Improving Visual Representation Learning through Perceptual Understanding

Samyakh Tukra Frederick Hoffman Ken Chatfield
Tractable AI
{samyakh.tukra,frederick.hoffman,ken}@tractable.ai

Abstract

We present an extension to masked autoencoders (MAE) which improves on the representations learnt by the model by explicitly encouraging the learning of higher scene-level features. We do this by: (i) the introduction of a perceptual similarity term between generated and real images (ii) incorporating several techniques from the adversarial training literature including multi-scale training and adaptive discriminator augmentation. The combination of these results in not only better pixel reconstruction but also representations which appear to capture better higher-level details within images. More consequentially, we show how our method, Perceptual MAE, leads to better performance when used for downstream tasks outperforming previous methods. We achieve 78.1% *top-1 accuracy linear probing on ImageNet-1K and up to* 88.1% *when fine-tuning, with similar results for other downstream tasks, all without use of additional pre-trained models or data.*

1. Introduction

Self-supervision provides a powerful framework for training deep neural networks without relying on explicit supervision or labels where learning proceeds by predicting one part of the input data from another. Approaches based on denoising autoencoders [45], where the input is masked and the missing parts reconstructed, have shown to be effective for pre-training in NLP with BERT [8], and more recently similar techniques have been applied for learning visual representations from images [1,4,17,30]. Such methods effectively use image reconstruction as a *pretext task* on the basis that by learning to predict missing patches useful representations can be learnt for downstream tasks.

One challenge when applying such techniques to images is that, unlike language where words contain some level of semantic meaning by design, the pixels in images are natural signals containing high-frequency variations. Therefore, image-based denoising autoencoders have been adapted to avoid learning trivial solutions to reconstruction based on local textures or patterns. BEiT [1] uses an intermediary codebook of patches such that pixels are not reconstructed directly, whilst MAE [17] masks a high proportion of the image to force the model to learn how to reconstruct whole scenes with limited context.

In this paper, we build upon MAE and ask how we can move beyond the implicit conditioning of high masking ratios to explicitly incorporate the learning of higher-order 'semantic' features into the learning objective. To do this, we focus on introducing scene-level information by adding a perceptual loss term [22]. This works by constraining feature map similarity with a second pre-trained network, a technique which has been shown empirically in the generative modelling literature to improve perceptual reconstruction quality [54]. In addition, this also provides a mechanism to incorporate relevant scene-level cues contained in the second network (which could be *e.g.* a strong ImageNet classifier or a pre-trained language-vision embedding).

One of the benefits of MAE is that it can rapidly learn strong representations using only self-supervision from the images in the target pre-training set. To maintain this property, we introduce a second idea: tying the features not with a separate network, but with an adversarial discriminator trained in parallel to distinguish between real and generated images. Both ideas combined result in not only a lower reconstruction error, but also learnt representations which better capture details of the scene layout and object boundaries (see Figure 3) without either explicit supervision or the use of hand-engineered inductive biases.

Finally, we build on these results and show that techniques from the generative modelling literature such as multi-scale gradients [24] and adaptive discriminator augmentation [25] can lead to further improvements in the learnt representation, and this also translates into a further boost in performance across downstream tasks. We hypothesise that the issues that these methods were designed to overcome, such as mode collapse during adversarial training and incomplete learning of the underlying data distribution, are related to overfitting on low-level image features.

Our contributions can be summarized as follows: (i) we introduce a simple and self-contained technique to improve the representations learnt by masked image modelling based

on perceptual loss and adversarial learning, (ii) we perform a rigorous evaluation of this method and variants, and set new state-of-the-art for masked image modelling without additional data for classification on ImageNet-1K, object detection on MS COCO and semantic segmentation on ADE20K, (iii) we demonstrate this approach can further draw on powerful pre-trained models when available, resulting in a further boost in performance and (iv) we show our approach leads qualitatively to more 'object-centric' representations and stronger performance with frozen features (in the linear probe setting) compared to MAE.

2. Related Work

2.1. Masked Image Modelling

Self-supervised learning has led to learning systems that do not depend on data labelling, where the raw data itself provides the supervisory signal for training. This results in models with feature representations that are generalisable to many tasks. Self-supervised learning has shown considerable success in Natural Language Processing (NLP) [2, 33, 44], where random parts of the input text are masked and the model is tasked with predicting the invisible content. This has become the de facto method of pre-training NLP models. Compared to this direct-prediction approach, the first performant approaches to self-supervised visual representation learning instead used predefined discriminative tasks such as estimating distortions of the input image [13, 15], patch re-ordering [11, 37], re-coloring a grayscale image input [53], and contrastive learning [5, 42].

Recently, inspired by NLP and facilitated by the advent of Transformer models (Vision Transformers [30] in particular) masked image modelling (MIM) returns to the idea of direct-prediction, randomly masking pixels of an input image before predicting the invisible content. Early work included iGPT [4] which downsized images and then directly predicted unknown pixel values in an autoregressive manner. Recent methods have moved towards predicting full resolution patches in an autoencoder configuration [1, 17, 51]. BEiT [1] relies on an additional generative model (*dVAE* [38]) pre-trained on a large corpus of images (250M) with the pretext task to predict for masked matches the closest visual token from a pre-trained codebook. MAE [17] takes a simpler approach, demonstrating that direct pixel prediction of masked regions using Mean Squared Error is also effective when a very large proportion of the image is masked out.

These recent methods produce strong performance when fine-tuning over downstream tasks, but generally discriminative self-supervision has continued to be more performant in a linear probe setting [17] suggesting focusing on the learning of features of the right level of abstraction remains a challenge. We seek to address this in our work.

2.2. Perceptual Similarity

The aim of perceptual similarity [54] is to mimic human visual perception. Humans are capable of understanding images on an abstract level, relying on high level concepts and semantic cues that define the underlying relationships between different entities in the frame. Perceptual similarity aims to mimic this human-like judgement by defining metrics which encode *perceptual distance*, with this being higher for image representations that similarly better capture visual semantic concepts.

Structural Similarity Index (SSIM), an early form of perceptual loss, attempts to capture properties of an image that when varied are perceived by humans as substantially different. Images are compared on three key features: (i) luminance (*i.e.* pixel intensity), (ii) contrast and (iii) structure [48]. Further work extended this to compute similarity at multiple scales [49]. In parallel other similar metrics have been proposed such as: Peak Signal to Noise ratio (PSNR) [21], Feature Similarity Index (FSIM) [52] and HDR-VDP-2 [34].

An alternative approach to capturing perceptual similarity is to not compare differences between pixels but instead compute the differences between the intermediary features learnt by a neural network and those extracted from a second fixed network pre-trained in a supervised manner on a large dataset, on the basis that these capture the higher-level semantically meaningful features required for accurate classification. This is the approach taken by [22] who pre-train a VGG network on ImageNet and then use this for learning. Such a *feature matching* based approach has been successfully applied to many tasks since in computer vision [41, 46, 47].

In our work, we draw on the feature matching based approach to perceptual similarity but remove the requirement for a pre-trained network, instead learning perceptual similarity dynamically. In parallel to this work, PeCo [12] also experiments with perceptual loss to prepare a perceptually aware codebook for masked image modelling. However, in our case we apply this directly during pre-training which is much more effective, as it enables the encoder to learn directly higher-level cues from the second network.

2.3. Generative modelling

If perceptual similarity tries to capture the semantic structure of images, generative models aim to capture the underlying distribution of the image data. An example is Generative Adversarial Networks (GANs) [23–25, 27]. The samples created by a generator model are evaluated by a separately trained discriminator model which is tasked with determining real images from generated images. This is trained in parallel with the generator using an adversarial loss function. Since GANs learn the underlying data distribution implicitly via a discriminator the original for-

mulation [16] could produce high-fidelity images but suffered from training instability and mode collapse, where the network was only able to capture a subset of the variance present in the data distribution. Subsequent work including Pro-GAN [24] and MSG-GAN [23] introduced the idea of generation at multiple scales to stabilise the generator, enabling a more complete capture of the underlying data distribution.

The StyleGAN family of papers [25–27] introduced several improvements to the learning of the discriminator further designed to improve the stability of generated images. Perceptual path length regularisation [54] enforces that small changes in the input latent code (in our work: the input to the decoder) lead to changes of a similar magnitude in the feature maps of the discriminator, thus ensuring good normalisation of the input codes. Adaptive Discriminator Augmentation (ADA) [25] enables the use of heavy augmentation when training the discriminator, avoiding overfitting even with smaller volumes of training data, whilst ensuring these augmentations do not affect the output of the generator. Both encourage the underlying feature space to be more stable to small changes in low-level image statistics. In this work, we explore if these additions, along with multi-scale learning described above additionally incorporated into the masked autoencoder architecture, can therefore help to learn richer, high-level representations when using adversarial training for masked image modelling.

An alternative to GANs for generative modelling is provided by explicit generative models which aim to capture the underlying data distribution directly. These methods avoid some of the issues such as mode collapse suffered by implicit modelling but, given the need to model the full distribution, generally are more sensitive to the volume of data used for training. Examples include Variational Auto-Encoders (VAE) [20, 29], Flow-based models [10, 28] and Diffusion models [9, 36]. VQ-VAE [43] builds on VAE by learning a discrete latent space rather than a continuous one by the creation of a codebook. Recently this was applied for masked image modelling in BEiT [1], where the rich latent representations learnt by a VQ-VAE model pre-trained on large data [38] are used as a prediction target. In this work, we also explore using such a large pre-trained VQ-VAE model when combined with perceptual similarity loss.

3. Methodology

The learning framework used for this work is based on MAE [17]. In Section 3.1 we describe how the MAE loss is extended with a perceptual loss term. In Section 3.2 we then describe variants of adversarial loss which is also added to the objective. In Section 3.3 we describe modifications to the MAE architecture to maximize learning in the encoder stage when using multi-scale gradients.

3.1. MAE with Perceptual Loss

The pixel reconstruction loss from the original MAE formulation is extended to include a perceptual loss term:

$$L^G = ||G(I_m) - I||_1 + L^G_{perceptual} \tag{1}$$

Where G is the MAE model, I is the original image and I_m is the original image randomly masked. We follow the convention in the generative modelling literature, and use L1 loss rather than L2 loss for the reconstruction term.

MS-SSIM: Our baseline perceptual loss is based on structural similarity index, specifically the multi-scale variant (MS-SSIM) [49]. The multi-scale component aids in reducing artefacts formed around the edges of the output reconstructed image I'. The perceptual loss term is thus:

$$L^G_{ssim} = \frac{1}{N}\sum_{i,j} \alpha \frac{1 - SSIM(G(I_m)_{ij}, I_{ij})}{2} \tag{2}$$

Where i, j are the pixel indexes and N is the total number of scales, set to 4. We use a 3×3 block filter for each scale. α is a weighting constant, with the L1 error weighted by the inverse $(1 - \alpha)$.

Feature matching: Based on the feature and style reconstruction losses of [22] our second perceptual loss relies on a separate *loss network* with the decoder network encouraged to have similar feature representations as each corresponding layer of the loss network ϕ.

In the original formulation ϕ is a fixed VGG network pre-trained on ImageNet as described in Section 2.2. To avoid this dependency on an external pre-trained network, we instead introduce an additional discriminator network D which will act as our loss network ϕ. This is trained in an adversarial setup to distinguish between the reconstructed image from the decoder G and the original image prior to masking. The intuition is that the features learnt through this task also contain higher-order perceptual cues which can be used to guide training of the decoder.

The perceptual loss comprises two parts. The first transfers high-level semantics (individual features are similar), with a second style term added which learns overall image statistics (correlations between features across the image are similar) giving:

$$\begin{aligned} L^G_{feat} =& \delta_f \sum_{j=1}^{J} \frac{1}{N_j}[||\phi^j(G(I_m)) - \phi^j(I)||_1] + \\ & \delta_s \sum_{j=1}^{J} \frac{1}{N_j}[||\Psi(\phi^j(G(I_m))) - \Psi(\phi^j(I))||_1] \end{aligned} \tag{3}$$

Where j is the index of the layer, N_j denotes the number of elements in each layer, Ψ is the Gram matrix function [14] and δ_f and δ_s are constant weighting factors.

Additionally, the adversarial loss is added to L^G. Any adversarial loss function can be used, but for our baseline experiments we use *LS-GAN* [35] which has shown to achieve more stable optimisation over the original min-max classification loss. This gives us the generator-discriminator loss pair:

$$L^D_{adv} = \frac{1}{2}[(D(I) - 1)^2] + [(D(G(I_m))^2] \tag{4}$$

$$L^G_{adv} = \frac{1}{2}[(D(G(I_m)) - 1)^2] \tag{5}$$

The full loss function for the decoder then becomes:

$$L^G = ||G(I_m) - I||_1 + L^G_{feat} + L^G_{adv} \tag{6}$$

Beyond the feature matching acting as a learnt perceptual loss, adding this term also has the further advantage of stabilising adversarial training.

dVAE perceptual: To provide a perceptual learning baseline also using the stronger supervision from a pre-trained network for comparison, we experiment with feature matching loss with the discrete variational autoencoder (dVAE) from [38]. For this, the same feature matching loss in Equation 3 above is used with the pre-trained dVAE encoder model component acting as the loss network ϕ. This then is the perceptual loss term, giving the full loss function for the decoder:

$$L^G = ||G(I_m) - I||_1 + L^G_{feat} \tag{7}$$

The dVAE was trained for image tokenization on the DALL-E dataset, comprising 250 million images. Therefore, the rich features encoded in its weights provide strong higher-order perceptual cues for decoder training.

3.2. Adversarial Training Variants

For feature matching based perceptual learning, any adversarial loss function can be used. In addition to the LS-GAN loss used in our baseline model, we experimented with two further variants, introduced below. Both were formulated to address issues with the original GAN formulation such as training instability and mode collapse [23]. We hypothesize that the richer distributions learnt by these methods will provide stronger cues for perceptual learning.

MSG-GAN: To stabilise the training of the generator, MSG-GAN [23] allows for the flow of gradients from the discriminator to the generator at multiple scales. This is done by adding skip connections from intermediate layers of the generator to intermediate layers of the discriminator. The loss function for training D and G remains unchanged.

StyleGANv2-ADA: We take all modifications made to the discriminator in the StyleGANv2-ADA [25] paper. Building on MSG-GAN, perceptual path regularisation between the decoder input and discriminator feature maps is added. Adaptive discriminator augmentation is also applied to all samples during training. The loss function for training D and G remains unchanged.

3.3. Model Architecture

One issue with the multi-scale GAN formulation used for both *MSG-GAN* and *StyleGANv2-ADA* methods is that the multi-scale learning occurs via skip connections between the discriminator D and decoder G. This means that only the decoder benefits from the multi-scale gradient penalty during training, and this is removed post pre-training, leaving only the encoder when adapting to downstream tasks.

To distribute the learning more evenly between encoder and decoder, similar to U-Net [39] we additionally introduce skip connections between intermediate encoder and decoder layers, as shown in Figure 1. In this modified architecture, which we term **MSG-MAE**, multi-scale signal is shared also with the encoder. In further detail: mask tokens are added to the encoder feature maps of dimension d along with positional embeddings. Via the skip connections, this is concatenated with the decoder feature map of the matching scale, creating a sequence of length $2d$. This combined feature is passed through a single learnt linear layer to return the dimension to d. The output is then processed by the decoder transformer layer, with the result forward propagated to the next layer and simultaneously converted to an RGB image at the required scale for the multi-scale loss.

4. Implementation Details

Pre-training: We use the ViT-B and ViT-L architectures from the MAE paper [17], with ViT-B and ViT-L trained for 300 and 1600 epochs respectively over the ImageNet-1K (IN1K) [7] training set. In each case, the input patch size is fixed to 16x16 and we mask 75% of input patches during training. For both ViT-B and ViT-L, the decoder architecture remains consistent. The model dimensions, hyperparameters and data augmentation strategies follow those of the original MAE paper [17] and we train with a batch size of 16. The Adam optimizer is used with a weighted decay, where the learning rate is 0.00015, weight decay is 0.05 (cosine strategy), 40 warm-up epochs are used and the momentum parameters β_1 and β_2 are 0.9 and 0.95.

In our experiments, the weighting factors in L^G_{feat} (where applied) is given a weighting factor δ_f of 0.05. The L^G_{ssim} weighting factor α is set to 0.85. In both cases, the result is to focus learning on the perceptual term, with the smaller δ_s value still resulting in a large weighting once accounting for the larger relative magnitude of the L^G_{feat} term. The parameter choice is based on other works in the literature [32, 41, 46]. δ_s is given a fixed value of 40. In order to give time for the discriminator to learn new features with which to compute perceptual similarity, a training schedule

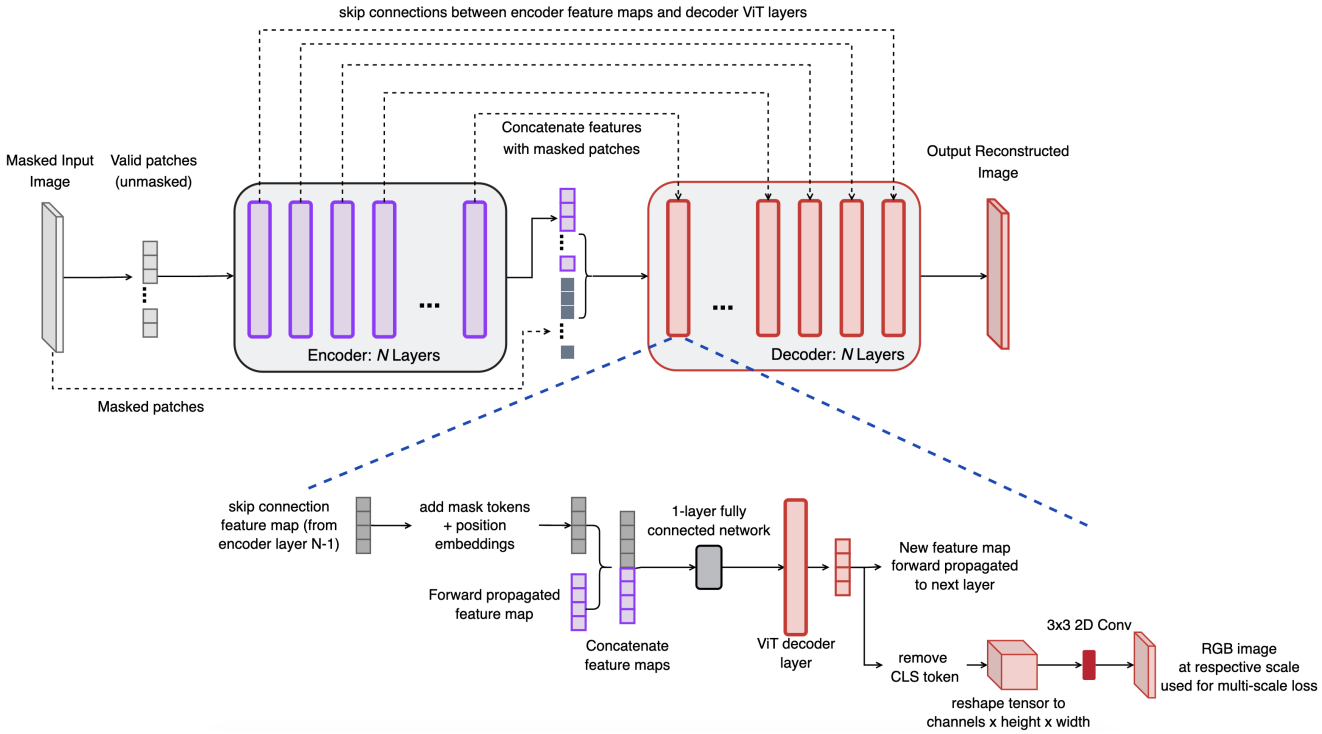

Figure 1. The Multi-Scale Gradient MAE (*MSG-MAE*) architecture. Dotted lines denote skip connections and solid forward propagation.

whereby the perceptual loss term L_G is applied only on even numbered epochs is used. This avoids the generator distribution collapsing to that of the discriminator and ensures well-balanced learning. All experiments were conducted on a GPU cluster consisting of 8xV100 Nvidia GPUs.

Fine-tuning: For fine-tuning, we take the pre-trained MAE encoder model and replace the decoder architecture with a task-specific head (initialised with random weights), similar to [17]. For image classification we replace the decoder with the original ViT model classification head [30]. For object detection and segmentation on MS-COCO [31] we use a Mask-RCNN [18] decoder model and for ADE-20K semantic segmentation [55] we adopt the UperNet model [50] as our decoder. When transferring to downstream tasks the intermediate feature maps of MSG-MAE are not used for the classification tasks. However, they are used for detection and segmentation, given the Mask-RCNN and UperNet architectures operate at multiple scales. All fine-tuned models are trained with an Adam optimizer with weighted decay. The learning rate is 0.001, weight decay is 0.05 (cosine strategy), warm-up epochs 5, and momentum parameters β_1 and β_2 are 0.9 and 0.95.

5. Experiments

In this section, we evaluate our models following self-supervised pre-training on the ImageNet-1K (IN1K) [7] training set. We first explore the main properties of the learnt representations in Section 5.1 in terms of (i) the fidelity of reconstructed output, (ii) the qualitative attention maps from the pre-trained model and (iii) linear probe results for downstream classification. Following this, in Section 5.2 we show the downstream performance of our models for transfer learning comparing this to previous work: fine-tuning on ImageNet-1K for classification, COCO for object detection and ADE20K for segmentation.

Table 1. Image reconstruction quality evaluation on ImageNet-1K. The ViT-B architecture is used. For columns with red headers, lower value is better and for the columns with green headers, higher value is better. The best result is highlighted in bold. Methods with † use the original MAE architecture, otherwise MSG-MAE is used.

Loss Function	L1	PSNR	SSIM	IS	FID
MSE †	0.25	0.38	0.76	6.33	42.7
MS-SSIM + L1 †	0.21	0.41	0.82	8.01	31.6
LS-GAN-P †	0.16	0.53	0.92	16.2	28.2
MSG-GAN-P	0.11	0.55	**0.94**	32.1	19.0
StyleGANv2-ADA-P	**0.06**	**0.58**	0.91	**36.8**	**10.3**

5.1. Main properties

Image reconstruction. In Table 1 we evaluate the reconstruction quality of the decoder stage of our models

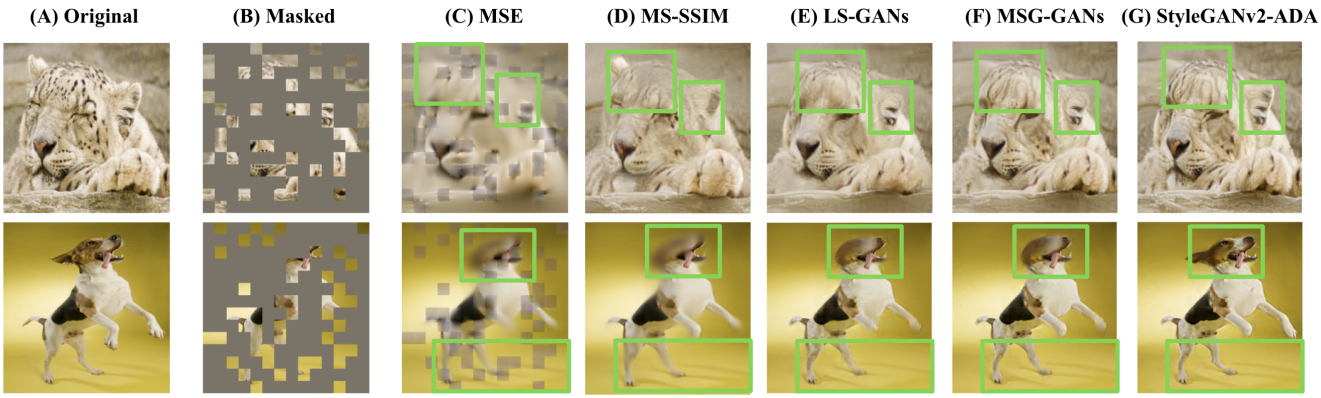

Figure 2. Differences in reconstruction quality of the model variants on samples from the ImageNet-1K validation set. The key areas of focus are highlighted in green. Columns (A-B) show the ground truth image and masking, (C-G) the reconstructed image for each method.

Table 2. **Classification performance on ImageNet1K** (IN1K). All models are pre-trained via self-supervision followed by either training of a linear probe or fine-tuning. Loss functions with '-P' include a perceptual loss term. The best result is highlighted in bold.

Method	Loss Function	Pre-training Data	Linear probe ViT-B	Fine-tuning ViT-B	Fine-tuning ViT-L
iGPT [4]	Cross Entropy	IN1K	–	66.5	–
DINO [3]	Cross Entropy	IN1K	**78.2**	82.8	–
MoCo v3 [6]	InfoNCE [42]	IN1K	76.7	83.2	84.1
BEiT [1]	Negative Log Likelihood	IN1K + DALL-E	56.7	83.2	85.2
MAE [17]	MSE	IN1K	67.8	83.6 [a]	85.9
MAE	MS-SSIM + L1	IN1K	71.2	84.1	86.3
MAE	LS-GAN-P	IN1K	72.5	84.5	86.5
MSG-MAE	MSG-GAN-P	IN1K	75.6	85.3 [b]	87.2
MSG-MAE	**StyleGANv2-ADA-P**	IN1K	**78.1**	**86.2**	**88.1**
MAE	dVAE-P	IN1K + DALL-E	**79.8**	**86.9**	**88.6**
MAE	*LS-GAN*	*IN1K*	–	*83.3*	*85.3*
MSG-MAE	*MSG-GAN*	*IN1K*	–	*84.7* [c]	*86.5*
MAE	*MSG-GAN-P*	*IN1K*	–	*83.2* [d]	*85.6*
MAE	*StyleGANv2-ADA-P*	*IN1K*	–	*84.5*	*86.2*

over the IN1K validation set using the following quantitative measures: L1 error, Peak Signal to Noise Ration (PSNR), Structural Similarity Index (SSIM) [49], Inception Score (IS) [40] and Fréchet inception distance (FID) [19]. These experiments use the ViT-B variant of the encoder, pre-trained using each of our perceptual losses.

For each of our methods, we observe a gradual increase in the fidelity of the reconstructed patches on the pixel-level measures (L1, PSNR, SSIM). However, what is particularly striking is the consistent boost of $+10\%$ for each method in FID score (with a similar pattern observed for IS). These methods compute a higher-level notion of perceptual similarity by comparing intermediary feature maps from a network pre-trained using a supervised objective for real and generated images, and suggests that through the introduction of a perceptual loss term the decoder learns a more generalisable notion of perceptual similarity. Examples of reconstructed patches are shown in Figure 2.

Self-attention maps. To evaluate qualitatively whether the features learnt by our approach properly capture the high-level semantics of the image over low-level details we visualise the attention maps from the final layer of our network, shown in Figure 3. Compared to the original MAE formulation, the combination of perceptual and ad-

Table 3. **Object detection and semantic segmentation performance on MS COCO and ADE20K.** All models were pre-trained using the ImageNet-1K training set (without labels). Loss functions with '-P' include a perceptual loss term. Best results are highlighted in bold.

Method	Loss Function	Pre-training Data	MS COCO mAP *Box*	MS COCO mAP *Mask*	ADE20K mIoU
DINO [3]	Cross Entropy	IN1K	–	–	44.1
MoCo v3 [6]	InfoNCE [42]	IN1K	47.9	42.7	47.3
BEiT [1]	Negative Log Likelihood	IN1K + DALL-E	49.8	44.4	47.1
MAE [17]	MSE	IN1K	50.3	44.9	48.1
MAE	MS-SSIM + L1	IN1K	50.8	45.1	48.8
MAE	LS-GAN-P	IN1K	51.4	45.4	49.2
MSG-MAE	MSG-GAN-P	IN1K	52.3	45.8	49.7 [b]
MSG-MAE	**StyleGANv2-ADA-P**	IN1K	**53.5**	**46.1**	**50.4**
MAE	dVAE-P	IN1K + DALL-E	**53.9**	**46.4**	**50.9**
MAE	*MSG-GAN-P*	*IN1K*	*50.9*	*45.9*	*49.3* [d]
MAE	*StyleGANv2-ADA-P*	*IN1K*	*51.8*	*45.5*	*49.1*

versarial loss (LS-GAN) leads to sharper focus on the object in the frame despite no supervision being used during training. The addition of multi-scale gradients (MSG-GAN) and adaptive discriminator augmentation and perceptual path length regularisation (StyleGANv2-ADA) brings further improvement.

In particular, with our best method, we achieve similar qualitative results to DINO [3] a self-supervised method which takes a contrastive approach to learning and requires careful balancing of the loss and sampling of image crops within batches compared to the simpler reconstruction-based approach used by our method.

Linear probing. To evaluate quantitatively the extent to which the features learnt by our approach capture useful semantic information, a common approach is to freeze the backbone of the pre-trained encoder model and train a simple linear classifier on top. We report the results for this over the IN1K validation set in Table 2, comparing to the MAE, BEiT and contrastive learning approach MoCo v3.

Our baseline model variant trained via MS-SSIM achieves 71.2% accuracy, 3% higher than the original MAE trained via MSE [17]. StyleGANv2-ADA-P attains $\mathbf{78.1}\%$, a boost of 10% compared to the original MAE. This significant increase suggests that our perceptual loss term leads to much more informative features being learnt without fine-tuning with labels being necessary. For comparison, we also include results when using perceptual loss computed instead against a pre-trained network. When adding this stronger supervision, the accuracy further improves to 79.8%, although this introduces a dependency on an external network and training data magnitudes larger than IN1K.

Table 4. **Computational cost.** The relative time to train per epoch. All figures computed with the ViT-B architecture using 8xV100s.

	MAE	LS-GAN-P	StyleGANv2-ADA-P
Rel. time / epoch	0.23	0.54	1
# Parameters	113M	113M	119M (+5%)

Computational properties. All experiments were run on 8xV100s, and took between 1-3 weeks to train to convergence for ViT-B. The relative training times and parameter counts are shown in Table 4. We also tried training MAE longer (*e.g.* 900 epochs instead of 300, to match the total time for StyleGANv2-ADA-P) and did not observe signficiant performance improvement.

5.2. Downstream Learning Results

Image classification. We fine-tune our models on IN1K, with the results shown in Table 2. Using ViT-B, we see a consistent boost in performance across the board by adding a perceptual loss term, obtaining an accuracy of 86.2% with our best method (StyleGANv2-ADA-P) and outperforming MAE and BEiT by 2.6% and 3% respectively.

Moving to ViT-L architecture, we obtain $\mathbf{88.1}\%$ accuracy, outperforming all previous methods training only on IN1K data and in particular resulting in comparable accuracy to that reported by MAE of 87.8% using the much larger ViT-H_{448} architecture (632M parameters vs 86M parameters, with input image of size 448 rather than 224).

If we use a pre-trained network with dVAE-P, we obtain a further boost of our best accuracy to 88.6%

Object detection and semantic segmentation. For object detection, we fine-tune a Mask R-CNN head on MS-

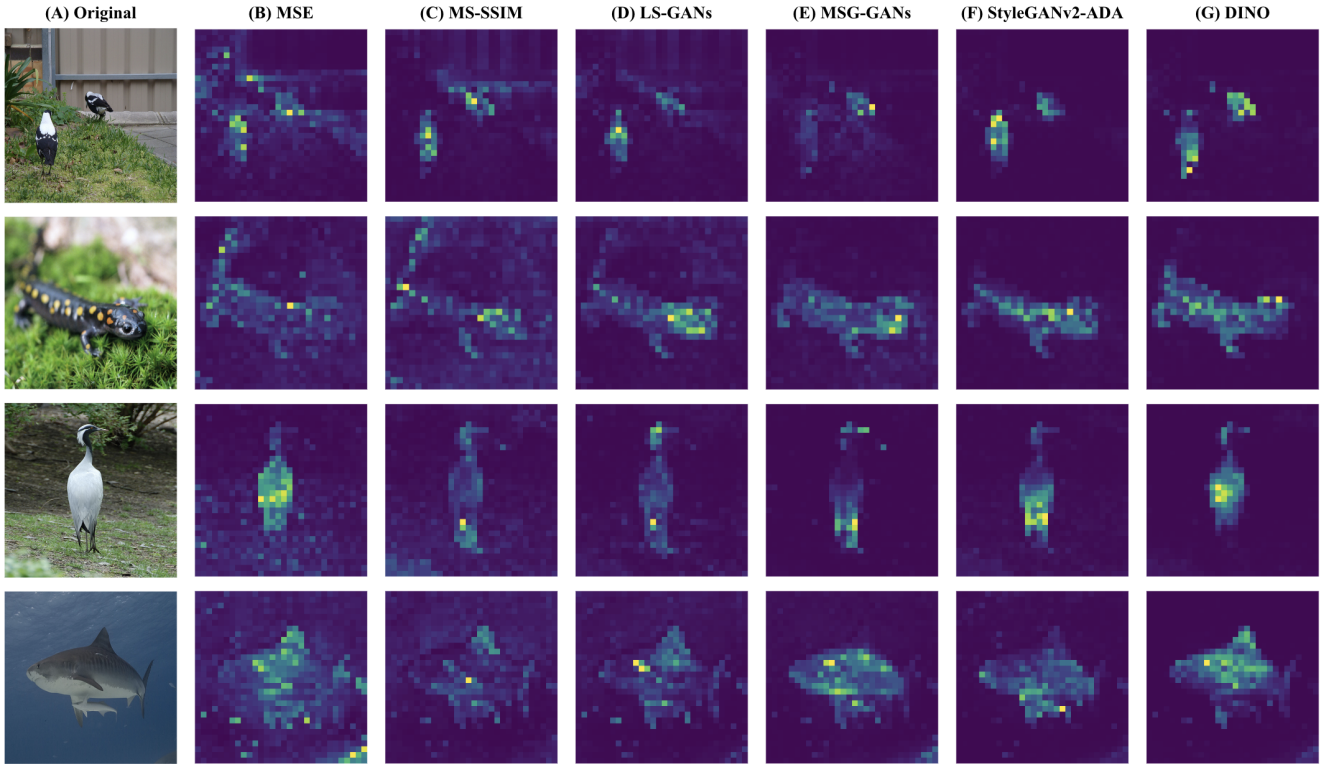

Figure 3. Attention maps of models pre-trained on ImageNet-1K without labels. We visualise the self-attention of the [CLS] token of the last layer. Sample images were selected randomly from the ImageNet-1K validation set. Column (A) is the original input image, (B-F) the outputs from our different losses and (G) is the output from DINO for comparison.

COCO with the results shown in Table 3. Our best method achieves 53.5 AP^{box} using a ViT-B architecture, which outperforms the previous best reported result of MAE trained with ViT-B by 3.2. Similarly, for semantic segmentation, we fine-tune an UperNet head on ADE20K with the results shown also in Table 3. Again using a ViT-B architecture, here we obtain up to 50.4 mIOU, 1.2 higher than for MAE.

Impact of perceptual loss term. When training a baseline LS-GAN without a perceptual loss term, we are unable to train a model that performs better than the baseline MSE loss [17] and observe clear stability issues during training. However, with an MSG-GAN loss training is much more stable. Referring to Table 2, for ViT-B with adversarial and reconstruction loss only we obtain a 1.1% boost (a *vs.* c) over the baseline MSE loss for image classification. However, this remains less than the 1.7% boost (a *vs.* b) with perceptual component added. An even larger gap is observed for ViT-L (0.6% *vs.* 1.3% boost). This suggests the perceptual loss term plays an important role not only for training stability but also is a large driver of performance.

Impact of multi-scale MAE. Training with a multi-scale loss without updating the MAE architecture as described in Section 3.3 results in a drop of performance of over 2% for image classification as seen in Table 2 (b *vs.* d), with a similar drop also observed for object detection and semantic segmentation in Table 3 (b *vs.* d).

6. Conclusion

We explored a method for incorporating the learning of higher-level features from images explicitly into the learning objective of a masked autoencoder. By introducing a perceptual loss term and adversarial training, we showed how the representations learnt by MAE [17] could be improved, boosting transfer performance for downstream tasks such as image classification, object detection and semantic segmentation. In particular, this performance boost is observed not only when fine-tuning, but also in the linear probe setting where contrastive methods have historically done better. This suggests that by combining the rich supervision of the pixel reconstruction task with a more focused higher-level learning signal we can greatly improve the data efficiency of the masked autoencoder approach.

This work also helps to start to address one of the key differences between the use of masked modelling for images and text: that images and image patches do not have inherent semantic meaning. Many questions remain about how to learn cues of the right level of abstraction directly from image data.

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, April 2022. 1, 2, 3, 6, 7

[2] Tom Brown, Benjamin Mann, Nick Ryder, and et. al. Subbiah. Language models are few-shot learners. volume 33, pages 1877–1901. Curran Associates, Inc., 2020. 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 6, 7

[4] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR, 13–18 Jul 2020. 1, 2, 6

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. 2

[6] Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021. 6, 7

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 4, 5

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

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. volume 34, pages 8780–8794. Curran Associates, Inc., 2021. 3

[10] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *ICLR*. OpenReview.net, 2017. 3

[11] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430, 2015. 2

[12] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. PeCo: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021. 2

[13] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. volume 27. Curran Associates, Inc., 2014. 2

[14] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423, 2016. 3

[15] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *ArXiv*, abs/1803.07728, 2018. 2

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. volume 27. Curran Associates, Inc., 2014. 3

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *CVPR*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2980–2988, 2017. 5

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. page 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. 6

[20] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017. 3

[21] Alain Horé and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *ICPR*, pages 2366–2369, 2010. 2

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 1, 2, 3

[23] Animesh Karnewar, Oliver Wang, and Raghu Sesha Iyengar. Msg-gan: Multi-scale gradient gan for stable image synthesis. In *CVPR*, 2020. 2, 3, 4

[24] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 1, 2, 3

[25] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. 2020. 1, 2, 3, 4

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4396–4405, 2019. 3

[27] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8107–8116, 2020. 2, 3

[28] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. volume 31. Curran Associates, Inc., 2018. 3

[29] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2014. 3

[30] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 2, 5

[31] 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*, Zürich, 2014. Oral. 5

[32] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *ECCV*, 2018. 4

[33] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. cite arxiv:1907.11692. 2

[34] Rafał Mantiuk, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graph.*, 30(4), jul 2011. 2

[35] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. pages 2813–2821, 2017. 4

[36] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. volume 162 of *Proceedings of Machine Learning Research*, pages 16784–16804. PMLR, 17–23 Jul 2022. 3

[37] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *ECCV*, pages 69–84, Cham, 2016. Springer International Publishing. 2

[38] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 18–24 Jul 2021. 2, 3, 4

[39] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]). 4

[40] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. volume 29. Curran Associates, Inc., 2016. 6

[41] Samyakh Tukra, Hani J. Marcus, and Stamatia Giannarou. See-through vision with unsupervised scene occlusion reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3779–3790, 2022. 2, 4

[42] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. 2, 6, 7

[43] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. volume 30. Curran Associates, Inc., 2017. 3

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. volume 30, 2017. 2

[45] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. 2008. 1

[46] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 2, 4

[47] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCV*, September 2018. 2

[48] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 2

[49] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, volume 2, pages 1398–1402 Vol.2, 2003. 2, 3, 6

[50] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*. Springer, 2018. 5

[51] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022. 2

[52] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011. 2

[53] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 2

[54] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1, 2, 3

[55] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 127(3):302–321, 2019. 5