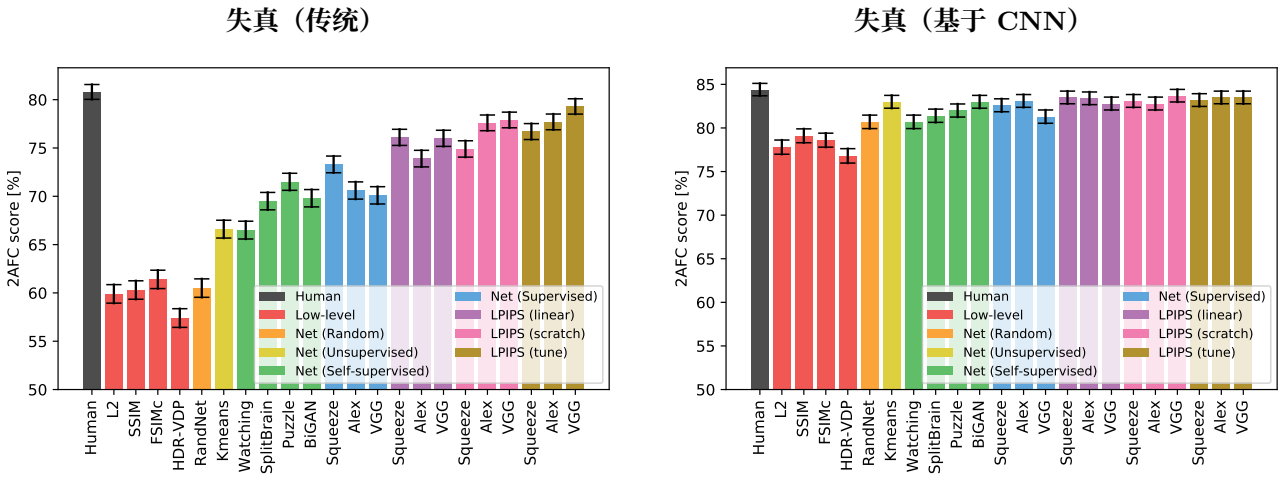


图 7: **单项结果**。（左）传统失真（右）基于 CNN 的失真

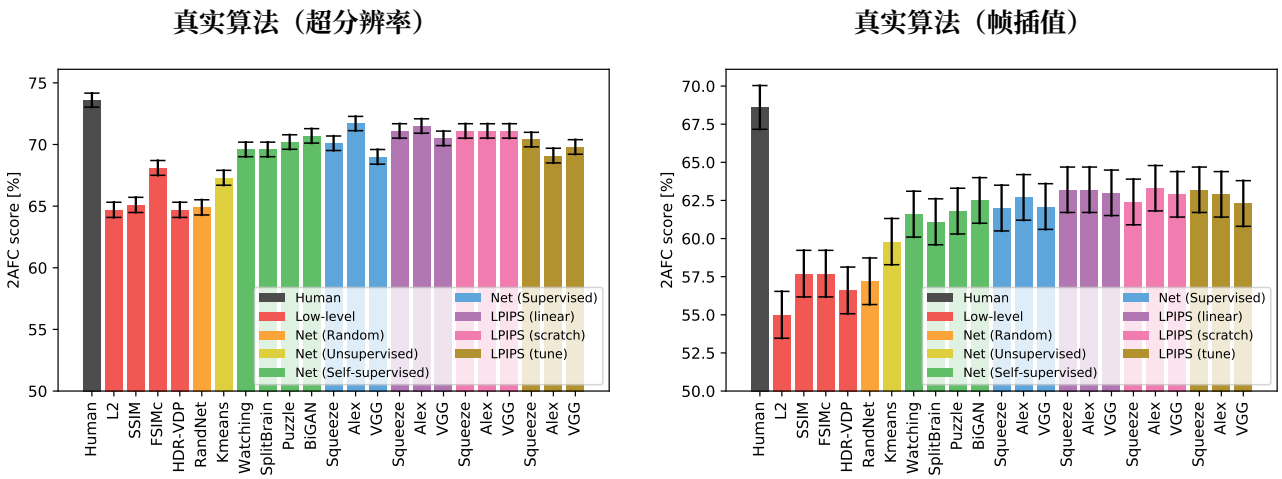


图 8: **单项结果**。（左）超分辨率（右）帧插值

的人类一致。对人类主体而言 $q = p$，故人类的期望得分为 $p^2 + (1-p)^2$。

对网络作线性校准 Linearly calibrating networks Learning linear weights on top of the ***Alex*** model achieves state-of-the-art results on the real algorithms test set. The ***linear*** models have a learned linear layer on top of each channel, whereas the out-of-the-box versions weight each channel equally. In Figure 10b, we show the learned weights for the ***Alex –frozen*** model. The `conv1-5` layers contain 64, 192, 384, 256, and 256 channels, respectively, for a total of 1152 weights. For each layer, `conv1-5`, 79.7%, 71.4%, 56.8%, 46.5%, 27.7%, respectively, of the weights are zero. This means that a majority of the `conv1` and `conv2` units are ignored, and almost all of the `conv5` units are used. Overall, about half of the units are ignored. Taking the cosine distance is equivalent to setting all weights to 1 (Figure 10a).

在 ***Alex*** 模型之上学习线性权重，在真实算法测试集上取得了当时最优的结果。***linear*** 模型在每个通道之上有一层学到的线性层，而开箱即用的版本对每个通道等权对待。在图 10b 中，我们展示 ***Alex –frozen*** 模型学到的权重。`conv1-5` 各层分别含 64、192、384、256 与 256 个通道，共计 1152 个权重。对 `conv1-5` 各层，权重为零的比例分别为 79.7%、71.4%、56.8%、46.5% 与 27.7%。这意味着 `conv1` 与 `conv2` 的多数单元被忽

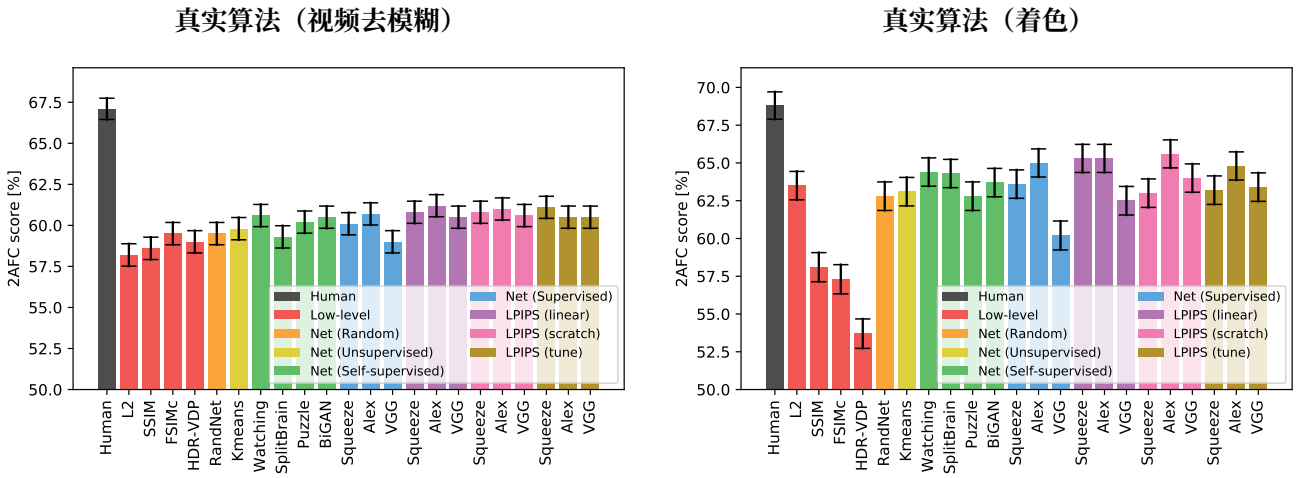


图 9: **单项结果**。（左）视频去模糊（右）着色

略，而 conv5 的单元几乎全部被使用；总体上约有一半单元被忽略。取余弦距离等价于把所有权重设为 1（图 10a）。

在失真上训练模型所需的数据量 Data quantity for training models on distortions The performance of the validation set on our distortions (80.6% and 81.4% for ***Alex*** + ***tune*** and ***VGG*** + ***tune***, respectively), is almost equal to human performance of 82.6%. This indicates that our training set size of 150k patch pairs and 300k judgments is nearly large enough to fully explore the traditional and CNN-based distortions which we defined. However, there is a small gap between the ***tune*** and ***scratch*** models (0.4% and 0.6% for ***Alex*** and ***VGG***, respectively).

在我们失真上的验证集表现（***Alex*** + ***tune*** 与 ***VGG*** + ***tune*** 分别为 80.6% 与 81.4%），几乎等同于 82.6% 的人类表现。这表明，我们 15 万对图块、30 万条判断的训练集规模，已几乎足以充分探索我们所定义的传统与基于 CNN 的失真。不过，***tune*** 与 ***scratch*** 模型之间仍有小幅差距（***Alex*** 与 ***VGG*** 分别为 0.4% 与 0.6%）。

B. 模型训练细节 Model Training Details

We illustrate the loss function for training the network in Figure 3 (right) and describe it further in the supplementary material. Given two distances, (d_0, d_1), we train a small network $\mathcal{G}$ on top to map to a score $\hat{h} \in (0, 1)$. The architecture uses two 32-channel FC-ReLU layers, followed by a 1-channel FC layer and a sigmoid. Our final loss function is shown in Equation 2.

我们在图 3（右）中示意了用于训练网络的损失函数，并在补充材料中进一步描述。给定两个距离 (d_0, d_1)，我们在其上训练一个小网络 $\mathcal{G}$，将其映射为一个分数 $\hat{h} \in (0, 1)$。该架构使用两个 32 通道的 FC-ReLU 层，随后接一个 1 通道的 FC 层与一个 sigmoid。最终的损失函数见式 2。

$$\begin{aligned}\mathcal{L}(x, x_0, x_1, h) = &-h \log \mathcal{G}(d(x, x_0), d(x, x_1)) \\ &-(1-h) \log(1 - \mathcal{G}(d(x, x_0), d(x, x_1)))\end{aligned} \tag{2}$$

In preliminary experiments, we also tried a ranking loss, which attempts to force a constant margin between patch pairs $d(x, x_0)$ and $d(x, x_1)$. We found that using a learned network, rather than enforcing the same margin in all cases, worked better.

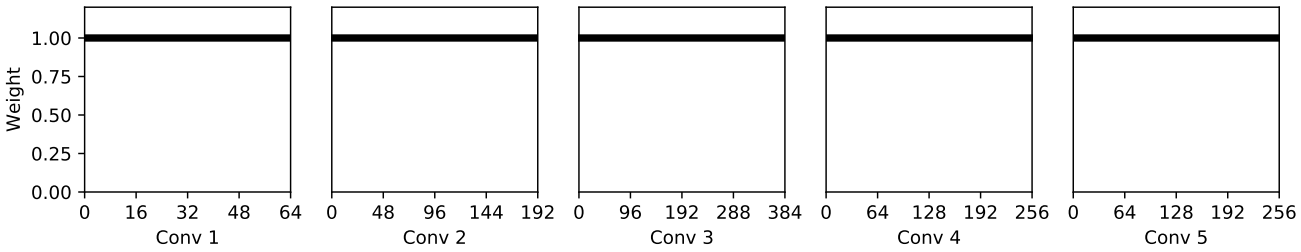
在初步实验中，我们还尝试过一种排序损失，它试图在图块对 $d(x, x_0)$ 与 $d(x, x_1)$ 之间强制一个恒定的间隔。我们发现，使用一个学到的网络，比在所有情况下都强制相同间隔效果更好。

Here, we provide some additional details on model training for our networks trained on distortions. We train with 5 epochs at initial learning rate 10^{-4}, 5 epochs with linear decay, and batch size 50. Each training patch pair is

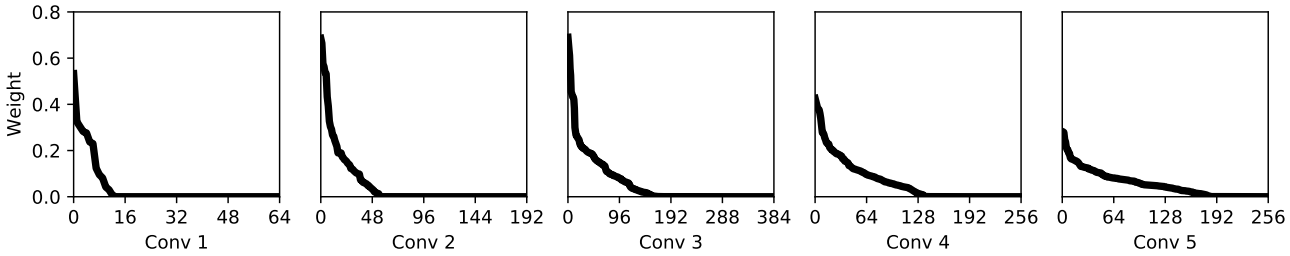
子类型	度量	失真			真实算法					全部
		传统	基于 CNN	全部	超分辨率	视频去模糊	着色	帧插值	全部	全部
理想上限	人类	80.8	84.4	82.6	73.4	67.1	68.8	68.6	69.5	73.9
低层	L2	59.9	77.8	68.9	64.7	58.2	63.5	55.0	60.3	63.2
	SSIM [58]	60.3	79.1	69.7	65.1	58.6	58.1	57.7	59.8	63.1
	FSIMc [62]	61.4	78.6	70.0	68.1	59.5	57.3	57.7	60.6	63.8
	HDR-VDP [34]	57.4	76.8	67.1	64.7	59.0	53.7	56.6	58.5	61.4
网络（随机）	Gaussian	60.5	80.7	70.6	64.9	59.5	62.8	57.2	61.1	64.3
网络（无监督）	K-means [26]	66.6	***83.0***	74.8	67.3	59.8	63.1	59.8	62.5	66.6
网络（自监督）	Watching [43]	66.5	80.7	73.6	69.6	60.6	64.4	61.6	64.1	67.2
	Split-Brain [64]	69.5	81.4	75.5	69.6	59.3	64.3	61.1	63.6	67.5
	Puzzle [40]	71.5	82.0	76.8	70.2	60.2	62.8	61.8	63.8	68.1
	BiGAN [13]	69.8	***83.0***	76.4	70.7	60.5	63.7	62.5	64.4	***68.4***
网络（监督）	SqueezeNet [20]	**73.3**	***82.6***	**78.0**	70.1	60.1	63.6	62.0	64.0	***68.6***
	AlexNet [27]	70.6	**83.1**	76.8	**71.7**	60.7	65.0	62.7	***65.0***	**68.9**
	VGG [52]	70.1	81.3	75.7	69.0	59.0	60.2	62.1	62.6	67.0
*LPIPS（学习式感知图像块相似性）	Squeeze – lin	76.1	83.5	79.8	71.1	***60.8***	***65.3***	***63.2***	***65.1***	70.0
	Alex – lin	73.9	83.4	78.7	***71.5***	**61.2**	***65.3***	***63.2***	**65.3**	69.8
	VGG – lin	76.0	82.8	79.4	70.5	60.5	62.5	***63.0***	64.1	69.2
	Squeeze – scratch	74.9	83.1	79.0	71.1	***60.8***	63.0	62.4	64.3	69.2
	Alex – scratch	77.6	82.8	80.2	71.1	***61.0***	**65.6**	**63.3**	***65.2***	**70.2**
	VGG – scratch	77.9	**83.7**	80.8	71.1	60.6	64.0	***62.9***	64.6	70.0
	Squeeze – tune	76.7	83.2	79.9	70.4	***61.1***	63.2	***63.2***	64.5	69.6
	Alex – tune	77.7	83.5	80.6	69.1	60.5	64.8	***62.9***	64.3	69.7
	VGG – tune	**79.3**	83.5	**81.4**	69.8	60.5	63.4	62.3	64.0	69.8

表 5: **结果**。我们展示了一系列方法与测试集上的 2AFC 分数（越高越好）。**加粗并加下划线** 的值为表现最高者；*加粗并斜体*的值处于最高值 0.5% 以内。*LPIPS 度量在相同的传统与基于 CNN 的失真上训练，因此在同类失真类型上测试时（即便测试图像未曾见过）相对其他方法占有优势。这些值以灰色标示；每列中最好的灰色值另以**加粗**表示。

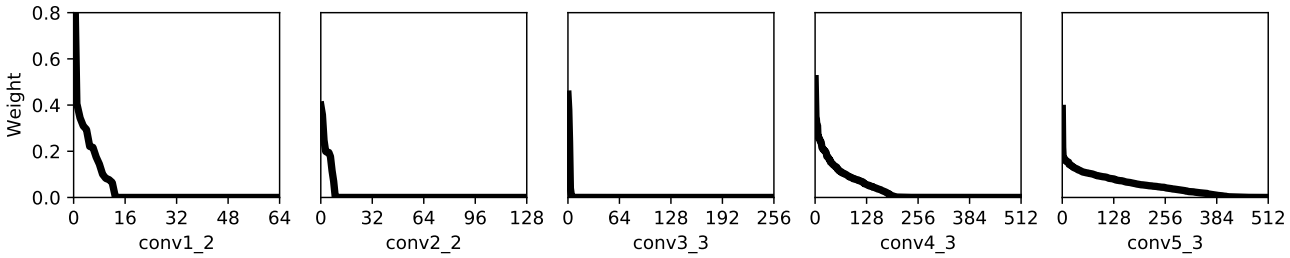
judged 2 times, and the judgments are grouped together. If, for example, the two judges are split, then the classification target (h in Figure 3) will be set at 0.5. We enforce non-negative weightings on the linear layer w, since larger distances in a certain feature should not result in two patches becoming closer in the distance metric. This is done by projecting the weights into the constraint set at every iteration. In other words, we check for any negative weights, and force them to be 0. The project was implemented using `PyTorch` [42].



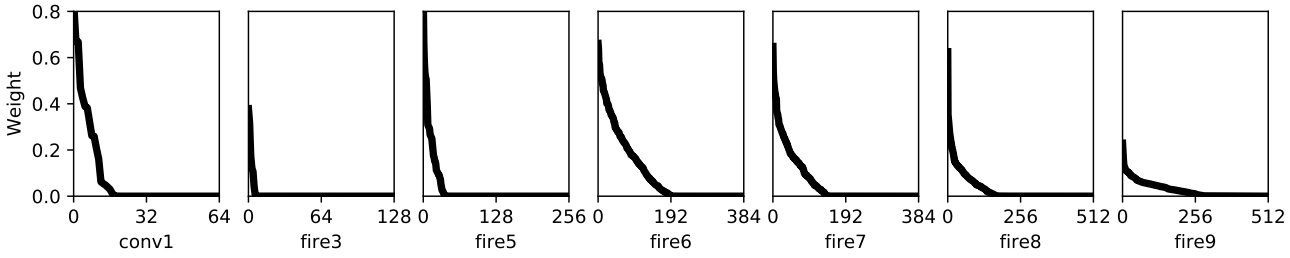
(a) **AlexNet 模型的未学习权重（余弦距离）**



(b) ***Alex–lin* 模型学到的权重**



(c) ***VGG–lin* 模型学到的权重**



(d) ***Squeeze–lin* 模型学到的权重**

图 10: **按层展示的学习到的线性权重**。 (a) 未学习权重对应于对每层每个通道都取权重 1，其结果即计算余弦距离。(b) 我们展示 *Alex–lin* 模型各层学到的权重，即图 3 中的 w 项。每个子图展示某一层的通道权重，按降序排列；横轴为通道编号，纵轴为权重。权重被限制为非负，因为图像块之间不应有负距离。(c,d) 同 (b)，但分别对应 *VGG–lin* 与 *Squeeze–lin* 模型。

这里，我们为在失真上训练的网络补充一些模型训练细节。我们以初始学习率 10^{-4} 训练 5 个 epoch，再以线性衰减训练 5 个 epoch，批大小为 50。每对训练图块被判断 2 次，并将两次判断归并在一起。举例来说，若两位判断者意见相左，则分类目标（图 3 中的 h）设为 0.5。我们对线性层 w 施加非负约束，因为某一特征上更大的距离不应导致两个图块在距离度量上变得更近。这通过在每次迭代把权重投影到约束集来实现；换言之，我们检查是否存在负权重，并强制将其置 0。该项目使用 `PyTorch` [42] 实现。

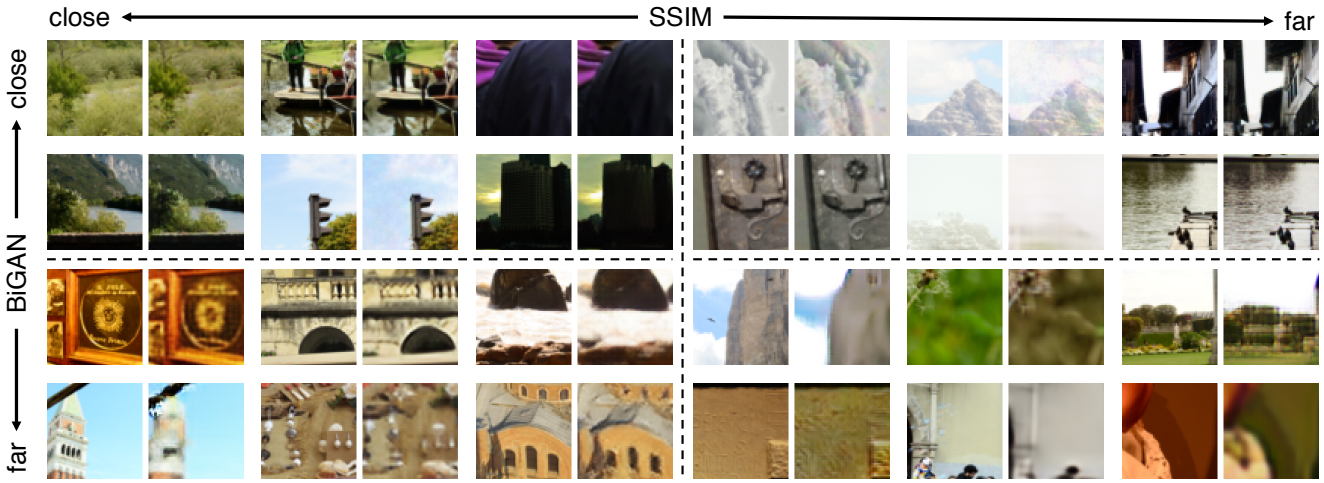


图 11: **失真上的定性比较**。我们用 SSIM [58] 度量与 BiGAN 网络 [13]，对基于 CNN 的失真作定性比较。我们既展示二者一致地认为图块更近或更远的例子，也展示二者产生分歧的例子。一个主要差异是，深度嵌入似乎对模糊更敏感。更多示例请见附录。

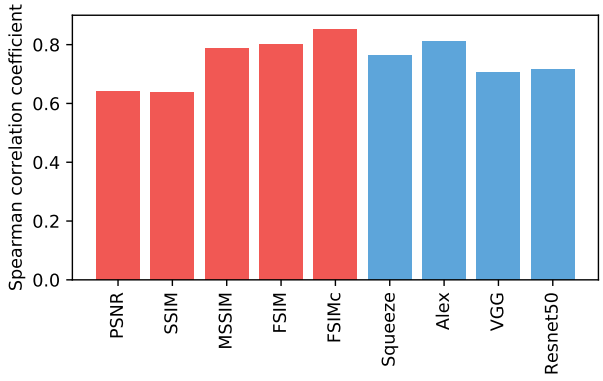


图 12: **TID 数据集**。我们展示各方法在 TID2013 数据集 [45] 上的 Spearman 相关系数。注意，为分类而训练的深度网络开箱即用便表现良好（蓝色）。

C. TID2013 数据集 TID2013 Dataset

In Figure 12, we compute scores on the TID2013 [45] dataset. We test the images at a different resolutions, using $\{128, 192, 256, 384, 512\}$ for the smaller dimension. We note that even averaging across all scales and layers, with no further calibration, the AlexNet [27] architecture gives scores near the highest metric, FSIMc [62]. On our traditional perturbations, the FSIMc metric achieves 61.4%, close to ℓ_2 at 59.9%, while the deep classification networks we tested achieved 73.3%, 70.6%, and 70.1%, respectively. The difference is likely due to the inclusion of geometric distortions in our dataset. Despite their frequent use in such situations, metrics such as SSIM were not designed to handle geometric distortions [49].

在图 12 中，我们计算在 TID2013 [45] 数据集上的分数。我们在不同分辨率下测试图像，较小的一维取 $\{128, 192, 256, 384, 512\}$。我们注意到，即便在所有尺度与层上取平均、且不作任何进一步校准，AlexNet [27] 架构给出的分数也接近最高的度量 FSIMc [62]。在我们的传统扰动上，FSIMc 度量达到 61.4%，接近 ℓ_2 的 59.9%，而我们测试的深度分类网络分别达到 73.3%、70.6% 与 70.1%。这一差异很可能源于我们的数据集纳入了几何失真。SSIM 等度量尽管常被用于此类情形，却并非为处理几何失真而设计 [49]。

D. 修订记录 Changelog

v1 initial preprint release

初始预印本发布

v2 CVPR camera ready; moved TID results (Appendix C), SSIM vs BiGAN (Figure 11), and some training details into the Appendix to fit into 8 page limit; clarified that SSIM was not designed to handle geometric distortions [49] and clarified that our dataset is a perceptual

similarity dataset (as opposed to an IQA dataset); added linear weights for *Squeeze-lin* and *VGG-lin* architectures in Figure 10; miscellaneous small edits to text.

CVPR 定稿版；为满足 8 页篇幅限制，将 TID 结果（附录 C）、SSIM 与 BiGAN 的对比（图 11）以及部分训练细节移入附录；澄清 SSIM 并非为处理几何失真而设计 [49]，并澄清我们的数据集是感知相似性数据集（而非 IQA 数据集）；在图 10 中新增 *Squeeze-lin* 与 *VGG-lin* 架构的线性权重；以及对正文的若干小改动。