Network Fusion for Content Creation with Conditional INNs

Robin Rombach* Patrick Esser* Björn Ommer
IWR, Heidelberg University, Germany

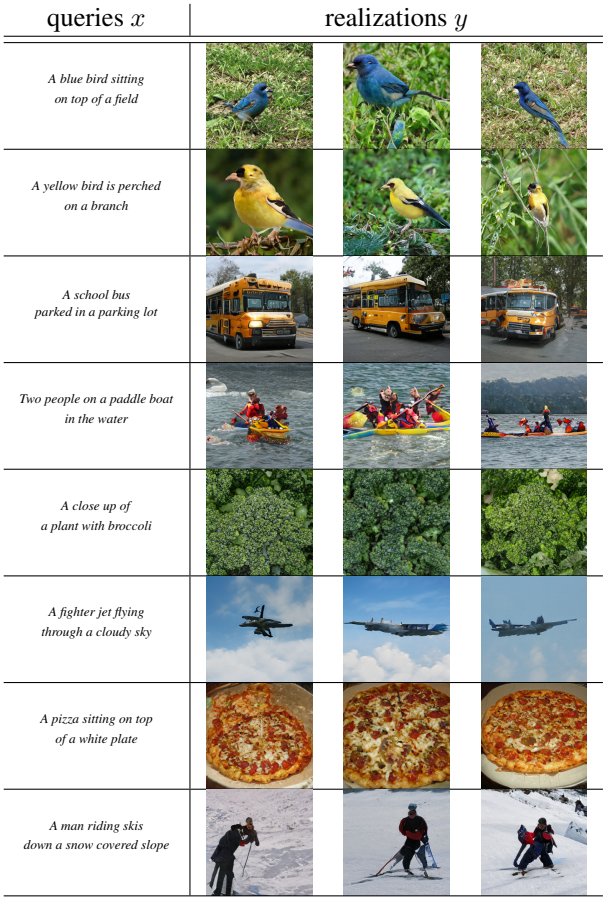
Abstract

Artificial Intelligence for Content Creation has the potential to reduce the amount of manual content creation work significantly. While automation of laborious work is welcome, it is only useful if it allows users to control aspects of the creative process when desired. Furthermore, widespread adoption of semi-automatic content creation depends on low barriers regarding the expertise, computational budget and time required to obtain results and experiment with new techniques. With state-of-the-art approaches relying on task-specific models, multi-GPU setups and weeks of training time, we must find ways to reuse and recombine them to meet these requirements. Instead of designing and training methods for controllable content creation from scratch, we thus present a method to repurpose powerful, existing models for new tasks, even though they have never been designed for them. We formulate this problem as a translation between expert models, which includes common content creation scenarios, such as text-to-image and image-to-image translation, as a special case. As this translation is ambiguous, we learn a generative model of hidden representations of one expert conditioned on hidden representations of the other expert. Working on the level of hidden representations makes optimal use of the computational effort that went into the training of the expert model to produce these efficient, low-dimensional representations. Experiments demonstrate that our approach can translate from BERT, a state-of-the-art expert for text, to BigGAN, a state-of-the-art expert for images, to enable text-to-image generation, which neither of the experts can perform on its own. Additional experiments show the wide applicability of our approach across different conditional image synthesis tasks and improvements over existing methods for image modifications.

1. Introduction

Neural Networks achieve superhuman performance in specific tasks [2, 45, 41, 1] but are far from being generally intelligent [52]. A state-of-the-art classifier might perfectly distinguish even slightest semantic variations in inputs, but will never be able to synthesize a description of its inner workings, unless explicitly trained to do so. Furthermore, such a model is typically well-performing on a certain domain (such as natural images), but cannot handle data from another domain, *e.g.* speech. These kind of problems can be summarized in the task of domain-to-

Table 1: *BERT* [8] to *BigGAN* [1] transfer: Our approach enables translation between expert models such as BERT and BigGAN. All samples are generated with the same transfer model. More results can be found on the project page at https://compvis.github.io/network-fusion/.



*Both authors contributed equally to this work.

domain translation and are subject to a large body of work [62, 63, 4]. A shortcoming of a lot of these approaches is their specificity: Specific loss functions and model classes are designed for specific problems, often resulting in non-transferable models. The breakthrough work of [28] introduced a general-purpose formulation based on conditional adversarial networks [39], which enabled supervised image-to-image translation without the need for a hand-crafted loss function. Following on from this method, we extend the task definition to arbitrary domain-to-domain translation in a paired fashion. To solve this problem, we utilize strong, pretrained individual networks, each an expert in its very specific task, and combine them by learning a translation between their hidden represensations with a conditionally invertible neural network (INN).

Using this aproach we are equipped with a single, general-purpose mechanism that enables generative fusion of arbitrary models, and by exploiting their individual capacities, yields a powerful task-transfer algorithm. Tab. 1 shows an example of such a network-to-network translation: Utilizing the transformer-based natural language model BERT [8] and a state-of-the-art GAN for ImageNet [7] generation, BigGAN [1], we train our approach to perform text-to-image translation. The figure shows a rich variety of generated examples, capturing both changes on a micro level (such as color in line 1-2) and the macro level (*e.g*. broccoli *vs*. school bus, l. 3&5).

Summarizing, our contributions are as follows: We (i) provide a general purpose approach that enables combination of arbitrary neural networks for multiple conditional image generation tasks trough a hidden bottleneck and maximum likelihood training, by (ii) learning a conditional generative model which models the distribution of realizations y corresponding to given input x, where x and y can be from different domains $\mathcal{D}_x$ and $\mathcal{D}_y$, and (iii) make transfer tasks on a rich variety of domains and datasets computationally affordable, since our method does not require any gradient computations w.r.t. the expert models.

2. Generative Models for Content Creation

The majority of approaches for deep-learning-based content creation rely on Variational Autoencoders (VAEs) [31, 50], Generative Adversarial Networks (GANs) [21], Autoregressive models [55], or normalizing flows [42] obtained with invertible neural networks (INNs) [9, 10]. These methods transform samples from a simple base distribution, mainly a standard normal or a uniform distribution, to a complex target distribution, *e.g*. the distribution of (a subset of) natural images. Sampling the base distribution then leads to the generation of novel content. Recent works [58, 16] also utilize INNs to transform the latent distribution of an autoencoder to the base distribution. A simple structure of the base distribution allows rudimentary control over the generative process in the form of vector arithmetic applied to samples [44, 48, 53, 20]. More generally, providing control over the generated content is formulated as conditional image synthesis.

In its most basic form, conditional image synthesis is achieved by generative models which, in addition to a sample from the base distribution, take class labels [39, 30] or attributes [24] into account. More complex conditioning information are considered in [62, 47], where textual descriptions provide more fine-grained control over the generative process. A wide range of approaches can be characterized as image-to-image translations where both the generated content and the conditioning information is given by images. Examples for conditioning images include grayscale images [64], low resolution images [32], edge images [28], segmentation maps [43, 3] or heatmaps of keypoints [17, 36, 14]. [28] introduced a common framework for image-to-image translation, which found widespread adoption among artists and designers. We argue that this success of [28] is caused by its unified treatment of image-to-image translation which allows artists to easily explore different ways to control the image synthesis without requiring deep-learning expertise. We take this unification one step further and provide a unified approach for a wide range of conditional content creation, including class labels, attributes, text and images as conditioning. In the case of image conditioning, our approach can be trained either with aligned image pairs as in [28, 3, 43] or with unaligned image pairs as in [66, 33, 5, 27, 15].

While many works on generative models focus on relatively simple datasets containing little variations, *e.g*. CelebA [35] containing only aligned images of faces, [1, 12] demonstrated the possibility to apply these models to large-scale datasets such as ImageNet [7]. However, such experiments require a computational effort which is typically far out of reach for individuals. Moreover, the need to retrain large models for experimentation prohibits rapid prototyping of new ideas for content creation. Making use of pre-trained neural networks can significantly reduce the computational budget and training time. For discriminative tasks, the ability to effectively reuse pre-trained neural networks has long been recognized [46, 11, 60]. For generative tasks, however, there are less works that aim to reuse pre-trained networks efficiently. Features obtained from pre-trained classifier networks are used to derive style and content losses for style transfer algorithms [18], and they have been demonstrated to measure perceptual similarity between images significantly better than pixelwise distances [37, 65]. [61, 38] find images which maximally activate neurons of pre-trained networks and [51] shows that improved synthesis results are obtained with adversarially robust classifiers. Instead of directly searching over images, [40] uses a pre-trained generator network of [13],

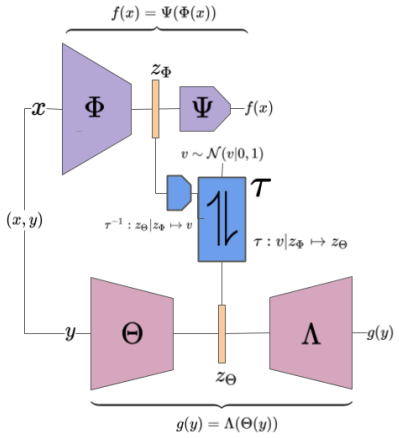


Figure 1: Proposed architecture. We provide post-hoc model fusion for two given deep networks $f = \Phi \circ \Psi$ and $g = \Theta \circ \Lambda$. For a deep representation $z_\Phi = \Phi(x)$ coming from an arbitrary layer, a conditional INN τ recovers the invariances v of the model Φ based on a representation z_Θ which contains *both* z_Φ and v in a generative fashion.

where it was used to reconstruct images from feature representations. However, these approaches are limited to neuron activation problems, rely on per-example optimization problems, which makes synthesis slow, and do not take into account the probabilistic nature of the conditional synthesis task, where a single conditioning corresponds to multiple outputs. In contrast, our approach efficiently utilizes pre-trained models for image-synthesis as well as for conditioning, such that their combination provides new generative capabilities for content creation through conditional sampling, without requiring the pre-trained models to be aware of these emerging capabilities.

3. Approach

Our goal is to learn a mapping between two domains $\mathcal{D}_x$ and $\mathcal{D}_y$: Given a *query* x, we aim to find a translation between $x \in \mathcal{D}_x$ and corresponding *realizations* $y \in \mathcal{D}_y$. More precisely, in our work $\mathcal{D}_x$ can contain a variety of entities such as textual descriptions, attributes, edge-images, segmentation maps or corrupted images, whereas $\mathcal{D}_y$ always contains natural images. This mapping is inherently multi-modal: As an example, consider mapping a query x from the domain of natural language to realizations y from the domain of natural images. Such visualizations typically show a rich variety in semantics, see Fig. 2 for an illustration. There exists a large body of fairly recent work covering the task of domain-to-domain translation, see Sec. 2 for details. Most of these methods, however, are highly specialized, domain-specific and have huge computational demands. A general-purpose algorithm for arbitrary domain-to-domain translation, such as Pix2Pix [28] or CycleGAN [67] for image-to-image translation, with low computational costs is currently missing in the literature.

The key insight of our work is that we can solve this task by making use of so-called *expert* models, which may achieve state-of-the-art performance on their respective domain, but are simultaneously restricted to this domain: For example, we aim to combine a transformer-based *language model* with a state-of-the art GAN for *image generation*. In a nutshell, we solve this problem by coupling such expert models via their *hidden representations* z.

To this end, let there be a joint distribution $p(x, y)$ from which queries x and realizations y can be sampled. Furthermore, let f denote the expert model acting on $\mathcal{D}_x$, while g denotes the model on $\mathcal{D}_y$. As we consider $\mathcal{D}_y$ to hold natural images, any well-performing model such as a generative adversarial network (GAN, [22]) or an autoencoder (AE) can be used to represent g. Because g does in general not know anything about $x \in \mathcal{D}_x$, g is *reusable* and can be coupled to various models f, which live on various query domains $\mathcal{D}_x$.

Additionally, we assume that we have access to query-realization pairs (x, y), where y is drawn from the distribution $p(y|x) = p(x, y)/p(x)$ given a query x. Our overall goal can then be expressed as learning an approximation $q(y|x)$ sucht that $q(y|x) \approx p(y|x)$. To do so, let us define that both expert models can be expressed as a composition of two functions $f(x) = \Psi(\Phi(x))$ and $g(y) = \Lambda(\Theta(y))$, such that we combine the models by learning a transformation τ that translates between their hidden representations $z_\Phi = \Phi(x)$ and $z_\Theta = \Theta(y)$.

3.1. Learning the translation $x \to y$

Using the above formulation, a high-level translation pipeline at inference time can be expressed as follows: Given a query x, we produce its latent embedding z_Φ, use τ to translate it to another model's hidden space z_Θ and finally use Λ to decode the translated representation into $\mathcal{D}_y$. We solve this task by learning a suitable transformation τ.

Figure 2: Sampling multiple *realizations* y given a *query* x. The given example corresponds to text-to-image creation.

Invariances of f enable control of content creation The above formulation contains a difficult challenge: The mapping from z_Φ to z_Θ is multi-modal, as (I) usually, multiple realizations y correspond to a single query x and (II) successful neural networks f learn invariances w.r.t. the input x. An example for the latter is a face recognition model, which, if trained successfully, should be invariant to pose, lighting, ... of an input image x. We thus have to approximate the distribution $p(z_\Theta|z_\Phi)$, such that sampling and decoding $z_\Theta \sim p(z_\Theta|z_\Phi) = p(z_\Theta|\Phi(x))$ through Λ enables *creation* of realizations of input queries x.

Learning a translation between model representations: To cover the invariances induced by both (I) and (II), we need a representation v of the remaining variance, such that, taken together, z_Φ and v uniquely determine z_Θ, *i.e.* there is a mapping τ s.t. $z_\Theta = \tau(v, z_\Phi)$, where sampling the invariances can be described by sampling v. Note that τ induces a distribution π, but for arbitrary τ, sampling from this induced distribution is just as hard as sampling z_Θ directly. However, there exist τ such that the induced distribution has the following nice properties: $v \sim \pi(v)$ is independent of z_Φ, π is easy to sample from and interpolations of samples are valid samples. One instantiation of such a distribution is given by a Gaussian distribution. We are thus looking for a τ such that the induced distribution is a (multivariate) normal distribution $\mathcal{N}(0, \mathbb{1})$.

We implement τ as a conditional invertible neural network (INN), such that by a change of variables

$$p(z_\Theta|z_\Phi) = \frac{p(v|z_\Phi)}{|\det \nabla(\tau(v|z_\Phi))|} \quad \text{where } v = \tau^{-1}(z_\Theta|z_\Phi). \tag{1}$$

Here, the denominator denotes the absolute value of the determinant of the Jacobian $\nabla(\tau)$ of $v \mapsto \tau(v|z_\Phi) = z_\Theta$, which can be effieciently computed for suitable invertible architectures.

By Eq. (1), $p(z_\Theta|z_\Phi)$ is expressed by means of the distribution $p(v|z_\Phi)$ of invariances, given a model's f representation $z_\Phi = f(x)$. As described above, the distribution $p(v|z_\Phi)$ is induced by τ: Thus, we identify $p(v|z_\Phi) = \pi(v) = \mathcal{N}(0, \mathbb{1})$. Note that we can assume such a simple Gaussian prior, as a powerful transformation τ can transform between two arbitrary densities. Given this prior, our task is then to learn the transformation τ that maps $\mathcal{N}(v|0, \mathbb{1})$ onto $p(z_\Theta|z_\Phi)$. To this end, we maximize the log-likelihood of z_Φ given z_Θ, obtained via paired training inputs $z_\Phi = \Phi(x)$ and $z_\Theta = \Theta(y)$, resulting in a per-example loss of

$$\begin{aligned} \ell(z_\Phi, z_\Theta) &= -\log p(z_\Theta|z_\Phi) \\ &= -\log \mathcal{N}(\tau^{-1}(z_\Theta|z_\Phi)) \\ &\quad -\log|\det \nabla\tau^{-1}(z_\Theta|z_\Phi)|. \end{aligned} \tag{2}$$

Minimizing this loss over the training data distribution $p(x, y)$ gives τ, a bijective mapping between (z_Φ, v) and z_Θ:

$$\begin{aligned} \mathcal{L}(\tau) &= \mathbb{E}_{x,y\sim p(x,y)}\left[\ell(\Phi(x), \Theta(y))\right] &(3)\\ &= \mathbb{E}_{x,y\sim p(x,y)}\left[1/2\|\tau^{-1}(\Theta(y)|\Phi(x))\|^2 + N_\Theta \log 2\pi \right. \\ &\quad \left. -\log|\det \nabla\tau^{-1}(\Theta(y)|\Phi(x))|\right] &(4) \end{aligned}$$

Note that both Φ and Θ remain fixed during minimization of $\mathcal{L}$.

Stacking the models Consequently, at inference time, we obtain translated samples z_Θ for given z_Φ by sampling from the invariant space v given z_Φ and then applying τ,

$$z_\Theta \sim p(z_\Theta|z_\Phi) \quad \Longleftrightarrow \quad v \sim \pi(v),\ z_\Theta = \tau(v|z_\Phi). \tag{5}$$

After training, translation between $\mathcal{D}_x$ and $\mathcal{D}_y$ is thus achieved by the following steps: (i) sample (x, y) from $p(x, y)$, (ii) encode x into the latent space $z_\Phi = \Phi(x)$ of expert model f, (iii) sample invariances v from the prior $\mathcal{N}(0, \mathbb{1})$, (iv) conditionally transform $z_\Theta = \tau(v|z_\Phi)$ and (v) decode z_Θ into the domain $\mathcal{D}_y$ of the second expert model: $y = \Lambda(z_\Theta)$. Note that this approach has multiple advantages: (i) hidden representations usually have lower dimensionality than x, which makes transfer between arbitrary complex domains affordable, (ii) the conditional INN τ can be be trained by minimizing the negative log-likelihood, independent of the domains $\mathcal{D}_x$ and $\mathcal{D}_y$, (iii) the approach does not require to take any gradients w.r.t. the expert models f and g, thus allowing post-hoc fusion of arbitrarily large models of interest, given that they obey some information bottleneck.

3.2. Building the INN τ

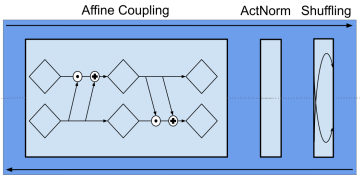


Figure 3: A single invertible block used to build our INN.

In our implementation, the conditionally invertible network τ is build from n blocks, each consisting of three invertible layers: affine coupling blocks [10], actnorm layers [29] and shuffling layers, which permute components of an input vector z in a fixed but randomly initialized manner, increasing overall expressivity of τ by mixing components for consecutive coupling layers. One invertible block is build from a sequence of these layers, *c.f.* Fig. 3.

4. Experiments

We investigate the wide applicability of our approach by performing experiments with multiple domains, datasets and models: (1) text-to-image translation by combination of BigGAN and BERT, (2) exploration of the use of combining standard ResNet-50 classifiers with BigGAN for image-to-image translation, (3) re-usability of a single generator for multiple translation taks and (4) comparison to existing methods for image modification.

Data requirements As our method does not require to compute gradients w.r.t. the models f and g, training can be conducted on a single GPU with about 10 GB VRAM.

4.1. Translation to BigGAN

This section is dedicated to the task of using a popular expert model as an *image generator*: BigGAN [1], achieving state-of-the-art FID scores on the ImageNet dataset. As most GAN frameworks in general and BigGAN in particular do not include an encoder, we aim to provide an encoding from an arbitrary domain by using an appropriate expert model f. Given the hidden representation $z_\Phi = \Phi(x)$, we aim to find a mapping between z_Φ and the latent space z_Θ of BigGAN's generator Λ. Thus, we identify $\Theta \equiv \mathbb{1}$ and $g = \Lambda$.

Here, z_Θ is the stacked vector

$$z_\Theta = [\tilde{z}, Wc]\,, \tag{6}$$

consisting of $\tilde{z} \sim \mathcal{N}(0, \mathbb{1})$, $\tilde{z} \in \mathbb{R}^{140}$, sampled from a multivariate normal distribution and $c \in \{0,1\}^K$, a one-hot vector specifying an ImageNet class ($K = 1000$ classes in total). The matrix W, a part of the generator Λ, maps the one-hot vector c to $h \in \mathbb{R}^{128}$, *i.e.* $h = Wc$. As c contains discrete labels, we have to avoid collapse of τ onto a single dimension of h during training. To this end, we pass the vector h through a small, fully connected variational autoencoder and replace h by its stochastic reconstruction, which effectively performs some kind of dequantization. Training of τ is then conducted by sampling z_Θ as described in Eq. (6) and minimizing the objective described in Eq. (4), *i.e.* finding a mapping τ that maps z_Θ to f's representations $z_\Phi = \Phi(x)$ and their corresponding invariances $v \sim \mathcal{N}(0, \mathbb{1})$.

The following sections present experiments in which the above approach is used to create novel content with a model-to-model transfer based on our conditional INN τ.

4.1.1 BERT-to-BigGAN translation

The emergence of transformer-based networks [56] has led to an immense leap in the field of natural language processing. One of the most widely used models is the so-called BERT (Bidirectional Encoder Representations from Transformers) model, an unsupervised model for learning language representations. Here, we make use of a variant of the original model, which modifies BERT such that it produces a latent space in which input sentences can be compared for similarity via the cosine-distance measure [49]. Thus, we train our model τ to map from these language representations $z_\Phi = \Phi(x)$ into the latent space z_Θ of BigGAN's generator, as described above. [1] During training, access to textual descriptions is obtained by using a captioning model as in [59], trained on the COCO [34] dataset. In a nutshell, at training time, we sample z_Θ as in Eq. (6), produce a corresponding image $\Lambda(z_\Theta)$), utilize [59] to produce a text-caption x describing the image and subsequently produce a sentence representation $z_\Phi = \Phi(x)$ which we use to minimize the overall objective Eq. (4).

Results can be found in Tab. 1. Our model captures both fine-grained and coarse descriptions and is able to synthesize images with highly different content, based on given textual queries x. We emphasize that all results from Tab. 1 are obtained with the transfer model τ, which shows the usefulness of combining highly specialized expert models for translation between their respective domains.

4.1.2 ResNet-to-BigGAN translation

Here, we train the INN τ conditioned on hidden representations of ResNet-50 from the penultimate layer (*i.e.* returned before being passed through the final classification layer) to show that standard classifiers, if trained in a suitable manner, can be employed for the task of domain-transfer. Reffering to Fig. 1, this means that f is represented by a ResNet classifier, whereas g is a BigGAN generator as already described.

To explore the utility of combining classifiers with GANs, we compare training with two ResNet-50 models with the same architecture, but trained with different training procedures. The first model is a vanilla ImageNet classifier, trained to perform class prediction on the ImageNet dataset. The second model, however, is trained on a *stylized* version of ImageNet. This is inspired by the work of [19], who showed that typical convolutional neural classification networks are biased towards texture when being trained on ImageNet. They proposed that this bias can be removed by training the CNNs on a stylized version of ImageNet instead, utilizing a simple neural AdaIN transfer algorithm [26] for stylization.

Examples conditioned on the latent representations of both a ResNet-50 trained on the stylized version of ImageNet and a ResNet-50 trained on standard ImageNet are displayed in Tab. 2. The results implicitly confirm the

[1] Note that we condition on the output of BERT, hence: $\Phi = f$ and $\Psi = \mathbb{1}$.

texture-bias hypothesis of [19]: Vanilla CNN-based classifiers are biased towards texture and do not classify input images based on their shape (as most humans would). Training the same CNN on a stylized version of the same dataset removes this bias. The figure demonstrates that such a classifier may be adopted for sketch-to-image translation and content creation, but success of adaption onto this task depends on the intrinsic properties of the classifier.

4.2. Reusing a single generator for different query domains

We evaluate the ability of our approach to combine a single autoencoder with different experts to solve a variety of image-to-image translation tasks.

We combine all carnivorous animal classes in ImageNet with images of the Animals with Attributes 2 dataset [57] and split the resulting *Animals* dataset into 211306 training images and 10000 testing images. For the autoencoder, we use a ResNet-101 [23] architecture as encoder, and the BigGAN architecture as the decoder. As we do not use class information, we feed the latent code z_Θ of the encoder also into a a fully-connected layer and use its softmax-activated output as a replacement for the one-hot class vector used in BigGAN. The encoder predicts mean $\Theta(y)_\mu$ and diagonal covariance $\Theta(y)_{\sigma^2}$ of a Gaussian distribution and we use the reparameterization trick to obtain samples $z_\Theta = \Theta(y)_\mu + \mathrm{diag}(\Theta(y)_{\sigma^2})\epsilon$ of the latent code, where $\epsilon \sim \mathcal{N}(\epsilon|0, \mathbb{1})$. For the reconstruction loss, we use a perceptual loss based on features of a pretrained VGG-16 network [54] for the reconstruction loss, and, following [6], include a learnable, scalar output variance γ. We use a PatchGAN discriminator [28] for improved image quality.

In Tab. 3, we consider the effects of fusing this autoencoder with different experts f using our conditional INN τ. In Tab. 3a, f is a segmentation network trained on COCOStuff, and $\Phi = f$, *i.e.* z_Φ is given by the final segmentation output of the network. This case corresponds to a translation from segmentation masks to images and we observe that our approach can successfully fuse the segmentation model with the autoencoder to obtain a wide variety of generated image samples corresponding to a given segmentation mask. Tab. 3b uses the same segmentation network for f, but this time, Φ consists of the logit predictions of the network (visualized by a random projection to RGB values). The diversity of generated samples is greatly reduced compared to Tab. 3a, which indicates that logits still contain a lot of information which are not strictly required for segmentation, *e.g.* the color of animals. This shows how different layers of an expert can be selected to obtain more control over the synthesis process.

In Tab. 3c, we consider the task of translating edge images to natural images. Here, x is obtained through the Sobel filter, and, based on insights of the previous section, we choose a ResNet pretrained for image classification on stylized ImageNet as a domain expert for edge images, as it has shown sensitivity to shapes. This combination again solves the translation task. Tab. 3d shows an image inpainting task, where x is a masked image. In this case, large portions of the shape are missing from the image but the unmasked regions contain texture patches. This makes a ResNet pretrained for image classification on ImageNet a suitable domain expert due to its texture bias. The samples demonstrate that textures are indeed faithfully preserved.

Note that all results in Tab. 3 were obtained by combining a single, generic autoencoder g, which has no conditioning capabilities on its own, and different domain experts f, which possess no generative capabilities at all. These results demonstrate the feasibility of solving a wide-range of image-to-image tasks through the fusion of pre-existing, task-agnostic experts on the domains $\mathcal{D}_x, \mathcal{D}_y$. Moreover, choosing different layers of the expert f provides additional, fine-grained control over the generation process.

4.3. Comparing image modification capabilities

To compare our approach to task-specific approaches, we compare its ability for attribute modification on face images to those of [4]. We train the same autoencoder as in the previous section on CelebA [35], and directly use attribute vectors for z_Φ. For an input y with attributes z_Φ, we synthesize versions with modified attributes z_Φ^*. In each column of Tab. 4.3, we flip the binary entry of the corresponding attribute to obtain z_Φ^*. To obtain the modified image, we first compute $z_\Theta = \Theta(y)$ and use its original attribute vector to obtain its attribute invariant representation $v = \tau^{-1}(z_\Theta|z_\Phi)$. We then mix it again with the modified attribute vector to obtain $z_\Theta^* = \tau(v|z_\Phi^*)$, which can be readily decoded to the modified image $y^* = \Lambda(z_\Theta^*)$.

Qualitative results in Tab. 4a demonstrate successful modification of attributes. In comparison to [4], our approach produces more coherent changes, *e.g.* changes in gender cause changes in hair length and changes in the beard attribute have no effect on female faces. This demonstrates the advantage of fusing attribute information on a low-dimensional representation of a generic autoencoder. Overall, our approach produces images of higher quality, as demonstrated by the FID scores [25] in Tab. 4b. Note that FID-scores are calculated w.r.t. the complete dataset, explaining the high FID scores for attribute *glasses*, where images consistently possess a large black area.

Additionally, Tab. 5 demonstrates results of our approach applied to unpaired image-to-image translation. Here, we use the same setup as for attribute modification, but train on a dataset containing images from the CelebA dataset, associated with a *human* attribute, and the Animal Faces-HQ dataset [5], associated with an *animal* attribute.

Table 2: Sketch-to-image transfer by combining variants of *ResNet-50* and *BigGAN*. Using a texture-agnostic classifier network (*left*), images can be created by coupling to the generator of BigGAN. This is not possible with a standard classifier, due to its bias towards texture (*right*).

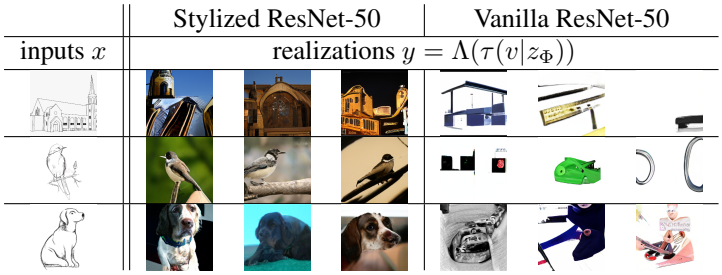
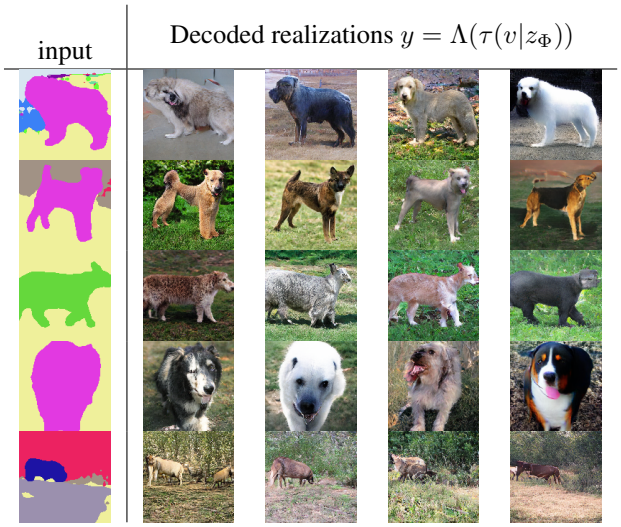
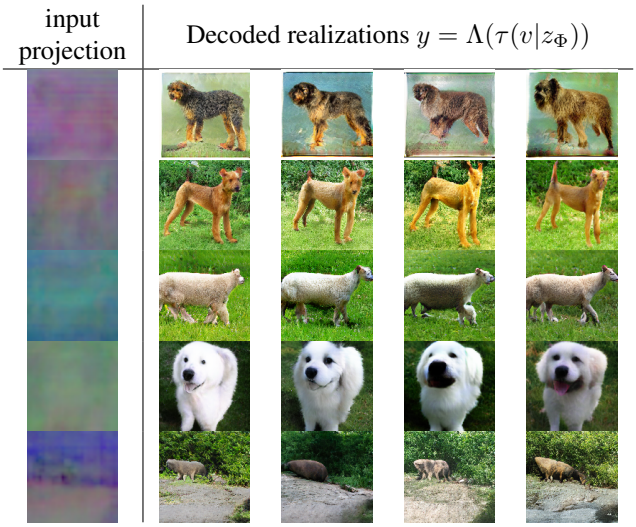


Table 3: Different Image-to-Image translation tasks solved with a single Autoencoder fused with different experts.

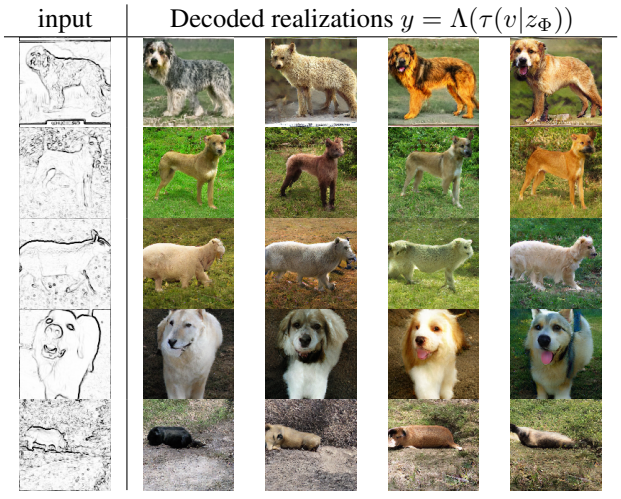
(a) Segmentation-to-Image transfer; *argmaxed logits* of expert.



(b) Segmentation-to-Image transfer; *logits* of segmentation expert.



(c) Edge-to-Image transfer using stylized ResNet classifier.



(d) Inpainting using vanilla ResNet classifier.

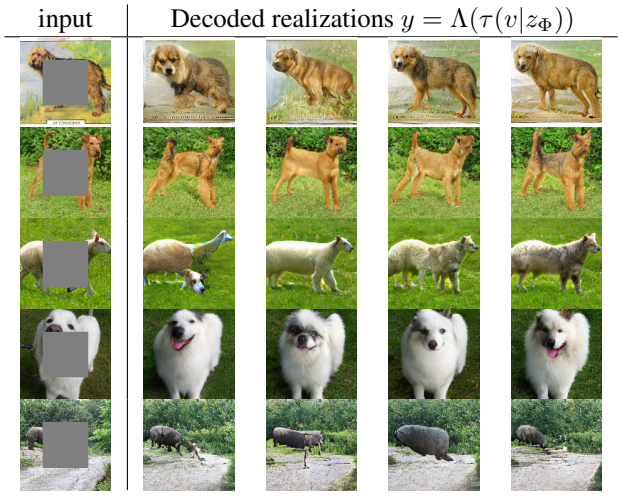
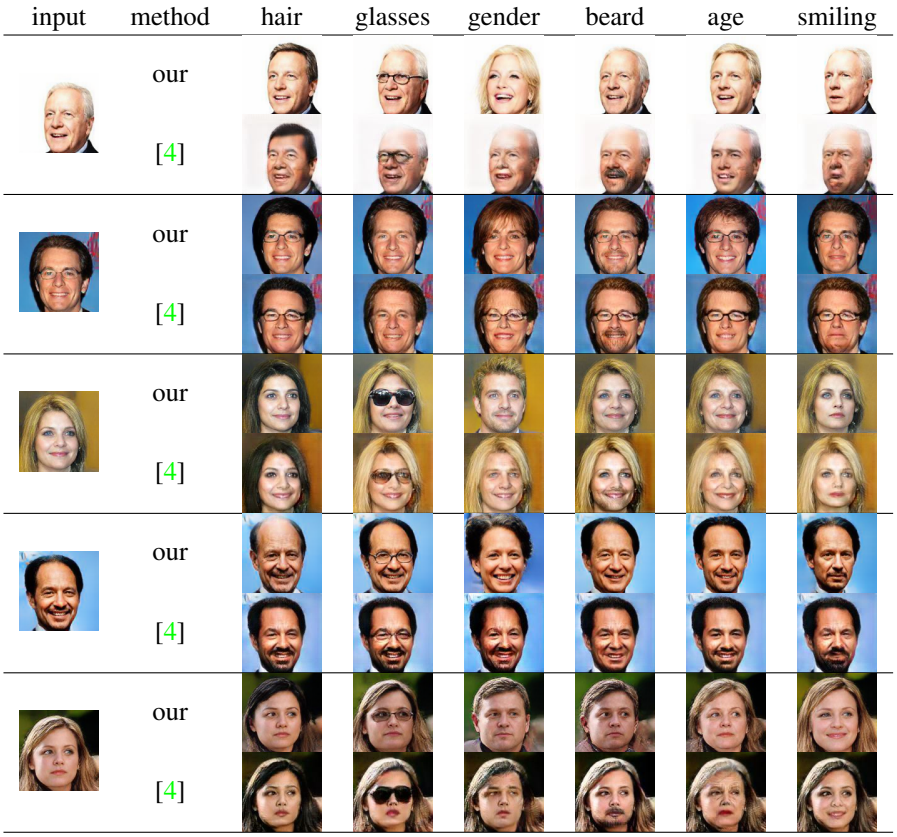


Table 4: Attribute Modification on CelebA.

(a) Qualitative results. Each column modifies a single attribute of the input.



(b) FID scores after modification of single attributes.

method	hair	glasses	gender	beard	age	smiling
our	**15.18**	**37.32**	**16.38**	**12.02**	**10.77**	**9.57**
[4]	20.94	41.27	20.04	19.88	21.77	14.47

5. Conclusion

We presented a new, unified approach to content creation through conditional image synthesis, based on a translation of representations obtained from pre-trained expert models. Our approach combines multiple desirable features, as it is (i) *affordable*: Individuals such as artists or scientists can utilize powerful, pretrained models such as BERT and BigGAN for new tasks, with just a single GPU instead of the full multi-GPU resources required for training such models from scratch; (ii) *flexible*: The objective is independent of translation domains $\mathcal{D}_x$ and $\mathcal{D}_y$. Training is always achieved by the maximum-likelihood principle which provides plug-and-play capabilities for new domains and experts to encourage creative applications; (iii): *powerful:* Using pretrained expert networks outsources the task of domain specific compression and understanding to these models. The INN can thus focus on the translation alone which leads to improvements over previous approaches. Interesting future applications include transfer between domains such as speech, music or brain signals.

Table 5: Swapping attributes *human* and *animal*

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1, 2, 5

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1

[3] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520, 2017. 2

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 6, 8

[5] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains, 2019. 2, 6

[6] Bin Dai and David Wipf. Diagnosing and enhancing vae models, 2019. 6

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 2

[9] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation, 2014. 2

[10] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp, 2016. 2, 4

[11] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition, 2013. 2

[12] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning, 2019. 2

[13] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks, 2016. 2

[14] Patrick Esser, Johannes Haux, Timo Milbich, and Björn Ommer. Towards learning a realistic rendering of human behavior. In *ECCV Workshops*, 2018. 2

[15] Patrick Esser, Johannes Haux, and Björn Ommer. Unsupervised robust disentangling of latent characteristics for image synthesis. In *Proceedings of the Intl. Conf. on Computer Vision (ICCV)*, 2019. 2

[16] Patrick Esser, Robin Rombach, and Björn Ommer. A disentangling invertible interpretation network for explaining latent representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[17] Patrick Esser and Ekaterina Sutter. A variational u-net for conditional appearance and shape generation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. 2

[18] Leon Gatys, Alexander Ecker, and Matthias Bethge. A neural algorithm of artistic style. *Journal of Vision*, 16(12):326, Sep 2016. 2

[19] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 5, 6

[20] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. *arXiv preprint arXiv:1906.10112*, 2019. 2

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. 6

[24] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019. 2

[25] 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, 2017. 6

[26] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 5

[27] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. *Lecture Notes in Computer Science*, page 179–196, 2018. 2

[28] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2, 3, 6

[29] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10215–10224, 2018. 4

[30] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014. 2

[31] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[32] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 2

[33] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. *Lecture Notes in Computer Science*, page 36–52, 2018. 2

[34] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[35] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 2, 6

[36] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *The IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[37] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015. 2

[38] Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120(3):233–255, 2016. 2

[39] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014. 2

[40] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks, 2016. 2

[41] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016. 1

[42] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference, 2019. 2

[43] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[44] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2015. 2

[45] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019. 1

[46] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: An astounding baseline for recognition. *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Jun 2014. 2

[47] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016. 2

[48] Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1252–1260. Curran Associates, Inc., 2015. 2

[49] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. 5

[50] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on International Conference on Machine Learning-Volume 32*, pages II–1278. JMLR. org, 2014. 2

[51] Shibani Santurkar, Dimitris Tsipras, Brandon Tran, Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Image synthesis with a single (robust) classifier, 2019. 2

[52] John R Searle. Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424, 1980. 1

[53] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. *arXiv preprint arXiv:1907.10786*, 2019. 2

[54] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[55] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems*, pages 4790–4798, 2016. 2

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 5

[57] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018. 6

[58] Zhisheng Xiao, Qing Yan, and Yali Amit. Generative latent flow, 2019. 2

[59] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 5

[60] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks?, 2014. 2

[61] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization, 2015. 2

[62] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2

[63] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018. 2

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

[65] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 2

[66] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[67] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 3