Artificial Intelligence Review (2022) 55:589–656
https://doi.org/10.1007/s10462-021-10039-7

Artificial intelligence in the creative industries: a review

Nantheera Anantrasirichai[1] · **David Bull**[1]

Published online: 2 July 2021

Abstract
This paper reviews the current state of the art in artificial intelligence (AI) technologies and applications in the context of the creative industries. A brief background of AI, and specifically machine learning (ML) algorithms, is provided including convolutional neural networks (CNNs), generative adversarial networks (GANs), recurrent neural networks (RNNs) and deep Reinforcement Learning (DRL). We categorize creative applications into five groups, related to how AI technologies are used: (i) content creation, (ii) information analysis, (iii) content enhancement and post production workflows, (iv) information extraction and enhancement, and (v) data compression. We critically examine the successes and limitations of this rapidly advancing technology in each of these areas. We further differentiate between the use of AI as a creative tool and its potential as a creator in its own right. We foresee that, in the near future, ML-based AI will be adopted widely as a tool or collaborative assistant for creativity. In contrast, we observe that the successes of ML in domains with fewer constraints, where AI is the 'creator', remain modest. The potential of AI (or its developers) to win awards for its original creations in competition with human creatives is also limited, based on contemporary technologies. We therefore conclude that, in the context of creative industries, maximum benefit from AI will be derived where its focus is human-centric—where it is designed to augment, rather than replace, human creativity.

Keywords Creative industries · Machine learning · Image and video enhancement

1 Introduction

The aim of new technologies is normally to make a specific process easier, more accurate, faster or cheaper. In some cases they also enable us to perform tasks or create things that were previously impossible. Over recent years, one of the most rapidly advancing scientific techniques for practical purposes has been Artificial Intelligence (AI). AI techniques enable machines to perform tasks that typically require some degree of human-like intelligence. With recent developments in high-performance computing and increased

✉ Nantheera Anantrasirichai
n.anantrasirichai@bristol.ac.uk

David Bull
dave.bull@bristol.ac.uk

[1] Bristol Vision Institute, University of Bristol, Bristol, UK

data storage capacities, AI technologies have been empowered and are increasingly being adopted across numerous applications, ranging from simple daily tasks, intelligent assistants and finance to highly specific command, control operations and national security. AI can, for example, help smart devices or computers to understand text and read it out loud, hear voices and respond, view images and recognize objects in them, and even predict what may happen next after a series of events. At higher levels, AI has been used to analyze human and social activity by observing their convocation and actions. It has also been used to understand socially relevant problems such as homelessness and to predict natural events. AI has been recognized by governments across the world to have potential as a major driver of economic growth and social progress (Hall and Pesenti 2018; NSTC 2016). This potential, however, does not come without concerns over the wider social impact of AI technologies which must be taken into account when designing and deploying these tools.

Processes associated with the creative sector demand significantly different levels of innovation and skill sets compared to routine behaviours. While AI accomplishments rely heavily on conformity of data, creativity often exploits the human imagination to drive original ideas which may not follow general rules. Basically, creatives have a lifetime of experiences to build on, enabling them to think 'outside of the box' and ask 'What if' questions that cannot readily be addressed by constrained learning systems.

There have been many studies over several decades into the possibility of applying AI in the creative sector. One of the limitations in the past was the readiness of the technology itself, and another was the belief that AI could attempt to replicate human creative behaviour (Rowe and Partridge 1993). A recent survey by Adobe[1] revealed that three quarters of artists in the US, UK, Germany and Japan would consider using AI tools as assistants, in areas such as image search, editing, and other 'non-creative' tasks. This indicates a general acceptance of AI as a tool across the community and reflects a general awareness of the state of the art, since most AI technologies have been developed to operate in closed domains where they can assist and support humans rather than replace them. Better collaboration between humans and AI technologies can thus maximize the benefits of the synergy. All that said, the first painting created solely by AI was auctioned for $432,500 in 2018.[2]

Applications of AI in the creative industries have dramatically increased in the last five years. Based on analysis of data from arXiv[3] and Gateway to Research,[4] Davies et al. (2020) revealed that the growth rate of research publications on AI (relevant to the creative industries) exceeds 500% in many countries (in Taiwan the growth rate is 1490%), and the most of these publications relate to image-based data. Analysis on company usage from the Crunchbase database[5] indicates that AI is used more in games and for immersive applications, advertising and marketing, than in other creative applications. Caramiaux et al. (2019) recently reviewed AI in the current media and creative industries across three areas: creation, production and consumption. They provide details of AI/ML-based research and development, as well as emerging challenges and trends.

[1] https://www.pfeifferreport.com/wp-content/uploads/2018/11/Creativity_and_AI_Report_INT.pdf.
[2] https://edition.cnn.com/style/article/obvious-ai-art-christies-auction-smart-creativity/index.html.
[3] https://arxiv.org/.
[4] https://gtr.ukri.org/.
[5] https://www.crunchbase.com/.

In this paper, we review how AI and its technologies are, or could be, used in applications relevant to creative industries. We first provide an overview of AI and current technologies (Sect. 1), followed by a selection of creative domain applications (Sect. 3). We group these into subsections[6] covering: (i) content creation: where AI is employed to generate original work, (ii) information analysis: where statistics of data are used to improve productivity, (iii) content enhancement and post production workflows: used to improve quality of creative work, (iv) information extraction and enhancement: where AI assists in interpretation, clarifies semantic meaning, and creates new ways to exhibit hidden information, and (v) data compression: where AI helps reduce the size of the data while preserving its quality. Finally we discuss challenges and the future potential of AI associated with the creative industries in Sect. 4.

2 An introduction to artificial intelligence

Artificial intelligence (AI) embodies a set of codes, techniques, algorithms and data that enables a computer system to develop and emulate human-like behaviour and hence make decisions similar to (or in some cases, better than) humans (Russell and Norvig 2020). When a machine exhibits full human intelligence, it is often referred to as 'general AI' or 'strong AI' (Bostrom 2014). However, currently reported technologies are normally restricted to operation in a limited domain to work on specific tasks. This is called 'narrow AI' or 'weak AI'. In the past, most AI technologies were model-driven; where the nature of the application is studied and a model is mathematically formed to describe it. Statistical learning is also data-dependent, but relies on rule-based programming (James et al. 2013). Previous generations of AI (mid-1950s until the late 1980s (Haugeland 1985)) were based on symbolic AI, following the assumption that humans use symbols to represent things and problems. Symbolic AI is intended to produce general, human-like intelligence in a machine (Honavar 1995), whereas most modern research is directed at specific sub-problems.

2.1 Machine learning, neurons and artificial neural networks

The main class of algorithms in use today are based on machine learning (ML), which is data-driven. ML employs computational methods to 'learn' information directly from large amounts of example data without relying on a predetermined equation or model (Mitchell 1997). These algorithms adaptively converge to an optimum solution and generally improve their performance as the number of samples available for learning increases. Several types of learning algorithms exist, including supervised learning, unsupervised learning and reinforcement learning. Supervised learning algorithms build a mathematical model from a set of data that contains both the inputs and the desired outputs (each output usually representing a classification of the associated input vector), while unsupervised learning algorithms model the problems on unlabeled data. Self-supervised learning is a form of unsupervised learning where the data provide the measurable structure to build a loss function. Semi-supervised learning employs a limited set of labeled data to label,

[6] While we hope that this categorization is helpful, it should be noted that several of the applications described could fit into, or span, multiple categories.

usually a larger amount of, unlabeled data. Then both datasets are combined to create a new model. Reinforcement learning methods learn from trial and error and are effectively self-supervised (Russell and Norvig 2020).

Modern ML methods have their roots in the early computational model of a neuron proposed by Warren MuCulloch (neuroscientist) and Walter Pitts (logician) in (1943). This is shown in Fig. 1a. In their model, the artificial neuron receives one or more inputs, where each input is independently weighted. The neuron sums these weighted inputs and the result is passed through a non-linear function known as an activation function, representing the neuron's action potential which is then transmitted along its axon to other neurons. The multi-layer perceptron (MLP) is a basic form of artificial neural network (ANN) that gained popularity in the 1980s. This connects its neural units in a multi-layered (typically one input layer, one hidden layer and one output layer) architecture (Fig. 1b). These neural layers are generally fully connected to adjacent layers, (i.e., each neuron in one layer is connected to all neurons in the next layer). The disadvantage of this approach is that the total number of parameters can be very large and this can make them prone to overfitting data.

For training, the MLP (and most supervised ANNs) utilizes error backpropagation to compute the gradient of a loss function. This loss function maps the event values from multiple inputs into one real number to represent the cost of that event. The goal of the training process is therefore to minimize the loss function over multiple presentations of the input dataset. The backpropagation algorithm was originally introduced in the 1970s, but peaked in popularity after (1986), when Rumelhart et al. described several neural networks where backpropagation worked far faster than earlier approaches, making ANNs applicable to practical problems.

2.2 An introduction to deep neural networks

Deep learning is a subset of ML that employs deep artificial neural networks (DNNs). The word 'deep' means that there are multiple hidden layers of neuron collections that have learnable weights and biases. When the data being processed occupies multiple dimensions (images for example), convolutional neural networks (CNNs) are often employed. CNNs are (loosely) a biologically-inspired architecture and their results are tiled so that they overlap to obtain a better representation of the original inputs.

The first CNN was designed by Fukushima (1980) as a tool for visual pattern recognition (Fig. 2a). This so called Neocognitron was a hierarchical architecture with multiple convolutional and pooling layers. LeCun et al. (1989) applied the standard backpropagation algorithm to a deep neural network with the purpose of recognizing handwritten ZIP codes. At that time, it took 3 days to train the network. Lecun et al. (1998) proposed LeNet5 (Fig. 2b), one of the earliest CNNs which could outperform other models for handwritten character recognition. The deep learning breakthrough occurred in the 2000s driven by the availability of graphics processing units (GPUs) that could dramatically accelerate training. Since around 2012, CNNs have represented the state of the art for complex problems such as image classification and recognition, having won several major international competitions.

A CNN creates its filters' values based on the task at hand. Generally, the CNN learns to detect edges from the raw pixels in the first layer, then uses those edges to detect simple shapes in the next layer, and so on building complexity through subsequent layers. The higher layers produce high-level features with more semantically relevant meaning. This

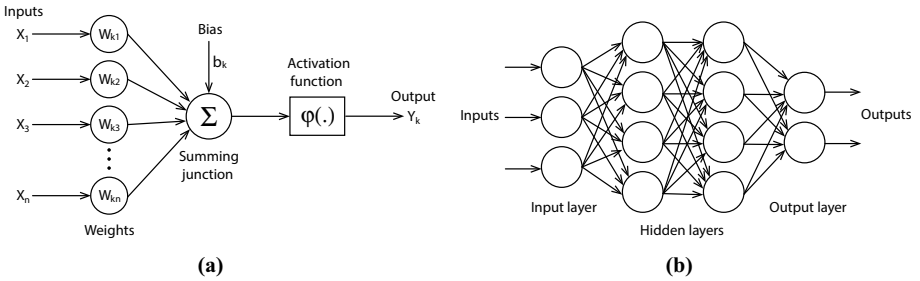

Fig. 1 **a** Basic neural network unit by MuCulloch and Pitts. **b** Basic multi-layer perceptron (MLP)

means that the algorithms can exploit both low-level features and a higher-level understanding of what the data represent. Deep learning has therefore emerged as a powerful tool to find patterns, analyze information, and to predict future events. The number of layers in a deep network is unlimited but most current networks contain between 10 and 100 layers.

Goodfellow et al. (2014) proposed an alternative form of architecture referred to as a Generative Adversarial Network (GAN). GANs consist of 2 AI competing modules where the first creates images (the generator) and the second (the discriminator) checks whether the received image is real or created from the first module. This competition results in the final picture being very similar to the real image. Because of their performance in reducing deceptive results, GAN technologies have become very popular and have been applied to numerous applications, including those related to creative practice.

While many types of machine learning algorithms exist, because of their prominence and performance, in this paper we place emphasis on deep learning methods. We will describe various applications relevant to the creative industries and critically review the methodologies that achieve, or have the potential to achieve, good performance.

2.3 Current AI technologies

This section presents state-of-the-art AI methods relevant to the creative industries. For those readers who prefer to focus on the applications, please refer to Sect. 3.

2.3.1 AI and the need for data

An AI system effectively combines a computational architecture and a learning strategy with a data environment in which it learns. Training databases are thus a critical component in optimizing the performance of ML processes and hence a significant proportion of the value of an AI system resides in them. A well-designed training database with appropriate size and coverage can help significantly with model generalization and avoiding problems of overfitting.

In order to learn without being explicitly programmed, ML systems must be trained using data having statistics and characteristics typical of the particular application domain under consideration. This is true regardless of training methods (see Sect. 2.1). Good datasets typically contain large numbers of examples with a statistical distribution matched to this domain. This is crucial because it enables the network to estimate gradients in the data (error) domain that enables it to converge to an optimum solution, forming robust decision

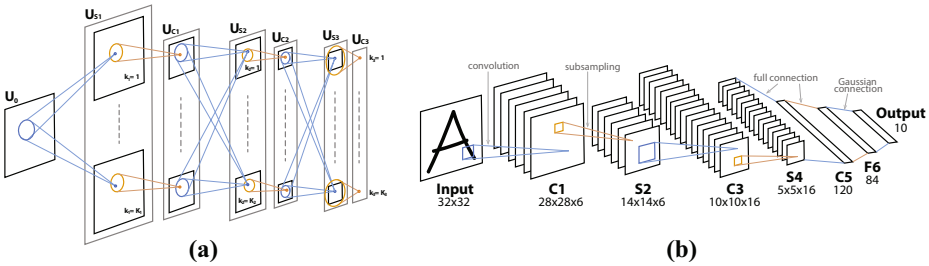

Fig. 2 **a** Neocognitron (Fukushima 1980), where U_s and U_c learn simple and complex features, respectively. **b** LeNet5 (Lecun et al. 1998), consisting of two sets of convolutional and average pooling layers, followed by a flattening convolutional layer, then two fully-connected layers and finally a softmax classifier

boundaries between its classes. The network will then, after training, be able to reliably match new unseen information to the right answer when deployed.

The reliability of training dataset labels is key in achieving high performance supervised deep learning. These datasets must comprise: i) data that are statistically similar to the inputs when the models are used in the real situations and ii) ground truth annotations that tell the machine what the desired outputs are. For example, in segmentation applications, the dataset would comprise the images and the corresponding segmentation maps indicating homogeneous, or semantically meaningful regions in each image. Similarly for object recognition, the dataset would also include the original images while the ground truth would be the object categories, e.g., car, house, human, type of animals, etc.

Some labeled datasets are freely available for public use,[7] but these are limited, especially in certain applications where data are difficult to collect and label. One of the largest, ImageNet, contains over 14 million images labeled into 22,000 classes. Care must be taken when collecting or using data to avoid imbalance and bias—skewed class distributions where the majority of data instances belong to a small number of classes with other classes being sparsely populated. For instance, in colorization, blue may appear more often as it is a color of sky, while pink flowers are much rarer. This imbalance causes ML algorithms to develop a bias towards classes with a greater number of instances; hence they preferentially predict majority class data. Features of minority classes are treated as noise and are often ignored.

Numerous approaches have been introduced to create balanced distributions and these can be divided into two major groups: modification of the learning algorithm, and data manipulation techniques (He and Garcia 2009). Zhang et al. (2016) solve the class-imbalance problem by re-weighting the loss of each pixel at train time based on the pixel color rarity. Recently, Lehtinen et al. (2018) have introduced an innovative approach to learning via their Noise2Noise network which demonstrates that it is possible to train a network without clean data if the corrupted data complies with certain statistical assumptions. However, this technique needs further testing and refinement to cope with real-world noisy data. Typical data manipulation techniques include downsampling majority classes, oversampling minority classes, or both. Two primary techniques are used to expand, adjust and rebalance the number of samples in the dataset and, in turn, to improve ML performance

[7] https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research, https://ieee-dataport.org/.

and model generalization: data augmentation and data synthesis. These are discussed further below.

2.3.1.1 Data augmentation Data augmentation techniques are frequently used to increase the volume and diversity of a training dataset without the need to collect new data. Instead, existing data are used to generate more samples, through transformations such as cropping, flipping, translating, rotating and scaling (Anantrasirichai et al. 2018; Krizhevsky et al. 2012). This can assist by increasing the representation of minority classes and also help to avoid overfitting, which occurs when a model memorizes the full dataset instead of only learning the main concepts which underlie the problem. GANs (see Section 2.3.3) have recently been employed with success to enlarge training sets, with the most popular network currently being CycleGAN (Zhu et al. 2017). The original CycleGAN mapped one input to only one output, causing inefficiencies when dataset diversity is required. Huang et al. (2018) improved CycleGAN with a structure-aware network to augment training data for vehicle detection. This slightly modified architecture is trained to transform contrast CT images (computed tomography scans) into non-contrast images (Sandfort et al. 2019). A CycleGAN-based technique has also been used for emotion classification, to amplify cases of extremely rare emotions such as disgust (Zhu et al. 2018). IBM Research introduced a Balancing GAN (Mariani et al. 2018), where the model learns useful features from majority classes and uses these to generate images for minority classes that avoid features close to those of majority cases. An extensive survey of data augmentation techniques can be found in Shorten and Khoshgoftaar (2019).

2.3.1.2 Data synthesis Scientific or parametric models can be exploited to generate synthetic data in those applications where it is difficult to collect real data, and where data augmentation techniques cannot increase variety in the dataset. Examples include signs of disease (Alsaih et al. 2017) and geological events that rarely happen (Anantrasirichai et al. 2019). In the case of creative processes, problems are often ill-posed as ground truth data or ideal outputs are not available. Examples include post-production operations such as deblurring, denoising and contrast enhancement. Synthetic data are often created by degrading the clean data. Su et al. (2017) applied synthetic motion blur on sharp video frames to train the deblurring model. LLNet (Lore et al. 2017), enhances low-light images, and is trained using a dataset generated with synthetic noise and intensity adjustment, while LLCNN (Tao et al. 2017) employs a gamma adjustment technique.

2.3.2 Convolutional neural networks (CNNs)

2.3.2.1 Basic CNNs Convolutional neural networks (CNNs) are a class of deep feed-forward ANN. They comprise a series of convolutional layers that are designed to take advantage of 2D structures, such as found in images. These employ locally connected layers that apply convolution operations between a predefined-size kernel and an internal signal; the output of each convolutional layer is the input signal modified by a convolution filter. The weights of the filter are adjusted according to a loss function that assesses the mismatch (during training) between the network output and the ground truth values or labels. Commonly used loss functions include ℓ_1, ℓ_2, SSIM (Tao et al. 2017) and perceptual loss (Johnson et al. 2016)). These errors are then backpropagated through multiple forward and backward iterations and the filter weights adjusted based on estimated gradients of the local error surface. This in turn drives what features are detected, associating them to the characteristics of the train-

ing data. The early layers in a CNN extract low-level features conceptually similar to visual basis functions found in the primary visual cortex (Matsugu et al. 2003).

The most common CNN architecture (Fig. 3a[8]) has the outputs from its convolution layers connected to a pooling layer, which combines the outputs of neuron clusters into a single neuron. Subsequently, activation functions such as *tanh* (the hyperbolic tangent) or *ReLU* (Rectified Linear Unit) are applied to introduce non-linearity into the network (Agostinelli et al. 2015). This structure is repeated with similar or different kernel sizes. As a result, the CNN learns to detect edges from the raw pixels in the first layer, then combines these to detect simple shapes in the next layer. The higher layers produce higher-level features, which have more semantic meaning. The last few layers represent the classification part of the network. These consist of fully connected layers (i.e. being connected to all the activation outputs in the previous layer) and a softmax layer, where the output class is modelled as a probability distribution - exponentially scaling the output between 0 and 1 (this is also referred to as a normalised exponential function).

VGG (Simonyan and Zisserman 2015) is one of the most common backbone networks, offering two depths: VGG-16 and VGG-19 with 16 and 19 layers respectively. The networks incorporate a series of convolution blocks (comprising convolutional layers, ReLU activations and a max-pooling layer), and the last three layers are fully connection with ReLU activations. VGG employs very small receptive fields (3×3 with a stride of 1) allowing deeper architectures than the older networks. DeepArt (Gatys et al. 2016) employs a VGG-Network without fully connected layers. It demonstrates that the higher layers in the VGG network can represent the content of an artwork. The pre-trained VGG network is widely used to provide a measure of perceptual loss (and style loss) during the training process of other networks (Johnson et al. 2016).

2.3.2.2 CNNs with reconstruction The basic structure of CNNs described in the previous section is sometimes called an 'encoder'. This is because the network learns a representation of a set of data, which often has fewer parameters than the input. In other words, it compresses the input to produce a code or a latent-space representation. In contrast, some architectures omit pooling layers in order to create dense features in an output with the same size as the input.

Alternatively, the size of the feature map can be enlarged to that of the input via deconvolutional layers or transposed convolution layers (Fig. 3b[9]). This structure is often referred to as a 'decoder' as it generates the output using the code produced by the encoder. Encoder-decoder architectures combine an encoder and a decoder. Autoencoders are a special case of encoder-decoder models, where the input and output are the same size. Encoder-decoder models are suitable for creative applications, such as style transfer (Zhang et al. 2016), image restoration (Nah et al. 2017; Yang and Sun 2018; Zhang et al. 2017), contrast enhancement (Lore et al. 2017; Tao et al. 2017), colorization (Zhang et al. 2016) and super-resolution (Shi et al. 2016).

Some architectures also add skip connections or a bridge section (Long et al. 2015) so that the local and global features, as well as semantics are connected and captured, providing improved pixel-wise accuracy. These techniques are widely used in object detection (Anantrasirichai and Bull 2019) and object tracking (Redmon and Farhadi 2018). U-Net

[8] https://uk.mathworks.com/solutions/deep-learning/convolutional-neural-network.html.
[9] https://uk.mathworks.com/solutions/image-video-processing/semantic-segmentation.html.

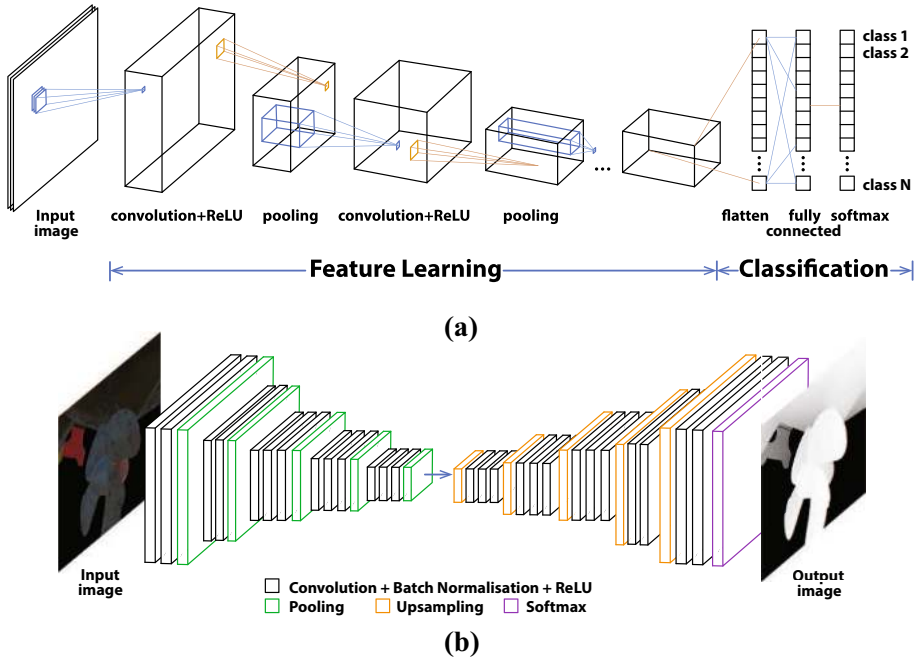

Fig. 3 CNN architectures for **a** object recognition adapted from [8], **b** semantic segmentation [9]

(Ronneberger et al. 2015) is perhaps the most popular network of this kind, even though it was originally developed for biomedical image segmentation. Its network consists of a contracting path (encoder) and an expansive path (decoder), giving it the u-shaped architecture. The contracting path consists of the repeated application of two 3×3 convolutions, followed by ReLU and a max-pooling layer. Each step in the expansive path consists of a transposed convolution layer for upsampling, followed by two sets of convolutional and ReLU layers, and concatenations with correspondingly-resolution features from the contracting path.

2.3.2.3 Advanced CNNs Some architectures introduce modified convolution operations for specific applications. For example, dilated convolution (Yu and Koltun 2016), also called atrous convolution, enlarges the receptive field, to support feature extraction locally and globally. The dilated convolution is applied to the input with a defined spacing between the values in a kernel. For example, a 3×3 kernel with a dilation rate of 2 has the same receptive field as a 5×5 kernel, but using 9 parameters. This has been used for colorization by Zhang et al. (2016) in the creative sector. ResNet is an architecture developed for residual learning, comprising several residual blocks (He et al. 2016). A single residual block has two convolution layers and a skip connection between the input and the output of the last convolution layer. This avoids the problem of vanishing gradients, enabling very deep CNN architectures. Residual learning has become an important part of the state of the art in many application, such as contrast enhancement (Tao et al. 2017), colorization (Huang et al. 2017), SR (Dai et al. 2019; Zhang et al. 2018a), object recognition (He et al. 2016), and denoising (Zhang et al. 2017).

Traditional convolution operations are performed in a regular grid fashion, leading to limitations for some applications, where the object and its location are not in the regular grid. Deformable convolution (Dai et al. 2017) has therefore been proposed to facilitate the region of support for the convolution operations to take on any shape, instead of just the traditional square shape. This has been used in object detection and SR (Wang et al. 2019a). 3D deformable kernels have also been proposed for denoising video content, as they can better cope with large motions, producing cleaner and sharper sequences (Xiangyu Xu 2019).

Capsule networks were developed to address some of the deficiencies with traditional CNNs (Sabour et al. 2017). They are able to better model hierarchical relationships, where each neuron (referred to as a capsule) expresses the likelihood and properties of its features, e.g., orientation or size. This improves object recognition performance. Capsule networks have been extended to other applications that deal with complex data, including multi-label text classification (Zhao et al. 2019), slot filling and intent detection (Zhang et al. 2019a), polyphonic sound event detection (Vesperini et al. 2019) and sign language recognition (Jalal et al. 2018).

2.3.3 Generative adversarial networks (GANs)

The generative adversarial network (GAN) is a recent algorithmic innovation that employs two neural networks: generative and discriminative. The GAN pits one against the other in order to generate new, synthetic instances of data that can pass for real data. The general GAN architecture is shown in Fig. 4a. It can be observed that the generative network generates new candidates to increase the error rate of the discriminative network until the discriminative network cannot tell whether these candidates are real or synthesized. The generator is typically a deconvolutional neural network, and the discriminator is a CNN. Recent successful applications of GANs include SR (Ledig et al. 2017), inpainting (Yu et al. 2019), contrast enhancement (Kuang et al. 2019) and compression (Ma et al. 2019a).

GANs have a reputation of being difficult to train since the two models are trained simultaneously to find a Nash equilibrium but with each model updating its cost (or error) independently. Failures often occur when the discriminator cannot feedback information that is good enough for the generator to make progress, leading to vanishing gradients. Wasserstein loss is designed to prevent this (Arjovsky et al. 2017; Frogner et al. 2015). A specific condition or characteristic, such as a label associated with an image, rather than a generic sample from an unknown noise distribution can be included in the generative model, creating what is referred to as a conditional GAN (cGAN) (Mirza and Osindero 2014). This improved GAN has been used in several applications, including pix2pix(Isola et al. 2017) and for deblurring (Kupyn et al. 2018).

Theoretically, the generator in a GAN will not learn to create new content, but it will just try to make its output look like the real data. Therefore, to produce creative works of art, the Creative Adversarial Network (CAN) has been proposed by Elgammal et al. (2017). This works by including an additional signal in the generator to prevent it from generating content that is too similar to existing examples. Similar to traditional CNNs, a perceptual loss based on VGG16 (Johnson et al. 2016) has become common in applications where new images are generated that have the same semantics as the input (Antic 2020; Ledig et al. 2017).

Most GAN-based methods are currently limited to the generation of relatively small square images, e.g., 256×256 pixels (Zhang et al. 2017). The best resolution created up

to the time of this review is 1024 × 1024-pixels, achieved by NVIDIA research. The team introduced the progressive growing of GANs (Karras et al. 2018) and showed that their method can generate near-realistic 1024 × 1024-pixel portrait images (trained for 14 days). However the problem of obvious artefacts at transition areas between foreground and background persists.

Another form of deep generative model is the Variational Autoencoder (VAE). A VAE is an autoencoder, where the encoding distribution is regularised to ensure the latent space has good properties to support the generative process. Then the decoder samples from this distribution to generate new data. Comparing VAEs to GANs, VAEs are more stable during training, while GANs are better at producing realistic images. Recently Deepmind (Google) has included vector quantization (VQ) within a VAE to learn a discrete latent representation (Razavi et al. 2019). Its performance for image generation are competitive with their BigGAN (Brock et al. 2019) but with greater capacity for generating a diverse range of images. There have also been many attempts to merge GANs and VAEs so that the end-to-end network benefits from both good samples and good representation, for example using a VAE as the generator for a GAN (Bhattacharyya et al. 2019; Wan et al. 2017). However, the results of this have not yet demonstrated significant improvement in terms of overall performance (Rosca et al. 2019), remaining an ongoing research topic.

A review of recent state-of-the-art GAN models and applications can be found in Foster (2019).

2.3.4 Recurrent neural networks (RNNs)

Recurrent neural networks (RNNs) have been widely employed to perform sequential recognition; they offer benefits in this respect by incorporating at least one feedback connection. The most commonly used type of RNN is the Long Short-Term Memory (LSTM) network (Hochreiter and Schmidhuber 1997), as this solves problems associated with vanishing gradients, observed in traditional RNNs. It does this by memorizing sufficient context information in time series data via its memory cell. Deep RNNs use their internal state to process variable length sequences of inputs, combining across multiple levels of representation. This makes them amenable to tasks such as speech recognition (Graves et al. 2013), handwriting recognition (Doetsch et al. 2014), and music generation (Briot et al. 2020). RNNs are also employed in image and video processing applications, where recurrency is applied to convolutional encoders for tasks such as drawing sketches (Ha and Eck 2018) and deblurring videos (Zhang et al. 2018). VINet (Kim et al. 2019) employs an encoder-decoder model using an RNN to estimate optical flow, processing multiple input frames concatenated with the previous inpainting results. An example network using an RNN is illustrated in Fig. 4b.

CNNs extract spatial features from its input images using convolutional filters and RNNs extract sequential features in time-series data using memory cells. In extension, 3D CNN, CNN-LSTM and ConvLSTM have been designed to extract spatial-temporal features from video sequences. The 3D activation maps produced in 3D CNNs are able to analyze temporal or volumetric context which are important in applications such as medical imaging (Lundervold and Lundervold 2019) and action recognition (Ji et al. 2013). The CNN-LSTM simply concatenates a CNN and an LSTM (the 1D output of the CNN is the input to the LSTM) to process time-series data. In contrast, ConvLSTM is another LSTM variant, where the internal matrix multiplications are replaced with convolution operations

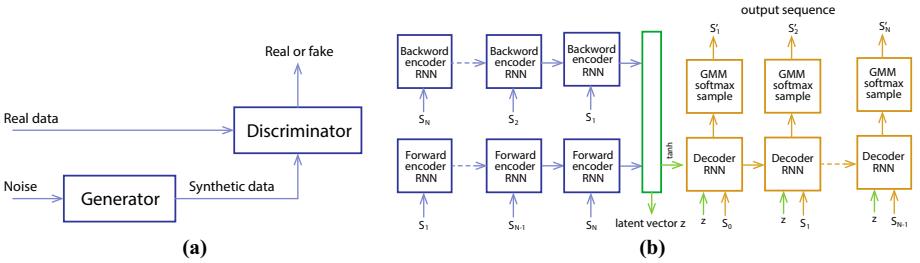

Fig. 4 Architectures of **a** GAN, **b** RNN for drawing sketches (Ha and Eck 2018)

at each gate of the LSTM cell so that the LSTM input can be in the form of multi-dimensional data (Shi et al. 2015).

2.3.5 Deep reinforcement learning (DRL)

Reinforcement learning (RL) is an ML algorithm trained to make a sequence of decisions. Deep reinforcement learning (DRL) combines ANNs with an RL architecture that enables RL agents to learn the best actions in a virtual environment to achieve their goals. The RL agents are comprised of a policy that performs a mapping from an input state to an output action and an algorithm responsible for updating this policy. This is done through leveraging a system of rewards and punishments to acquire useful behaviour—effectively a trial-and-error process. The framework trains using a simulation model, so it does not require a predefined training dataset, either labeled or unlabeled.

However, pure RL requires an excessive number of trials to learn fully, something that may be impractical in many (especially real-time) applications if training from scratch (Hessel et al. 2018). AlphaGo, a computer program developed by DeepMind Technologies that can beat a human professional Go player, employs RL on top of a pre-trained model to improve its play strategy to beat a particular player.[10] RL could be useful in creative applications, where there may not be a predefined way to perform a given task, but where there are rules that the model has to follow to perform its duties correctly. Current applications involve end-to-end RL combined with CNNs, including gaming (Mnih et al. 2013), and RLs with GANs in optimal painting stroke in stroke-based rendering (Huang et al. 2019). Recently RL methods have been developed using a graph neural network (GNN) to play Diplomacy, a highly complex 7-player (large scale) board game (Anthony et al. 2020).

Temporal difference (TD) learning (Gregor et al. 2019; Chen et al. 2018; Nguyen et al. 2020) has recently been introduced as a model-free reinforcement learning method that learns how to predict a quantity that depends on future values of a given signal. That is, the model learns from an environment through episodes with no prior knowledge of the environment. This may well have application in the creative sector for storytelling, caption-from-image generation and gaming.

[10] https://ai.googleblog.com/2016/01/alphago-mastering-ancient-game-of-go.html.

3 AI for the creative industries

AI has increasingly (and often mistakenly) been associated with human creativity and artistic practice. As it has demonstrated abilities to 'see', 'hear', 'speak', 'move', and 'write', it has been applied in domains and applications including: audio, image and video analysis, gaming, journalism, script writing, filmmaking, social media analysis and marketing. One of the earliest AI technologies, available for more than two decades, is Autotune, which automatically fixes vocal intonation errors (Hildebrand 1999). An early attempt to exploit AI for creating art occurred in 2016, when a three-dimensional (3D) printed painting, the Next Rembrandt,[11] was produced solely based on training data from Rembrandt's portfolio. It was created using deep learning algorithms and facial recognition techniques.

Creativity is defined in the Cambridge Dictionary as 'the ability to produce original and unusual ideas, or to make something new or imaginative'. Creative tasks generally require some degree of original thinking, extensive experience and an understanding of the audience, while production tasks are, in general, more repetitive or predictable, making them more amenable to being performed by machines. To date, AI technologies have produced mixed results when used for generating original creative works. For example, GumGum[12] creates a new piece of art following the input of a brief idea from the user. The model is trained by recording the preferred tools and processes that the artist uses to create a painting. A Turing test revealed that it is difficult to distinguish these AI generated products from those painted by humans. AI methods often produce unusual results when employed to create new narratives for books or movie scripts. Botnik[13] employs an AI algorithm to automatically remix texts of existing books to create a new chapter. In one experiment, the team fed the seven Harry Potter novels through their predictive text algorithm, and the 'bot' created rather strange but amusing sentences, such as "*Ron was standing there and doing a kind of frenzied tap dance. He saw Harry and immediately began to eat Hermione's family*" (Sautoy 2019). However, when AI is used to create less structured content (e.g., some forms of 'musical' experience), it can demonstrate pleasurable difference (Briot et al. 2020).

In the production domain, Twitter has applied automatic cropping to create image thumbnails that show the most salient part of an image (Theis et al. 2018). The BBC has created a proof-of-concept system for automated coverage of live events. In this work, the AI-based system performs shot framing (wide, mid and close-up shots), sequencing, and shot selection automatically (Wright et al. 2020). However, the initial results show that the algorithm needs some improvement if it is to replace human operators. Nippon Hoso Kyokai (NHK, Japan's Broadcasting Corporation), has developed a new AI-driven broadcasting technology called "Smart Production". This approach extracts events and incidents from diverse sources such as social media feeds (e.g., Twitter), local government data and interviews, and integrates these into a human-friendly accessible format (Kaneko et al. 2020).

In this review, we divide creative applications into five major categories: content creation, information analysis, content enhancement and post production workflows, information extraction and enhancement, and data compression. However, it should be noted that

[11] https://www.nextrembrandt.com/.
[12] https://gumgum.com/artificial-creativity.
[13] https://botnik.org/.

Table 1 Creative applications and corresponding AI-based methods

Application		Technology			
		CNN	GAN	RNN	other
Creation	Content generation for text, audio, video and game	Gatys et al. (2016), Karpathy et al. (2019), Quesnel et al. (2018) Zhang (2020), Zhang et al. (2016)	Brock et al. (2019), Donahue et al. (2019), Engel et al. (2019) He et al. (2019), Isola et al. (2017), Jin et al. (2017) Karras et al. (2018), Kim et al. (2020b), Li et al. (2019b, 2018b), Radford et al. (2016), Song et al. (2018b, 2018b), Subramanian et al. (2018), Wang et al. (2018), Yi et al. (2017), Zakharov et al. (2019), Zhu et al. (2017)	Dzmitry Bahdanau (2015), Ha and Eck (2018), Kim et al. (2020b) Li et al. (2019b), Mao et al. (2018), Sturm et al. (2016), Subramanian et al. (2018), Venugopalan et al. (2015)	TM (Dörr 2016), RL (Chen et al. 2020; Gregor et al. 2019; Chen et al. 2018; Nguyen et al. 2020; Wang et al. 2017), BERT (Devlin et al. 2019), VAE (Razavi et al. 2019), NEAT (Stanley et al. 2009), Graph-based (Chaplot et al. 2020)
	Animation	Holden et al. (2015), Nalbach et al. (2017), Starke et al. (2019, 2020), Tesfaldet et al. (2018)	Nagano et al. (2018), Wei et al. (2019)	Lee et al. (2018), Siyao et al. (2021)	VAE (Wei et al. 2019)
	AR/VR	Anantrasirichai et al. (2016), Panphattarasap and Calway (2018)			
	Deepfakes		Chan et al. (2019)	Suwajanakorn et al. (2017)	VAE (Kietzmann et al. 2020)
	Content and captions	Chen et al. (2018), Chen et al. (2019), Pu et al. (2016)	Li et al. (2019c), Mansimov et al. (2016), Zhang et al. (2017),	Chen et al. (2018), Xia and Wang (2005), Xu et al. (2017b)	VAE (Pu et al. 2016)
Information Analysis	Text categorization	Johnson and Zhang (2015)	Li et al. (2018b)	Chen et al. (2017), Gunasekara and Nejadgholi (2018), Truşcă et al. (2020)	SOM (Pawar and Gawande 2012), ELM (Rezaei-Ravari et al. 2021)
	Ads/film analysis	Young et al. (2018)	Young et al. (2018)	Young et al. (2018)	GP (Lacerda et al. 2006)

Table 1 (continued)

Application		Technology			
		CNN	GAN	RNN	other
	Content retrieval	Amato et al. (2017),Gordo et al. (2016), Wan et al. (2014),Wu et al. (2015)	Song et al. (2018a)	Jabeen et al. (2018)	PM (Jeon et al. 2003)
	Fake detection	Güera and Delp (2018), Li and Lyu (2019)		Güera and Delp (2018)	Blockchain (Hasan and Salah 2019)
	Recommendation	Deldjoo et al. (2018)		See et al. (2017), Li et al. (2019a), Rush et al. (2015), Yi et al. (2020)	Deep belief net (Batmaz et al. 2019), regularization (Chen et al. 2014)
Content	Contrast	Lore et al. (2017)	Anantrasirichai and Bull (2021), Jiang et al. (2021) and Kuang et al. (2019)		Histogram (Pizer et al. 1987)
Enhancement and Post	Colorization	Cheng et al. (2015), Limmer and Lensch (2016), Ronneberger et al. (2015), Xu et al. (2020)	Anantrasirichai and Bull (2021) Antic (2020), Kuang et al. (2020), Suarez et al. (2017) and Zhang et al. (2019)		
Production	Super-resolution	Caballero et al. (2017) Dai et al. (2017, 2019), Dong et al. (2014), Haris et al. (2019), Huang et al. (2015) Kappeler et al. (2016), Kim et al. (2016), Liu et al. (2018a) Sajjadi et al. (2017), Shi et al. (2016), Tai et al. (2017), Wang et al. (2019a), Wang et al. (2019), Zhang et al. (2018a)	Ledig et al. (2017) and Wang et al. (2018)	Haris et al. (2019) and Huang et al. (2015)	

Table 1 (continued)

Application		Technology			
		CNN	GAN	RNN	other
	Deblurring	Gao et al. (2019), Hradis et al. (2015), Hyun Kim et al. (2017), Nah et al. (2017), Schuler et al. (2016), Su et al. (2017), Tao et al. (2018), Zhang et al. (2018) and Zhou et al. (2019)	Kupyn et al. (2018)	Tao et al. (2018)	Statistic model (Biemond et al. 1990; Hyun Kim et al. 2017; Nah et al. 2019; Zhang et al. 2018), BD (Jia 2007; Krishnan et al. 2011)
	Denoising	Brooks et al. (2019), Chen et al. (2018a, 2018b), Claus and van Gemert (2019), Davy et al. (2019), Krull et al. (2019), Lehtinen et al. (2018), Lempitsky et al. (2018), Li et al. (2018a) Liu et al. (2018b), Xiangyu Xu (2019), Xue et al. (2019), Zhang et al. (2017), Zhang et al. (2018), Zhao et al. (2019a)	Chen et al. (2018b) and Yang et al. (2018)	Anantrasirichai and Bull (2021), Maas et al. (2012) and Zhang et al. (2018b)	Filtering (Buades and Duran 2019; Maggioni et al. 2012; Malm et al. 2007; Yahya et al. 2016; Zuo et al. 2013)
	Dehazing	Cai et al. (2016), Hu et al. (2018), Li et al. (2017), Li et al. (2016) and Yang and Sun (2018)	Engin et al. (2018) and Tang et al. (2019)		
	Turbulence removal	Gao et al. (2019) and Nieuwenhuizen and Schutte (2019)	Chak et al. (2018)		Fusion (Anantrasirichai et al. 2013),

Table 1 (continued)

Application		Technology			
		CNN	GAN	RNN	other
					BD (Xie et al. 2016; Zhu and Milanfar 2013)
	Inpainting	Hong et al. (2019), Kim et al. (2019) and Xie et al. (2012)	Chang et al. (2019), Yu et al. (2018) and Yu et al. (2019)	Kim et al. (2019)	Sparse coding (Xie et al. 2012)
	VFX	Hu et al. (2017), Torrejon et al. (2020)			Filtering (Barber et al. 2016)
Information Extraction	Segmentation	Asgari Taghanaki et al. (2021), Kirillov et al. (2020), Long et al. (2015), Noh et al. (2015), Qi et al. (2017) and Ronneberger et al. (2015)	Isola et al. (2017)		
	Recognition	Cai and Pu (2019), Chen et al. (2019), He et al. (2017), Jalal et al. (2018), Ji et al. (2013), Kazakos et al. (2019), Kim et al. (2020a), Peng et al. (2018), Redmon and Farhadi (2018), Redmon et al. (2016), Ren et al. (2017), Shillingford et al. (2019), Sun et al. (2018), Adithya and Rajesh (2020), Wang et al. (2016), Yang et al. (2020b) and Zhen et al. (2019)		Shillingford et al. (2019)	

Table 1 (continued)

Application		Technology			
		CNN	GAN	RNN	other
	Tracking	Bochkovskiy et al. (2020), Borysenko et al. (2020), Li et al. (2018), Liu et al. (2020) and Wang et al. (2019b)		Fang (2016), Gordon et al. (2018) and Milan et al. (2017)	
	SOD	Fan et al. (2019), Fan et al. (2020) and Hou et al. (2019)	Jiang et al. (2020), Mejjati et al. (2020) and Wang et al. (2020a)	Fan et al. (2019)	graph-cut (Cheng et al. 2010), multi-scale features (Gupta et al. 2013)
	Fusion	Liu et al. (2017) and Prabhakar et al. (2017)	Ma et al. (2019c) and Wu et al. (2019)		Filtering (Anantrasirichai et al. 2020b; Li et al. 2013; Ma et al. 2019b)
	3D Reconstruction	Anantrasirichai et al. (2021), Bulat and Tzimiropoulos (2017), Chang and Chen (2018), Cheng et al. (2019), Flynn et al. (2019), Gao and Grauman (2019), Gkioxari et al. (2019), Jackson et al. (2017), Kanazawa et al. (2018), Mescheder et al. (2019), Morgado et al. (2018), Newell et al. (2016), Shi et al. (2020), Tewari et al. (2020), Vasudevan et al. (2020), Xie et al. (2019), Zhang et al. (2019)	Jiang et al. (2018), Shimada et al. (2019), Tian et al. (2018), Wu et al. (2017) and Yang et al. (2017)		VAE (Soltani et al. 2017)

Table 1 (continued)

Application	Technology			
	CNN	GAN	RNN	other
Compression	Han et al. (2019), Jiang et al. (2018), Li et al. (2018), Liu et al. (2018), Lu et al. (2020) Ma et al. (2019a), Oh et al. (2009), Schiopu et al. (2019), Stankiewicz (2019), Xue and Su (2019), Zhang et al. (2019b), Zhang et al. (2020) and Zhao et al. (2019a)	Ma et al. (2019a) and Ma et al. (2020a)	Goyal et al. (2019)	VAE (Han et al. 2019)

CNN Convolutional neural network, *GAN* generative adversarial network

RNN recurrent neural network, *RL* reinforcement learning, *PM* probabilistic model

BERT bidirectional encoder representations from transformers, *TM* text mining

VAE variational autoencoders, *AR* augmented reality, *VR* virtual reality

GP genetic programming, *BD* blind deconvolution, *VFX* visual effects, *SOM* self-organizing map

NEAT NeuroEvolution of augmenting topologies, *ELM* extreme learning machine

many applications exploit several categories in combination. For instance, post-production tools (discussed in Sects. 3.3 and 3.4) frequently combine information extraction and content enhancement techniques. These combinations can together be used to create new experiences, enhance existing material or to re-purpose archives (e.g., 'Venice Through a VR Lens, 1898' directed by BDH Immersive and Academy 7 Production[14]). These workflows may employ AI-enabled super-resolution, colorization, 3D reconstruction and frame rate interpolation methods. Gaming is another important example that has been key for the development of AI. It could be considered as an 'all-in-one' AI platform, since it combines rendering, prediction and learning.

We categorize the applications and the corresponding AI-based solutions as shown in Table 1. For those interested, a more detailed overview of contemporary Deep Learning systems is provided in Sect. 2.3.

3.1 Content creation

Content creation is a fundamental activity of artists and designers. This section discusses how AI technologies have been employed both to support the creative process and as a creator in their own right.

3.1.1 Script and movie generation

The narrative or story underpins all forms of creativity across art, fiction, journalism, gaming, and other forms of entertainment. AI has been used both to create stories and to optimize the use of supporting data, for example organizing and searching through huge archives for documentaries. The script of a fictional short film, Sunspring (2016),[15] was entirely written by an AI machine, known as Benjamin, created by New York University. The model, based on a recurrent neural network (RNN) architecture, was trained using science fiction screenplays as input, and the script was generated with random seeds from a sci-fi filmmaking contest. Sunspring has some unnatural story lines. In the sequel, It's No Game (2017), Benjamin was then used only in selected areas and in collaboration with humans, producing a more fluid and natural plot. This reinforces the notion that the current AI technology can work more efficiently in conjunction with humans rather than being left to its own devices. In 2016, IBM Watson, an AI-based computer system, composed the 6-min movie trailer of a horror film, called Morgan.[16] The model was trained with more than 100 trailers of horror films enabling it to learn the normative structure and pattern. Later in 2018, Benjamin was used to generate a new film 'Zone Out' (produced within 48 h). The project also experimented further by using face-swapping, based on a GAN and voice-generating technologies. This film was entirely directed by AI, but includes many artefacts and unnatural scenes as shown in Fig. 5a.[17] Recently, ScriptBook[18] introduced a story-awareness concept for AI-based storytelling. The generative models focus on three

[14] https://www.bdh.net/immersive/venice-1898-through-the-lens.
[15] https://www.imdb.com/title/tt5794766/.
[16] https://www.ibm.com/blogs/think/2016/08/cognitive-movie-trailer/.
[17] https://www.youtube.com/watch?v=vUgUeFu2Dcw.
[18] https://www.scriptbook.io.

aspects: awareness of characters and their traits, awareness of a script's style and theme, and awareness of a script's structure, so the resulting script is more natural.

In gaming, AI has been used to support design, decision-making and interactivity (Justesen et al. 2020). Interactive narrative, where users create a storyline through actions, has been developed using AI methods over the past decade (Riedl and Bulitko 2012). For example, MADE (Massive Artificial Drama Engine for non-player characters) generates procedural content in games (Héctor 2014), and deep reinforcement learning has been employed for personalization (Wang et al. 2017). AI Dungeon[19] is a web-based game that is capable of generating a storyline in real time, interacting with player input. The underlying algorithm requires more than 10,000 label contributions for training to ensure that the model produces smooth interaction with the players. Procedural generation has been used to automatically randomize content so that a game does not present content in the same order every time (Short and Adams 2017). Modern games often integrate 3D visualization, augmented reality (AR) and virtual reality (VR) techniques, with the aim of making play more realistic and immersive. Examples include Vid2Vid (Wang et al. 2018) which uses a deep neural network, trained on real videos of cityscapes, to generate a synthetic 3D gaming environment. Recently, NVIDIA Research has used a generative model [GameGAN by Kim et al. (2020b)], trained on 50,000 PAC-MAN episodes, to create new content, which can be used by game developers to automatically generate layouts for new game levels in the future.

3.1.2 Journalism and text generation

Natural language processing (NLP) refers to the broad class of computational techniques for incorporating speech and text. It analyzes natural language data and trains machines to perceive and to generate human language directly. NLP algorithms frequently involve speech recognition (Sect. 3.4), natural language understanding [e.g., BERT by Google AI (Devlin et al. 2019)], and natural language generation (Leppänen et al. 2017). Automated journalism, also known as robot journalism, describes automated tools that can generate news articles from structured data. The process scans large amounts of assorted data, orders key points, and inserts details such as names, places, statistics, and some figures (Cohen 2015). This can be achieved through NLP and text mining techniques (Dörr 2016).

AI can help to break down barriers between different languages with machine translation (Dzmitry Bahdanau 2015). A conditioned GAN with an RNN architecture has been proposed for language translation by Subramanian et al. (2018). It was used for the difficult task of generating English sentences from Chinese poems; it creates understandable text but sometimes with grammatical errors. CNN and RNN architectures are employed to translate video into natural language sentences (Venugopalan et al. 2015). AI can also be used to rewrite one article to suit several different channels or audience tastes.[20] A survey of recent deep learning methods for text generation by Iqbal and Qureshi (2020) concludes that text generated from images could be most amenable to GAN processing while topic-to-text translation is likely to be dominated by variational autoencoders (VAE).

[19] https://aidungeon.io/.

[20] https://www.niemanlab.org/2016/10/the-ap-wants-to-use-machine-learning-to-automate-turning-print-stories-into-broadcast-ones/.

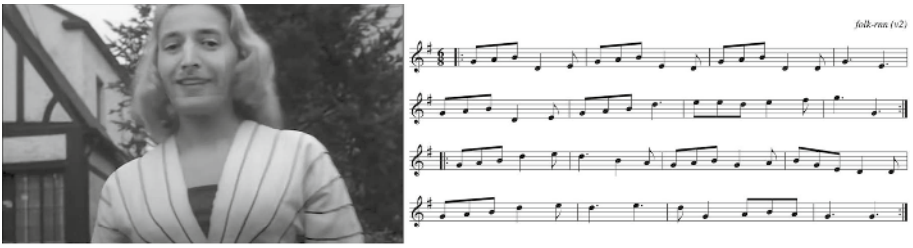

Fig. 5 **a** A screenshot from 'Zone Out', where the face of the woman was replaced with a man's mouth[17]. **b** Music transcription generated by AI algorithm[31]

Automated journalism is now quite widely used. For example, BBC reported on the UK general election in 2019 using such tools.[21] Forbes uses an AI-based content management system, called Bertie, to assist in providing reporters with the first drafts and templates for news stories.[22] The Washington Post also has a robot reporting program called Heliograf.[23] Microsoft has announced in 2020 that they use automated systems to select news stories to appear on MSN website.[24] This application of AI demonstrates that current AI technology can be effective in supporting human journalists in constrained cases, increasing production efficiency.

3.1.3 Music generation

There are many different areas where sound design is used in professional practice, including television, film, music production, sound art, video games and theatre. Applications of AI in this domain include searching through large databases to find the most appropriate match for such applications (see Sect. 3.2.3), and assisting sound design. Currently, several AI assisted music composition systems support music creation. The process generally involves using ML algorithms to analyze data to find musical patterns, e.g., chords, tempo, and length from various instruments, synthesizers and drums. The system then suggests new composed melodies that may inspire the artist. Example software includes Flow Machines by Sony,[25] Jukebox by OpenAI[26] and NSynth by Google AI.[27] In 2016, Flow Machines launched a song in the style of The Beatles, and in 2018 the team released the first AI album, 'Hello World', composed by an artist, SKYGGE (Benoit Carré), using an AI-based tool.[28] Coconet uses a CNN to infill missing pieces of music.[29] Modelling music

[21] https://www.bbc.com/news/technology-50779761.
[22] https://www.forbes.com/sites/nicolemartin1/2019/02/08/did-a-robot-write-this-how-ai-is-impacting-journalism/#5292ab617795.
[23] https://www.washingtonpost.com/pr/wp/2016/10/19/the-washington-post-uses-artificial-intelligence-to-cover-nearly-500-races-on-election-day/.
[24] https://www.bbc.com/news/world-us-canada-52860247.
[25] http://www.flow-machines.com/.
[26] https://openai.com/blog/jukebox/.
[27] https://magenta.tensorflow.org/nsynth.
[28] https://www.helloworldalbum.net/.
[29] https://magenta.tensorflow.org/coconet.

creativity is often achieved using Long Short-Term Memory (LSTM), a special type of RNN architecture (Sturm et al. 2016) (an example of the output of this model is shown in Fig. 5b[30] and the reader can experience AI-based music at Ars Electronica Voyages Channel[31]). The model takes a transcribed musical idea and transforms it in meaningful ways. For example, DeepJ composes music conditioned on a specific mixture of composer styles using a Biaxial LSTM architecture (Mao et al. 2018). More recently, generative models have been configured based on an LSTM neural network to generate music (Li et al. 2019b).

Alongside these methods of musical notation based audio synthesis, there also exists a range of direct waveform synthesis techniques that learn and/or act directly on the waveform of the audio itself [for example (Donahue et al. 2019; Engel et al. 2019]. A more detailed overview of Deep Learning techniques for music generation can be found in Briot et al. (2020).

3.1.4 Image generation

AI can be used to create new digital imagery or art-forms automatically, based on selected training datasets, e.g., new examples of bedrooms (Radford et al. 2016), cartoon characters (Jin et al. 2017), celebrity headshots (Karras et al. 2018). Some applications produce a new image conditioned to the input image, referred to as image-to-image translation, or '*style transfer*'. It is called translation or transfer, because the image output has a different appearance to the input but with similar semantic content. That is, the algorithms learn the mapping between an input image and an output image. For example, grayscale tones can be converted into natural colors (Zhang et al. 2016), using eight simple convolution layers to capture localized semantic meaning and to generate *a* and *b* color channels of the CIELAB color space. This involves mapping class probabilities to point estimates in *ab* space. DeepArt (Gatys et al. 2016) transforms the input image into the style of the selected artist by combining feature maps from different convolutional layers. A stroke-based drawing method trains machines to draw and generalise abstract concepts in a manner similar to humans using RNNs (Ha and Eck 2018).

A Berkeley AI Research team has successfully used GANs to convert between two image types (Isola et al. 2017), e.g., from a Google map to an aerial photo, a segmentation map to a real scene, or a sketch to a colored object (Fig. 6). They have published their pix2pix codebase[32] and invited the online community to experiment with it in different application domains, including depth map to street view, background removal and pose transfer. For example pix2pix has been used[33] to create a Renaissance portrait from a real portrait photo. Following pix2pix, a large number of research works have improved the performance of style transfer. Cycle-consistent adversarial networks (CycleGAN) (Zhu et al. 2017) and DualGAN (Yi et al. 2017) have been proposed for unsupervised learning. Both algorithms are based on similar concepts—the images of both groups are translated twice (e.g., from group A to group B, then translated back to the original group A) and the loss function compares the input image and its reconstruction, computing what is referred to as cycle-consistency loss. Samsung AI has shown, using GANs, that it is possible to turn

[30] https://folkrnn.org/.
[31] https://ars.electronica.art/keplersgardens/en/folk-algorithms/.
[32] https://phillipi.github.io/pix2pix/.
[33] https://ai-art.tokyo/en/.

a portrait image, such as the Mona Lisa, into a video where the portrait's face speaks in the style of a guide (Zakharov et al. 2019). Conditional GANs can be trained to transform a human face into one of a different age (Song et al. 2018b), and to change facial attributes, such as the presence of a beard, skin condition, hair style and color (He et al. 2019).

Several creative tools have employed ML-AI methods to create new unique artworks. For example, Picbreeder[34] and EndlessForms[35] employ Hypercube-based NeuroEvolution of Augmenting Topologies (Stanley et al. 2009) as a generative encoder that exploits geometric regularities. Artbreeder[36] and GANVAS Studio[37] employ BigGAN (Brock et al. 2019) to generate high-resolution class-conditional images and also to mix two images together to create new interesting work.

3.1.5 Animation

Animation is the process of using drawings and models to create moving images. Traditionally this was done by hand-drawing each frame in the sequence and rendering these at an appropriate rate to give the appearance of continuous motion. In recent years, AI methods have been employed to automate the animation process making it easier, faster and more realistic than in the past. A single animation project can involve several shot types, ranging from simple camera pans on a static scene, to more challenging dynamic movements of multiple interacting characters [e.g basketball players (Starke et al. 2020)]. ML-based AI is particularly well suited to learning models of motion from captured real motion sequences. These motion characteristics can be learnt using deep learning-based approaches, such as autoencoders (Holden et al. 2015), LSTMs (Lee et al. 2018), and motion prediction networks (Starke et al. 2019). Then, the inference applies these characteristics from the trained model to animate characters and dynamic movements. In simple animation, the motion can be estimated using a single low-cost camera. For example, Google research has created software for pose animation that turns a human pose into a cartoon animation in real time[38]. This is based on PoseNet (estimating pose position[39]) and FaceMesh (capturing face movement (Kartynnik et al. 2019)) as shown in Fig. 7a. Adobe has also created Character Animator software[40] offering lip synchronisation, eye tracking and gesture control through webcam and microphone inputs in real-time. This has been adopted by Hollywood studios and other online content creators.

AI has also been employed for rendering objects and scenes. This includes the synthesis of 3D views from motion capture or from monocular cameras (see Sect. 3.4.6), shading (Nalbach et al. 2017) and dynamic texture synthesis (Tesfaldet et al. 2018). Creating realistic lighting in animation and visual effects has also benefited by combining traditional geometrical computer vision with enhanced ML approaches and multiple depth sensors (Guo et al. 2019). Animation is not only important within the film industry; it also plays an important role in the games industry, responsible for the portrayal of movement and behaviour. Animating characters, including their faces and postures, is a key component in

[34] http://picbreeder.org/.
[35] http://endlessforms.com/.
[36] https://www.artbreeder.com/.
[37] https://ganvas.studio/.
[38] https://github.com/yemount/pose-animator/.
[39] https://www.tensorflow.org/lite/models/pose_estimation/overview.
[40] https://www.adobe.com/products/character-animator.html.

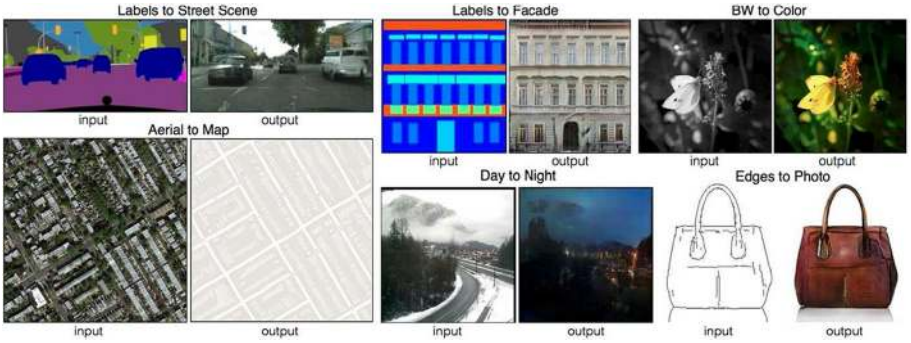

Fig. 6 Example applications of pix2pix framework (Isola et al. 2017)

Fig. 7 **a** Real-time pose animator[38]. **b** Deepfake applied to replaces Alden Ehrenreich with young Harrison Ford in Solo: a star wars story by derpfakes[50]

a game engine. AI-based technologies have enabled digital characters and audiences to co-exist and interact.[41] Avatar creation has also been employed to enhance virtual assistants,[42] e.g., using proprietary photoreal AI face synthesis technology (Nagano et al. 2018). Facebook Reality Labs have employed ML-AI techniques to animate realistic digital humans, called Codec Avatars, in real time using GAN-based style transfer and using a VAE to extract avatar parameters (Wei et al. 2019). AI is also employed to up-sample frame rate in animation (Siyao et al. 2021).

3.1.6 Augmented, virtual and mixed reality (VR, AR, MR)

AR and VR use computer technologies to create a fully simulated environment or one that is real but augmented with virtual entities. AR expands the physical world with digital layers via mobile phones, tablets or head mounted displays, while VR takes the user into immersive experiences via a headset with a 3D display that isolates the viewer (at least in an audio-visual sense) from the physical world (Milgram et al. 1995).

[41] https://cubicmotion.com/persona/.
[42] https://www.pinscreen.com/.

Significant predictions have been made about the growth of AR and VR markets in recent years but these have not realised yet.[43] This is due to many factors including equipment cost, available content and the physiological effects of 'immersion' (particularly over extended time periods) due to conflicting sensory interactions (Ng et al. 2020). VR can be used to simulate a real workspace for training workers for the sake of safety and to prevent the real-world consequences of failure (Laver et al. 2017). In the healthcare industry, VR is being increasingly used in various sectors, ranging from surgical simulation to physical therapy (Keswani et al. 2020).

Gaming is often cited as a major market for VR, along with related areas such as previsualisation of designs or creative productions (e.g., in building, architecture and filmmaking). A good list of VR games can be found in many article.[44] Deep learning technologies have been exploited in many aspects of gaming, for example in VR/AR game design (Zhang 2020) and emotion detection while using VR to improve the user's immersive experience (Quesnel et al. 2018). More recently AI gaming methods have been extended into the area of virtual production, where the tools are scaled to produce dynamic virtual environments for filmmaking

AR perhaps has more early potential for growth than VR and uses have been developed in education and to create shared information, work or design spaces, where it can provide added 3D realism for the users interacting in the space (Palmarini et al. 2018). AR has also gained interest in augmenting experiences in movie and theatre settings.[45] A review of current and future trends of AR and VR systems can be found in Bastug et al. (2017).

MR combines the real world with digital elements (or the virtual world) (Milgram and Kishino 1994). It allows us to interact with objects and environments in both the real and virtual world by using touch technology and other sensory interfaces, to merge reality and imagination and to provide more engaging experiences. Examples of MR applications include the 'MR Sales Gallery' used by large real estate developers.[46] It is a virtual sample room that simulates the environment for customers to experience the atmosphere of an interactive residential project. The growth of VR, AR and MR technologies is described by Immerse UK in their recent report on the immersive economy in the UK 2019.[47] Extended reality (XR) is a newer technology that combines VR, AR and MR with internet connectivity, which opens further opportunities across industry, education, defence, health, tourism and entertainment (Chuah 2018).

An immersive experience with VR or MR requires good quality, high-resolution, animated worlds or 360-degree video content (Ozcinar and Smolic 2018). This poses new problems for data compression and visual quality assessment, which are the subject of increased research activity currently (Xu et al. 2020). AI technologies have been employed to make AR/VR/MR/XR content more exciting and realistic, to robustly track and localize objects and users in the environment. For example, automatic map reading using image-based localization (Panphattarasap and Calway 2018), and gaze estimation (Anantrasirichai

[43] https://www.marketresearchfuture.com/reports/augmented-reality-virtual-reality-market-6884.

[44] https://www.forbes.com/sites/jessedamiani/2020/01/15/the-top-50-vr-games-of-2019/?sh=42279941322d.

[45] https://www.factor-tech.com/feature/lifting-the-curtain-on-augmented-reality-how-ar-is-bringing-theatre-into-the-future/.

[46] https://dynamics.microsoft.com/en-gb/mixed-reality/overview/.

[47] https://www.immerseuk.org/wp-content/uploads/2019/11/The-Immersive-Economy-in-the-UK-Report-2019.pdf.

et al. 2016; Soccini 2017). Oculus Insight, by Facebook, uses visual-inertial SLAM (simultaneous localization and mapping) to generate real-time maps and position tracking.[48] More sophisticated approaches, such as Neural Topological SLAM, leverage semantics and geometric information to improve long-horizon navigation (Chaplot et al. 2020). Combining audio and visual sensors can further improve navigation of egocentric observations in complex 3D environments, which can be done through deep reinforcement learning approach (Chen et al. 2020).

3.1.7 Deepfakes

Manipulations of visual and auditory media, either for amusement or malicious intent, are not new. However, advances in AI and ML methods have taken this to another level, improving their realistism and providing automated processes that make them easier to render. Text generator tools, such as those by OpenAI, can generate coherent paragraphs of text with basic comprehension, translation and summarization but have also been used to create fake news or abusive spam on social media.[49] Deepfake technologies can also create realistic fake videos by replacing some parts of the media with synthetic content. For example, substituting someone's face while hair, body and action remain the same (Fig. 7b[50]). Early research created mouth movement synthesis tools capable of making the subject appear to say something different from the actual narrative, e.g., President Barack Obama is lip-synchronized to a new audio track in Suwajanakorn et al. (2017). More recently, DeepFaceLab (Perov et al. 2020) provided a state-of-the-art tool for face replacement; however manual editing is still required in order to create the most natural appearance. Whole body movements have been generated via learning from a source video to synthesize the positions of arms, legs and body of the target (Chan et al. 2019).

Deep learning approaches to Deepfake generation primarily employ generative neural network architectures, e.g., VAEs (Kietzmann et al. 2020) and GANs (Zakharov et al. 2019). Despite rapid progress in this area, the creation of perfectly natural figures remains challenging; for example deepfake faces often do not blink naturally. Deepfake techniques have been widely used to create pornographic images of celebrities, to cause political distress or social unrest, for purposes of blackmail and to announce fake terrorism events or other disasters. This has resulted in several countries banning non-consensual deepfake content. To counter these often malicious attacks, a number of approaches have been reported and introduced to detect fake digital content (Güera and Delp 2018; Hasan and Salah 2019; Li and Lyu 2019).

3.1.8 Content and captions

There are many approaches that attempt to interpret an image or video and then automatically generate captions based on its content (Pu et al. 2016; Xia and Wang 2005; Xu et al. 2017b). This can successfully be achieved through object recognition (see Sect. 3.4); YouTube has provided this function for both video-on-demand and livestream videos.[51]

[48] https://ai.facebook.com/blog/powered-by-ai-oculus-insight/.
[49] https://talktotransformer.com/.
[50] https://www.youtube.com/watch?time_continue=2&v=ANXucrz7Hjs.
[51] https://support.google.com/youtube/answer/6373554?hl=en.

The other way around, AI can also help to generate a new image from text. However, this problem is far more complicated; attempts so far have been based on GANs. Early work by Mansimov et al. (2016) was capable of generating background image content with relevant colors but with blurred foreground details. A conditioning augmentation technique was proposed to stabilize the training process of the conditional GAN, and also to improve the diversity of the generated samples (Zhang et al. 2017). Recent methods with significantly increased complexity are capable of learning to generate an image in an object-wise fashion, leading to more natural-looking results (Li et al. 2019c). However, limitations remain, for example artefacts often appear around object boundaries or inappropriate backgrounds can be produced if the words of the caption are not given in the correct order.

3.2 Information analysis

AI has proven capability to process and adapt to large amounts of training data. It can learn and analyze the characteristics of these data, making it possible to classify content and predict outcomes with high levels of confidence. Example applications include advertising and film analysis, as well as image or video retrieval, for example enabling producers to acquire information, analysts to better market products or journalists to retrieve content relevant to an investigation.

3.2.1 Text categorization

Text categorization is a core application of NLP. This generic text processing task is useful in indexing documents for subsequent retrieval and content analysis (e.g., spam detection, sentiment classification, and topic classification). It can be thought of as the generation of summarised texts from full texts. Traditional techniques for both multi-class and multi-label classifications include decision trees, support vector machines (Kowsari et al. 2019), term frequency–inverse document frequency (Azam and Yao 2012), and extreme learning machine (Rezaei-Ravari et al. 2021). Unsupervised learning with self-organizing maps has also been investigated (Pawar and Gawande 2012). Modern NLP techniques are based on deep learning, where generally the first layer is an embedding layer that converts words to vector representations. Additional CNN layers are then added to extract text features and learn word positions (Johnson and Zhang 2015). RNNs (mostly based on LSTM architectures) have also been concatenated to learn sentences and give prediction outputs (Chen et al. 2017; Gunasekara and Nejadgholi 2018). A category sentence generative adversarial network has also been proposed that combines GAN, RNN and reinforcement learning to enlarge training datasets, which improves performance for sentiment classification (Li et al. 2018b). Recently, an attention layer has been integrated into the network to provide semantic representations in aspect-based sentiment analysis (Truşcă et al. 2020). The artist, Vibeke Sorensen, has applied AI techniques to categorize texts from global social networks such as Twitter into six live emotions and display the 'Mood of the Planet' artistically using six different colors.[52]

[52] http://vibeke.info/mood-of-the-planet/.

3.2.2 Advertisements and film analysis

AI can assist creators in matching content more effectively to their audiences, for example recommending music and movies in a streaming service, like Spotify or Netflix. Learning systems have also been used to characterize and target individual viewers, optimizing the time they spend on advertising (Lacerda et al. 2006). This approach assesses what users look at and how long they spend browsing adverts, participating on social media platforms. In addition, AI can be used to inform how adverts should be presented to help boost their effectiveness, for example by identifying suitable customers and showing the ad at the right time. This normally involves gathering and analysing personal data in order to predict preferences (Golbeck et al. 2011).

Contextualizing social-media conversations can also help advertisers understand how consumers feel about products and to detect fraudulent ad impressions (Ghani et al. 2019). This can be achieved using NLP methods (Young et al. 2018). Recently, an AI-based data analysis tool has been introduced to assist filmmaking companies to develop strategies for how, when and where prospective films should be released (Dodds 2020). The tool employs ML approaches to model the patterns of historical data about film performances associating with the film's content and themes. This is also used in gaming industries, where the behaviour of each player is analyzed so that the company can better understand their style of play and decide when best to approach them to make money.[53]

3.2.3 Content retrieval

Data retrieval is an important component in many creative processes, since producing a new piece of work generally requires undertaking a significant amount of research at the start. Traditional retrieval technologies employ metadata or annotation text (e.g., titles, captions, tags, keywords and descriptions) to the source content (Jeon et al. 2003). The manual annotation process needed to create this metadata is however very time-consuming. AI methods have enabled automatic annotation by supporting the analysis of media based on audio and object recognition and scene understanding (Amato et al. 2017; Wu et al. 2015).

In contrast to traditional concept-based approaches, content-based image retrieval (or query by image content (QBIC)) analyzes the content of an image rather than its metadata. A reverse image search technique (one of the techniques Google Images uses[54]) extracts low-level features from an input image, such as points, lines, shapes, colors and textures. The query system then searches for related images by matching these features within the search space. Modern image retrieval methods often employ deep learning techniques, enabling image to image searching by extracting low-level features and then combining these to form semantic representations of the reference image that can be used as the basis of a search (Wan et al. 2014). For example, when a user uploads an image of a dog to Google Images, the search engine will return the dog breed, show similar websites by searching with this keyword, and also show selected images that are visually similar to that dog, e.g., with similar colors and background. These techniques have been further improved by exploiting features at local, regional and global image levels (Gordo et al. 2016). GAN

[53] https://www.bloomberg.com/news/articles/2017-10-23/game-makers-tap-ai-to-profile-each-player-and-keep-them-hooked.

[54] https://images.google.com/.

approaches are also popular, associated with learning-based hashing which was proposed for scalable image retrieval (Song et al. 2018a). Video retrieval can be more challenging due to the requirement for understanding activities, interactions between objects and unknown context; RNNs have provided a natural extension that supports the extraction of sequential behaviour in this case (Jabeen et al. 2018).

Music information retrieval extracts features of sound, and then converts these to a meaningful representation suitable for a query engine. Several methods for this have been reported, including automatic tagging, query by humming, search by sound and acoustic fingerprinting (Kaminskas and Ricci 2012).

3.2.4 Recommendation services

A recommendation engine is a system that suggests products, services, information to users based on analysis of data. For example, a music curator creates a soundtrack or a playlist that has songs with similar mood and tone, bringing related content to the user. Curation tools, capable of searching large databases and creating recommendation shortlists, have become popular because they can save time, elevate brand visibility and increase connection to the audience. The techniques used in recommendation systems generally fall into three categories: (i) content-based filtering, which uses a single user's data, (ii) collaborative filtering, the most prominent approach, that derives suggestions from many other users, and (iii) knowledge-based system, based on specific queries made by the user, which is generally employed in complex domains, where the first two cannot be applied. The approach can be hybrid; for instance where content-based filtering exploits individual metadata and collaborative filtering finds overlaps between user playlists. Such systems build a profile of what the users listen to or watch, and then look at what other people who have similar profiles listen to or watch. ESPN and Netflix have partnered with Spotify to curate playlists from the documentary 'The Last Dance'. Spotify has created music and podcast playlists that viewers can check out after watching the show.[55]

Content summarization is a fundamental tool that can support recommendation services. Text categorization approaches extract important content from a document into key indices (see Sect. 3.2.1). RNN-based models incorporating attention models have been employed to successfully generate a summary in the form of an abstract (Rush et al. 2015), short paragraph (See et al. 2017) or a personalized sentence (Li et al. 2019a). The gaze behavior of an individual viewer has also been included for personalised text summarization (Yi et al. 2020). The personalized identification of key frames and start points in a video has also been framed as an optimization problem in Chen et al. (2014). ML approaches have been developed to perform content-based recommendations. Multimodal features of text, audio, image, and video content are extracted and used to seek similar content in Deldjoo et al. (2018). This task is relevant to content retrieval, as discussed in Sect. 3.2.3. A detailed review of deep learning for recommendation systems can be found in Batmaz et al. (2019).

3.2.5 Intelligent assistants

Intelligent Assistants employ a combination of AI tools, including many of those mentioned above, in the form of a software agent that can perform tasks or services for an

[55] https://open.spotify.com/show/3ViwFAdff2YaXPygfUuv51.

individual. These virtual agents can access information via digital channels to answer questions relating to, for example, weather forecasts, news items or encyclopaedic enquiries. They can recommend songs, movies and places, as well as suggest routes. They can also manage personal schedules, emails, and reminders. The communication can be in the form of text or voice. The AI technologies behind the intelligent assistants are based on sophisticated ML and NLP methods. Examples of current intelligent assistants include Google Assistant,[56] Siri,[57] Amazon Alexa and Nina by Nuance.[58] Similarly, chatbots and other types of virtual assistants are used for marketing, customer service, finding specific content and information gathering (Xu et al. 2017a).

3.3 Content enhancement and post production workflows

It is often the case that original content (whether images, videos, audio or documents) is not fit for the purpose of its target audience. This could be due to noise caused by sensor limitations, the conditions prevailing during acquisition, or degradation over time. AI offers the potential to create assistive intelligent tools that improve both quality and management, particularly for mass-produced content.

3.3.1 Contrast enhancement

The human visual system employs many opponent processes, both in the retina and visual cortex, that rely heavily on differences in color, luminance or motion to trigger salient reactions (Bull and Zhang 2021). Contrast is the difference in luminance and/or color that makes an object distinguishable, and this is an important factor in any subjective evaluation of image quality. Low contrast images exhibit a narrow range of tones and can therefore appear flat or dull. Non-parametric methods for contrast enhancement involve histogram equalisation which spans the intensity of an image between its bit depth limits from 0 to a maximum value (e.g., 255 for 8 bits/pixel). Contrast-limited adaptive histogram equalisation (CLAHE) is one example that is commonly used to adjust an histogram and reduce noise amplification (Pizer et al. 1987). Modern methods have further extended performance by exploiting CNNs and autoencoders (Lore et al. 2017), inception modules and residual learning (Tao et al. 2017). Image Enhancement Conditional Generative Adversarial Networks (IE-CGANs) designed to process both visible and infrared images have been proposed by Kuang et al. (2019). Contrast enhancement, along with other methods to be discussed later, suffer from a fundamental lack of data for supervised training because real image pairs with low and high contrast are unavailable (Jiang et al. 2021). Most of these methods therefore train their networks with synthetic data (see Sect. 2.3.1).

3.3.2 Colorization

Colorization is the process that adds or restores color in visual media. This can be useful in coloring archive black and white content, enhancing infrared imagery (e.g., in low-light natural history filming) and also in restoring the color of aged film. A good example is the

[56] https://assistant.google.com/.
[57] https://www.apple.com/siri/.
[58] https://www.nuance.com/omni-channel-customer-engagement/digital/virtual-assistant/nina.html.

recent film "They Shall Not Grow Old" (2018) by Peter Jackson, that colorized (and corrected for speed and jerkiness, added sound and converted to 3D) 90 minutes of footage from World War One. The workflow was based on extensive studies of WW1 equipment and uniforms as a reference point and involved a time-consuming use of post production tools.

The first AI-based techniques for colorization used a CNN with only three convolutional layers to convert a grayscale image into chrominance values and refined them with bilateral filters to generate a natural color image (Cheng et al. 2015). A deeper network, but still only with eight dilated convolutional layers, was proposed a year later (Zhang et al. 2016). This network captured better semantics, resulting in an improvement on images with distinct foreground objects. Encoder-decoder networks are employed in Xu et al. (2020).

Colorization remains a challenging problem for AI as recognized in the recent Challenge in Computer Vision and Pattern Recognition Workshops (CVPRW) in 2019 (Nah et al. 2019). Six teams competed and all of them employed deep learning methods. Most of the methods adopted an encoder-decoder or a structure based on U-Net (Ronneberger et al. 2015). The deep residual net (NesNet) architecture (He et al. 2016) and the dense net (DenseNet) architecture (Huang et al. 2017) have both demonstrated effective conversion of gray scale to natural-looking color images. More complex architectures have been developed based on GAN structures (Zhang et al. 2019), for example DeOldify and NoGAN (Antic 2020). The latter model was shown to reduce temporal color flickering on the video sequence, which is a common problem when enhancing colors on an individual frame by frame basis. Infrared images have also been converted to natural color images using CNNs (e.g., Limmer and Lensch 2016) (Fig. 8a) and GANs (e.g., Kuang et al. 2020; Suarez et al. 2017).

3.3.3 Upscaling imagery: super-resolution methods

Super-resolution (SR) approaches have gained popularity in recent years, enabling the upsampling of images and video spatially or temporally. This is useful for up-converting legacy content for compatibility with modern formats and displays. SR methods increase the resolution (or sample rate) of a low-resolution (LR) image (Fig. 8b) or video. In the case of video sequences, successive frames can, for example, be employed to construct a single high-resolution (HR) frame. Although the basic concept of the SR algorithm is quite simple, there are many problems related to perceptual quality and restriction of available data. For example, the LR video may be aliased and exhibit sub-pixel shifts between frames and hence some points in the HR frame do not correspond to any information from the LR frames.

With deep learning-based technologies, the LR and HR images are matched and used for training architectures such as CNNs, to provide high quality upscaling potentially using only a single LR image (Dong et al. 2014). Sub-pixel convolution layers can be introduced to improve fine details in the image, as reported by Shi et al.. Residual learning and generative models are also employed, (e.g., Kim et al. 2016; Tai et al. 2017). A generative model with a VGG-based[59] perceptual loss function has been shown to significantly improve quality and sharpness when used with the SRGAN by Ledig et al. (2017). Wang et al. (2018)

[59] VGG is a popular CNN, originally developed for object recognition by the Visual Geometry Group at the University of Oxford (Simonyan and Zisserman 2015). See Sect. 2.3.2 for more detail.

Fig. 8 Image enhancement. **a** Colorization for infrared image (Limmer and Lensch 2016). **b** Super-resolution (Ledig et al. 2017)

proposed a progressive multi-scale GAN for perceptual enhancement, where pyramidal decomposition is combined with a DenseNet architecture (Huang et al. 2017). The above techniques seek to learn implicit redundancy that is present in natural data to recover missing HR information from a single LR instance. For single image SR, the review by Yang et al. (2019) suggests that methods such as EnhanceNet (Sajjadi et al. 2017) and SRGAN (Ledig et al. 2017), that achieve high subjective quality with good sharpness and textural detail, cannot simultaneously achieve low distortion loss (e.g., mean absolute error (MAE) or peak signal-to-noise-ratio (PSNR)). A comprehensive survey of image SR is provided by Wang et al. (2020b). This observes that more complex networks generally produce better PSNR results and that most state-of-the-art methods are based on residual learning and use ℓ_1 as one of training losses (e.g., Dai et al. 2019; Zhang et al. 2018a).

When applied to video sequences, super-resolution methods can exploit temporal correlations across frames as well as local spatial correlations within them. Early contributions applying deep learning to achieve video SR gathered multiple frames into a 3D volume which formed the input to a CNN (Kappeler et al. 2016). Later work exploited temporal correlation via a motion compensation process before concatenating multiple warped frames into a 3D volume (Caballero et al. 2017) using a recurrent architecture (Huang et al. 2015). The framework proposed by Liu et al. (2018a) upscales each frame before applying another network for motion compensation. The original target frame is fed, along with its neighbouring frames, into intermediate layers of the CNN to perform inter-frame motion compensation during feature extraction (Haris et al. 2019). EDVR (Wang et al. 2019), the winner of the NTIRE19 video restoration and enhancement challenges in 2019,[60] employs a deformable convolutional network (Dai et al. 2017) to align two successive frames. Deformable convolution is also employed in DNLN (Deformable Non-Local Network) (Wang et al. 2019a). At the time of writing, EDVR (Wang et al. 2019) and DNLN (Wang et al. 2019a) are reported to outperform other methods for video SR, followed by the method of Haris et al. (2019). This suggests that deformable convolution plays an important role in overcoming inter-frame misalignment, producing sharp textural details.

3.3.4 Restoration

The quality of a signal can often be reduced due to distortion or damage. This could be due to environmental conditions during acquisition (low light, atmospheric distortions

[60] https://data.vision.ee.ethz.ch/cvl/ntire19/.

Fig. 9 Restoration for **a** deblurring (Zhang et al. 2018), **b** denoising with DnCNN (Zhang et al. 2017), and **c** turbulence mitigation (Anantrasirichai et al. 2013). Left and right are the original degraded images and the restored images respectively

or high motion), sensor characteristics (quantization due to limited resolution or bit-depth or electronic noise in the sensor itself) or ageing of the original medium such as tape of film. The general degradation model can be written as $I_{obs} = h * I_{ideal} + n$, where I_{obs} is an observed (distorted) version of the ideal signal I_{ideal}, h is the degradation operator, $*$ represents convolution, and n is noise. The restoration process tries to reconstruct I_{ideal} from I_{obs}. h and n are values or functions that are dependent on the application. Signal restoration can be addressed as an inverse problem and deep learning techniques have been employed to solve it. Below we divide restoration into four classes that relate to work in the creative industries with examples illustrated in Fig. 9. Further details of deep learning for inverse problem solving can be found in Lucas et al. (2018).

3.3.4.1 Deblurring Images can be distorted by blur, due to poor camera focus or camera or subject motion. Blur-removal is an ill-posed problem represented by a point spread function (PSF) h, which is generally unknown. Deblurring methods sharpen an image to increase subjective quality, and also to assist subsequent operations such as optical character recognition (OCR) (Hradis et al. 2015) and object detection (Kupyn et al. 2018). Early work in this area analyzed the statistics of the image and attempted to model physical image and camera properties (Biemond et al. 1990). More sophisticated algorithms such as blind deconvolution (BD), attempt to restore the image and the PSF simultaneously (Jia 2007; Krishnan et al. 2011). These methods however assume a space-invariant PSF and the process generally involves several iterations.

As described by the image degradation model, the PSF (h) is related to the target image via a convolution operation. CNNs are therefore inherently applicable for solving blur problems (Schuler et al. 2016). Deblurring techniques based on CNNs (Nah et al. 2017) and GANs (Kupyn et al. 2018) usually employ residual blocks, where skip connections are inserted every two convolution layers (He et al. 2016). Deblurring an image from coarse-to-fine scales is proposed in Tao et al. (2018), where the outputs are upscaled and are fed back to the encoder-decoder structure. The high-level features of each iteration are linked in a recurrent manner, leading to a recursive process of learning sharp images from blurred ones. Nested skip connections were introduced by Gao et al. (2019), where feature maps from multiple convolution layers are merged before applying them to the next convolution layer (in contrast to the residual block approach where one feature map is merged at the next input). This more complicated architecture improves information flow and results in sharper images with fewer ghosting artefacts compared to previous methods.

In the case of video sequences, deblurring can benefit from the abundant information present across neighbouring frames. The DeBlurNet model (Su et al. 2017) takes a stack of nearby frames as input and uses synthetic motion blur to generate a training dataset. A Spatio-temporal recurrent network exploiting a dynamic temporal blending network is proposed by Hyun Kim et al. (2017). Zhang et al. (2018) have concatenated an encoder, recurrent network and decoder to mitigate motion blur. Recently a recurrent network with iterative updating of the hidden state was trained using a regularization process to create sharp images with fewer ringing artefacts (Nah et al. 2019), denoted as IFI-RNN. A Spatio-Temporal Filter Adaptive Network (STFAN) has been proposed (Zhou et al. 2019), where the convolutional kernel is acquired from the feature values in a spatially varying manner. IFI-RNN and STFAN produce comparable results and hitherto achieve the best performances in terms of both subjective and objective quality measurements [the average PSNRs of both methods are higher than that of Hyun Kim et al. (2017) by up to 3 dB].

3.3.4.2 Denoising Noise can be introduced from various sources during signal acquisition, recording and processing, and is normally attributed to sensor limitations when operating under extreme conditions. It is generally characterized in terms of whether it is additive, multiplicative, impulsive or signal dependent, and in terms of its statistical properties. Not only visually distracting, but noise can also affect the performance of detection, classification and tracking tools. Denoising nodes are therefore commonplace in post production workflows, especially for challenging low light natural history content (Anantrasirichai et al. 2020a). In addition, noise can reduce the efficiency of video compression algorithms, since the encoder allocates wasted bits to represent noise rather than signal, especially at low compression levels. This is the reason that film-grain noise suppression tools are employed in certain modern video codecs (Such as AV1) prior to encoding by streaming and broadcasting organisations.

The simplest noise reduction technique is weighted averaging, performed spatially and/or temporally as a sliding window, also known as a moving average filter (Yahya et al. 2016). More sophisticated methods however perform significantly better and are able to adapt to change noise statistics. These include adaptive spatio-temporal smoothing through anisotropic filtering (Malm et al. 2007), nonlocal transform-domain group filtering (Maggioni et al. 2012), Kalman-bilateral mixture model (Zuo et al. 2013), and spatio-temporal patch-based filtering (Buades and Duran 2019). Prior to the introduction of deep neural network denoising, methods such as BM3D (block matching 3-D) (Dabov et al. 2007) represented the state of the art in denoising performance.

Recent advances in denoising have almost entirely been based on deep learning approaches and these now represent the state of the art. RNNs have been employed successfully to remove noise in audio (Maas et al. 2012; Zhang et al. 2018b). A residual noise map is estimated in the Denoising Convolutional Neural Network (DnCNN) method (Zhang et al. 2017) for image based denoising, and for video based denoising, a spatial and temporal network are concatenated (Claus and van Gemert 2019) where the latter handles brightness changes and temporal inconsistencies. FFDNet is a modified form of DnCNN that works on reversibly downsampled subimages (Zhang et al. 2018). Liu et al. (2018b) developed MWCNN; a similar system that integrates multiscale wavelet transforms within the network to replace max pooling layers in order to better retain visual information. This integrated a wavelet/CNN denoising system and currently provides the state-of-the-art performance for Additive Gaussian White Noise (AGWN). VNLnet combines a non-local patch search module with DnCNN. The first part extracts features, while

the latter mitigates the remaining noise (Davy et al. 2019). Zhao et al. (2019a) proposed a simple and shallow network, SDNet, uses six convolution layers with some skip connection to create a hierarchy of residual blocks. TOFlow (Xue et al. 2019) offers an end-to-end trainable convolutional network that performs motion analysis and video processing simultaneously. GANs have been employed to estimate a noise distribution which is subsequently used to augment clean data for training CNN-based denoising networks (such as DnCNN) (Chen et al. 2018b). GANs for denoising data have been proposed for medical imaging (Yang et al. 2018), but they are not popular in the natural image domain due to the limited data resolution of current GANs. However, CycleGAN has recently been modified to attempt denoising and enhancing low-light ultra-high-definition (UHD) videos using a patch-based strategy (Anantrasirichai and Bull 2021).

Recently, the Noise2Noise algorithm has shown that it is possible to train a denoising network without clean data, under the assumption that the data is corrupted by zero-mean noise (Lehtinen et al. 2018). The training pair of input and output images are both noisy and the network learns to minimize the loss function by solving the point estimation problem separately for each input sample. However, this algorithm is sensitive to the loss function used, which can significantly influence the performance of the model. Another algorithm, Noise2Void (Krull et al. 2019), employs a novel blind-spot network that does not include the current pixel in the convolution. The network is trained using the noisy patches as input and output within the same noisy patch. It achieves comparable performance to Noise2Noise but allows the network to learn noise characteristics in a single image.

NTIRE 2020 held a denoising grand challenge within the IEEE CVPR conference that compared many contemporary high performing ML denoising methods on real images (Abdelhamed et al. 2020). The best competing teams employed a variety of techniques using variants on CNN architectures such as U-Net (Ronneberger et al. 2015), ResNet (He et al. 2016) and DenseNet (Huang et al. 2017), together with ℓ_1 loss functions and ensemble processing including flips and rotations. The survey by Tian et al. (2020) states that SDNet (Zhao et al. 2019a) achieves the best results on ISO noise, and FFDNet (Zhang et al. 2018) offers the best denoising performance overall, including Gaussian noise and spatially variant noise (non-uniform noise levels).

Neural networks have also been used for other aspects of image denoising: Chen et al. (2018a) have developed specific low light denoising methods using CNN-based methods; Lempitsky et al. (2018) have developed a deep learning prior that can be used to denoise images without access to training data; and Brooks et al. (2019) have developed specific neural networks to denoise real images through 'unprocessing', i.e. they re-generate raw captured images by inverting the processing stages in a camera to form a supervised training system for raw images.

3.3.4.3 Dehazing In certain situations, fog, haze, smoke and mist can create mood in an image or video. In other cases, they are considered as distortions that reduce contrast, increase brightness and lower color fidelity. Further problems can be caused by condensation forming on the camera lens. The degradation model can be represented as: $I_{obs} = I_{ideal}t + A(1-t)$ where A is atmospheric light and t is medium transmission. The transmission t is estimated using a dark channel prior based on the observation that the lowest value of each color channel of haze-free images is close to zero (He et al. 2011). Berman et al. (2016), the true colors are recovered based on the assumption that an image can be faithfully represented with just a few hundred distinct colors. The authors showed that tight color clusters change because of haze and form lines in RGB space enabling them to be readjusted. The scene

radiance (I_{ideal}) is attenuated exponentially with depth so some work has included an estimate of the depth map corresponding to each pixel in the image (Kopf et al. 2008). CNNs are employed to estimate transmission t and dark channel by Yang and Sun (2018). CycleDehazing (Engin et al. 2018) is used to enhance GAN architecture in CycleGAN (Zhu et al. 2017). This formulation combines cycle-consistency loss (see Sect. 3.1.4) and perceptual loss (see Sect. 2.3.2) in order to improve the quality of textural information recovery and generate visually better haze-free images (Engin et al. 2018). A comprehensive study and an evaluation of existing single-image dehazing CNN-based algorithms are reported by Li et al. (2019). It concludes that DehazeNet (Cai et al. 2016) performs best in terms of perceptual loss, MSCNN (Tang et al. 2019) offers the best subjective quality and superior detection performance on real hazy images, and AOD-Net (Li et al. 2017) is the most efficient.

A related application is underwater photography (Li et al. 2016) as commonly used in natural history filmmaking. CNNs are employed to estimate the corresponding transmission map or ambient light of an underwater hazy image in Shin et al. (2016). More complicated structures merging U-Net, multi-scale estimation, and incorporating cross layer connections to produce even better results are reported by Hu et al. (2018).

3.3.4.4 Mitigating atmospheric turbulence When the temperature difference between the ground and the air increases, the air layers move upwards rapidly, leading to a change in the interference pattern of the light refraction. This is generally observed as a combination of blur, ripple and intensity fluctuations in the scene. Restoring a scene distorted by atmospheric turbulence is a challenging problem. The effect, which is caused by random, spatially varying, perturbations, makes a model-based solution difficult and, in most cases, impractical. Traditional methods have involved frame selection, image registration, image fusion, phase alignment and image deblurring (Anantrasirichai et al. 2013; Xie et al. 2016; Zhu and Milanfar 2013). Removing the turbulence distortion from a video containing moving objects is very challenging, as generally multiple frames are used and they are needed to be aligned. Temporal filtering with local weights determined from optical flow is employed to address this by Anantrasirichai et al. (2018). However, artefacts in the transition areas between foreground and background regions can remain. Removing atmospheric turbulence based on single image processing is proposed using ML by Gao et al. (2019). Deep learning techniques to solve this problem are still in their early stages. However, one method reported employs a CNN to support deblurring (Nieuwenhuizen and Schutte 2019) and another employs multiple frames using a GAN architecture (Chak et al. 2018). This however appears only to work well for static scenes.

3.3.5 Inpainting

Inpainting is the process of estimating lost or damaged parts of an image or a video. Example applications for this approach include the repair of damage caused by cracks, scratches, dust or spots on film or chemical damage resulting in image degradation. Similar problems arise due to data loss during transmission across packet networks. Related applications include the removal of unwanted foreground objects or regions of an image and video; in this case the occluded background that is revealed must be estimated. An example of inpainting is shown in Fig. 10. In digital photography and video editing, perhaps the most

Fig. 10 Example of inpainting, (left-right) original image, masking and inpainted image

widely used tool is Adobe Photoshop,[61] where inpainting is achieved using content-aware interpolation by analysing the entire image to find the best detail to intelligently replace the damaged area.

Recently AI technologies have been reported that model the missing parts of an image using content in proximity to the damage, as well as global information to assist extracting semantic meaning. Xie et al. (2012) combine sparse coding with deep neural networks pre-trained with denoising auto-encoders. Dilated convolutions are employed in two concatenated networks for spatial reconstruction in the coarse and fine details (Yu et al. 2018). Some methods allow users to interact with the process, for example inputting information such as strong edges to guide the solution and produce better results. An example of this image inpainting with user-guided free-form is given by Yu et al. (2019). Gated convolution is used to learn the soft mask automatically from the data and the content is then generated using both low-level features and extracted semantic meaning. Chang et al. (2019) extend the work by Yu et al. (2019) to video sequences using a GAN architecture. Video Inpainting, VINet, as reported by Kim et al. (2019) offers the ability to remove moving objects and replace them with content aggregated from both spatial and temporal information using CNNs and recurrent feedback. Black et al. (2020) evaluated state-of-the-art methods by comparing performance based on the classification and retrieval of fixed images. They reported that DFNet (Hong et al. 2019), based on U-Net (Ronneberger et al. 2015) adding fusion blocks in the decoding layers, outperformed other methods over a wide range of missing pixels.

3.3.6 Visual special effects (VFX)

Closely related to animation, the use of ML-based AI in VFX has increased rapidly in recent years. Examples include BBC's His Dark Materials and Avengers Endgame (Marvel).[62] These both use a combination of physics models with data driven results from AI algorithms to create high fidelity and photorealistic 3D animations, simulations and renderings. ML-based tools transform the actor's face into the film's character using head-mounted cameras and facial tracking markers. With ML-based AI, a single image can be turned into a photorealistic and fully-clothed production-level 3D avatar in real-time (Hu

[61] https://www.adobe.com/products/photoshop/content-aware-fill.html.
[62] https://blogs.nvidia.com/blog/2020/02/07/ai-vfx-oscars/.

et al. 2017). Other techniques related to VFX can be found in Sect. 3.1 (e.g., style transfer and deepfakes), Sect. 3.3 (e.g., colorization and super-resolution) and Sect. 3.4 (e.g tracking and 3D rendering). AI techniques[63] are increasingly being employed to reduce the human resources needed for certain labour-intensive or repetitive tasks such as matchmove, tracking, rotoscoping, compositing and animation (Barber et al. 2016; Torrejon et al. 2020).

3.4 Information extraction and enhancement

AI methods based on deep learning have demonstrated significant success in recognizing and extracting information from data. They are well suited to this task since successive convolutional layers efficiently perform statistical analysis from low to high level, progressively abstracting meaningful and representative features. Once information is extracted from a signal, it is frequently desirable to enhance it or transform it in some way. This may, for example, make an image more readily interpretable through modality fusion, or translate actions from a real animal to an animation. This section investigates how AI methods can utilize explicit information extracted from images and videos to construct such information and reuse it in new directions or new forms.

3.4.1 Segmentation

Segmentation methods are widely employed to partition a signal (typically an image or video) into a form that is semantically more meaningful and easier to analyze or track. The resulting segmentation map indicates the locations and boundaries of semantic objects or regions with parametric homogeneity in an image. Pixels within a region could therefore represent an identifiable object and/or have shared characteristics, such as color, intensity, and texture. Segmentation boundaries indicate the shape of objects and this, together with other parameters, can be used to identify what the object is. Segmentation can be used as a tool in the creative process, for example assisting with rotoscoping, masking, cropping and for merging objects from different sources into a new picture. Segmentation, in the case of video content, also enables the user to change the object or region's characteristics over time, for example through blurring, color grading or replacement.[64]

Classification systems can be built on top of segmentation in order to detect or identify objects in a scene (Fig. 11a). This can be compared with the way that humans view a photograph or video, to spot people or other objects, to interpret visual details or to interpret the scene. Since different objects or regions will differ to some degree in terms of the parameters that characterize them, we can train a machine to perform a similar process, providing an understanding of what the image or video contains and activities in the scene. This can in turn support classification, cataloguing and data retrieval. Semantic segmentation classifies all pixels in an image into predefined categories, implying that it processes segmentation and classification simultaneously. The first deep learning approach to semantic segmentation employed a fully convolutional network (Long et al. 2015). In the same year, the encoder-decoder model in Noh et al. (2015) and the U-Net architecture (Ronneberger et al. 2015) were introduced. Following these, a number of modified networks

[63] https://www.vfxvoice.com/the-new-artificial-intelligence-frontier-of-vfx/.
[64] https://support.zoom.us/hc/en-us/articles/210707503-Virtual-Background.

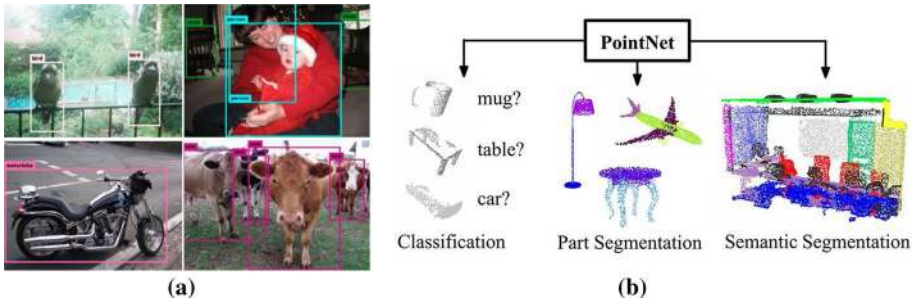

Fig. 11 Segmentation and recognition. **a** Object recognition (Kim et al. 2020a). **b** 3D semantic segmentation (Qi et al. 2017)

based on them architectures have been reported Asgari Taghanaki et al. (2021). GANs have also been employed for the purpose of image translation, in this case to translate a natural image into a segmentation map (Isola et al. 2017). The semantic segmentation approach has also been applied to point cloud data to classify and segment 3D scenes, e.g., Fig. 11b (Qi et al. 2017).

3.4.2 Recognition

Object recognition has been one of the most common targets for AI in recent years, driven by the complexity of the task but also by the huge amount of labeled imagery available for training deep networks. The performance in terms of mean Average Precision (mAP) for detecting 200 classes has increased more than 300% over the last 5 years (Liu et al. 2020). The Mask R-CNN approach (He et al. 2017) has gained popularity due to its ability to separate different objects in an image or a video giving their bounding boxes, classes and pixel-level masks, as demonstrated by Ren et al. (2017). Feature Pyramid Network (FPN) is also a popular backbone for object detection (Lin et al. 2017). An in-depth review of object recognition using deep learning can be found in Zhao et al. (2019b) and Liu et al. (2020).

YOLO and its variants represent the current state of the art in real-time object detection and tracking (Redmon et al. 2016). A state-of-the-art, real-time object detection system, You Only Look Once (YOLO), works on a frame-by-frame basis and is fast enough to process at typical video rates (currently reported up to 55 fps). YOLO divides an image into regions and predicts bounding boxes using a multi-scale approach and gives probabilities for each region. The latest model, YOLOv4, (Bochkovskiy et al. 2020), concatenates YOLOv3 (Redmon and Farhadi 2018) with a CNN that is 53 layers deep, with SPP-blocks (He et al. 2015) or SAM-blocks (Woo et al. 2018) and a multi-scale CNN backbone. YOLOv4 offers real-time computation and high precision [up to 66 mAP on Microsoft's COCO object dataset (Lin et al. 2014)].

On the PASCAL visual object classes (VOC) Challenge datasets (Everingham et al. 2012), YOLOv3 is the leader of object detection on the VOC2010 dataset a with mAP of 80.8% (YOLOv4 performance on this dataset had not been reported at the time of writing)

and NAS-Yolo is the best for VOC2012 dataset with a mAP of 86.5%[65] (the VOC2012 dataset has a larger number of segmentations than VOC2010). NAS-Yolo (Yang et al. 2020b) employs Neural Architecture Search (NAS) and reinforcement learning to find the best augmentation policies for the target. In the PASCAL VOC Challenge for semantic segmentation, FlatteNet (Cai and Pu 2019) and FDNet (Zhen et al. 2019) lead the field achieving the mAP of 84.3 and 84.0% on VOC2012 data, respectively. FlatteNet integrates fully convolutional network with pixel-wise visual descriptors converting from feature maps. FDNet links all feature maps from the encoder to each input of the decoder leading to really dense network and precise segmentation. On the Microsoft COCO object dataset, MegDetV2 (Li et al. 2019d) ranks first on both the detection leaderboard and the semantic segmentation leaderboard. MegDetV2 combines ResNet with FPN and uses deformable convolution to train the end-to-end network with large mini-batches.

Recognition of speech and music has also been successfully achieved using deep learning methods. Mobile phone apps that capture a few seconds of sound or music, such as Shazam,[66] characterize songs based on an audio fingerprint using a spectrogram (a time-frequency graph) that is used to search for a matching fingerprint in a database. Houndify by SoundHound[67] exploits speech recognition and searches content across the internet. This technology also provides voice interaction for in-car systems. Google proposed a full visual-speech recognition system that maps videos of lips to sequences of words using spatiotemporal CNNs and LSTMs (Shillingford et al. 2019).

Emotion recognition has also been studied for over a decade. AI methods have been used to learn, interpret and respond to human emotion, via speech (e.g., tone, loudness, and tempo) (Kwon et al. 2003), face detection (e.g., eyebrows, the tip of nose, the corners of mouth) (Ko 2018), and both audio and video (Hossain and Muhammad 2019). Such systems have also been used in security systems and for fraud detection.

A further task, relevant to video content, is action recognition. This involves capturing spatio-temporal context across frames, for example: jumping into a pool, swimming, getting out of the pool. Deep learning has again been extensively exploited in this area, with the first report based on a 3D CNN (Ji et al. 2013). An excellent state-of-the-art review on action recognition can be found in Yao et al. (2019). More recent advances include temporal segment networks (Wang et al. 2016) and temporal binding networks, where the fusion of audio and visual information is employed (Kazakos et al. 2019). EPIC-KITCHENS, is a large dataset focused on egocentric vision that provides audio-visual, non-scripted recordings in native environments (Damen et al. 2018); it has been extensively used to train action recognition systems. Research on sign language recognition is also related to creative applications, since it studies body posture, hand gesture, and face expression, and hence involves segmentation, detection, classification and 3D reconstruction (Jalal et al. 2018; Kratimenos et al. 2020; Adithya and Rajesh 2020). Moreover, visual and linguistic modelling has been combined to enable translation between spoken/written language and continuous sign language videos (Bragg et al. 2019).

[65] http://host.robots.ox.ac.uk:8080/leaderboard/main_bootstrap.php.
[66] https://www.shazam.com/gb/company.
[67] https://www.soundhound.com/.

3.4.3 Salient object detection

Salient object detection (SOD) is a task based on visual attention mechanisms, in which algorithms aim to identify objects or regions that are likely to be the focus of attention. SOD methods can benefit the creative industries in applications such as image editing (Cheng et al. 2010; Mejjati et al. 2020), content interpretation (Rutishauser et al. 2004), egocentric vision (Anantrasirichai et al. 2018), VR (Ozcinar and Smolic 2018), and compression (Gupta et al. 2013). The purpose of SOD differs from fixation detection, which predicts where humans look, but there is a strong correlation between the two (Borji et al. 2019). In general, the SOD process involves two tasks: saliency prediction and segmentation. Recent supervised learning technologies have significantly improved the performance of SOD. Hou et al. (2019) merge multi-level features of a VGG network with fusion and cross-entropy losses. A survey by Wang et al. (2021) reveals that most SOD models employ VGG and ResNet as backbone architectures and train the model with the standard binary cross-entropy loss. More recent work has developed the end-to-end framework with GANs (Wang et al. 2020a) and some works include depth information from RGB-D cameras (Jiang et al. 2020). More details on the recent SOD on RGB-D data can be found in (Zhou et al. 2021). When detecting salient objects in the video, an LSTM module is used to learn saliency shifts (Fan et al. 2019). The SOD approach has also been extended to co-salient object detection (CoSOD), aiming to detect the co-occurring salient objects in multiple images (Fan et al. 2020).

3.4.4 Tracking

Object tracking is the temporal process of locating objects in consecutive video frames. It takes an initial set of object detections (see Sect. 3.4), creates a unique ID for each of these initial detections, and then tracks each of the objects, via their properties, over time. Similar to segmentation, object tracking can support the creative process, particularly in editing. For example, a user can identify and edit a particular area or object in one frame and, by tracking the region, these adjusted parameters can be applied to the rest of the sequence regardless of object motion. Semi-supervised learning is also employed in SiamMask (Wang et al. 2019b) offering the user an interface to define the object of interest and to track it over time.

Similar to object recognition, deep learning has become an effective tool for object tracking, particularly when tracking multiple objects in the video (Liu et al. 2020). Recurrent networks have been integrated with object recognition methods to track the detected objects over time (e.g., Fang 2016; Gordon et al. 2018; Milan et al. 2017). VOT benchmarks (Kristan et al. 2016) have been reported for real-time visual object tracking challenges run in both ICCV and ECCV conferences, and the performance of tracking has been observed to improve year on year. The best performing methods include Re^3 (Gordon et al. 2018) and Siamese-RPN (Li et al. 2018) achieving 150 and 160 fps at the expected overlap of 0.2, respectively. MOTChallenge[68] and KITTI[69] are the most commonly used datasets for training and testing multiple object tracking (MOT). At the time of publishing, ReMOTS (Yang et al. 2020a) is currently the best performer with a mask-based MOT

[68] https://motchallenge.net/.

[69] http://www.cvlibs.net/datasets/kitti/eval_tracking.php.

Original image 1 Original image 2 Fused image Original 2D face Reconstructed 3D face

(a) **(b)**

Fig. 12 Information Enhancement. **a** Multifocal image fusion. **b** 2D to 3D face conversion generated using the algorithm proposed by Jackson et al. (2017)

accuracy of 83.9%. ReMOTS fuses the segmentation results of the Mask R-CNN (He et al. 2017) and a ResNet-101 (He et al. 2016) backbone extended with FPN.

3.4.5 Image fusion

Image fusion provides a mechanism to combine multiple images (or regions therein, or their associated information) into a single representation that has the potential to aid human visual perception and/or subsequent image processing tasks. A fused image (e.g., a combination of IR and visible images) aims to express the salient information from each source image without introducing artefacts or inconsistencies. A number of applications have exploited image fusion to combine complementary information into a single image, where the capability of a single sensor is limited by design or observational constraints. Existing pixel-level fusion schemes range from simple averaging of the pixel values of registered (aligned) images to more complex multiresolution pyramids, sparse methods (Anantrasirichai et al. 2020b) and methods based on complex wavelets (Lewis et al. 2007). Deep learning techniques have been successfully employed in many image fusion applications. An all-in-focus image is created using multiple images of the same scene taken with different focal settings (Liu et al. 2017) (Fig. 12a). Multi-exposure deep fusion is used to create high-dynamic range images by Prabhakar et al. (2017). A review of deep learning for pixel-level image fusion can be found in Liu et al. (2018). Recently, GANs have also been developed for this application (e.g., Ma et al. 2019c), with an example of image blending using a guided mask (e.g., Wu et al. 2019).

The performance of a fusion algorithm is difficult to quantitatively assess as no ground truth exists in the fused domain. Ma et al. (2019b) shows that a guided filtering-based fusion (Li et al. 2013) achieves the best results based on the visual information fidelity (VIF) metric, but proposed that fused images with very low correlation coefficients, measuring the degree of linear correlation between the fused image its source images, also works well compared to subjective assessment.

3.4.6 3D reconstruction and rendering

In the human visual system, a stereopsis process (together with many other visual cues and priors (Bull and Zhang 2021) creates a perception of three-dimensional (3D) depth from the combination of two spatially separated signals received by the visual cortex from our retinas. The fusion of these two slightly different pictures gives the sensation of strong three-dimensionality by matching similarities. To provide stereopsis in machine vision applications, images are captured simultaneously from two cameras with parallel camera

geometry, and an implicit geometric process is used to extract 3D information from these images. This process can be extended using multiple cameras in an array to create a full volumetric representation of an object. This approach is becoming increasingly popular in the creative industries, especially for special effects that create digital humans[70] in high end movies or live performance.

To convert 2D to 3D representations (including 2D+t to 3D), the first step is normally depth estimation, which is performed using stereo or multi-view RGB camera arrays. Consumer RGB-D sensors can also be used for this purpose (Maier et al. 2017). Depth estimation, based on disparity can also be assisted by motion parallax (using a single moving camera), focus, and perspective. For example, motion parallax is learned using a chain of encoder-decoder networks by Ummenhofer et al. (2017). Google Earth has computed topographical information from images captured using aircraft and added texture to create a 3D mesh. As the demands for higher depth accuracy have increased and real-time computation has become feasible, deep learning methods (particularly CNNs) have gained more attention. A number of network architectures have been proposed for stereo configurations, including a pyramid stereo matching network (PSMNet) (Chang and Chen 2018), a stacked hourglass architecture (Newell et al. 2016), a sparse cost volume network (SCV-Net) (Lu et al. 2018), a fast densenet (Anantrasirichai et al. 2021) and a guided aggregation net (GA-Net) (Zhang et al. 2019). On the KITTI Stereo dataset benchmark (Geiger et al. 2012), the team, called LEAStereo from Monash University, ranks 1st at the time of writing (the number of erroneous pixels reported as 1.65%). They exploit neural architecture search (NAS) technique[71] to build the best network designed by another neural network.

3D reconstruction is generally divided into: volumetric, surface-based, and multi-plane representations. Volumetric representations can be achieved by extending the 2D convolutions used in image analysis. Surface-based representations, e.g., meshes, can be more memory-efficient, but are not regular structures and thus do not easily map onto deep learning architectures. The state-of-the-art methods for volumetric and surface-based representations are Pix2Vox (Xie et al. 2019) and AllVPNet (Soltani et al. 2017) reporting an Intersection-over-Union (IoU) measure of 0.71 and 0.83 constructed from 20 views on the ShapeNet dataset benchmark (Chang et al. 2015)). GAN architectures have been used to generate non-rigid surfaces from a monocular image (Shimada et al. 2019). The third type of representation is formed from multiple planes of the scene. It is a trade-off between the first two representations—efficient storage and amenable to training with deep learning. The method in Flynn et al. (2019), developed by Google, achieves view synthesis with learned gradient descent. A review of state-of-the-art 3D reconstruction from images using deep learning can be found in Han et al. (2019).

Recently, low-cost video plus depth (RGB-D) sensors have become widely available. Key challenges related to RGB-D video processing have included synchronisation, alignment and data fusion between multimodal sensors (Malleson et al. 2019). Deep learning approaches have also been used to achieve semantic segmentation, multi-model feature matching and noise reduction for RGB-D information (Zollhöfer et al. 2018). Light field cameras, that capture the intensity and direction of light rays, produce denser data than the RGB-D cameras. Depth information of a scene can be extracted from the displacement of the image array, and 3D rendering has been reported using deep learning approaches in Shi

[70] https://www.dimensionstudio.co/solutions/digital-humans.

[71] http://www.cvlibs.net/datasets/kitti/eval_scene_flow.php?benchmark=stereo.

et al. (2020). Recent state-of-the-art light field methods can be found in the review by Jiang et al. (2020).

3D reconstruction from a single image is an ill-posed problem. However, it is possible with deep learning due to the network's ability to learn semantic meaning (similar to object recognition, described in Sect. 3.4). Using a 2D RGB training image with 3D ground truth, the model can predict what kind of scene and objects are contained in the test image. Deep learning-based methods also provide state-of-the-art performance for generating the corresponding right view from a left view in a stereo pair (Xie et al. 2016), and for converting 2D face images to 3D face reconstructions using CNN-based encoder-decoder architectures (Bulat and Tzimiropoulos 2017; Jackson et al. 2017), autoencoders (Tewari et al. 2020) and GANs (Tian et al. 2018) (Fig. 12b). Creating 3D models of bodies from photographs is the focus of (Kanazawa et al. 2018). Here, a CNN is used to translate a single 2D image of a person into parameters of shape and pose, as well as to estimate camera parameters. This is useful for applications such as virtual modelling of clothes in the fashion industry. A recent method reported by Mescheder et al. (2019) is able to generate a realistic 3D surface from a single image intruding the idea of a continuous decision boundary within the deep neural network classifier. For 2D image to 3D object generation, generative models offer the best performance to date, with the state-of-the-art method, GAL (Jiang et al. 2018), achieving an average IoU of 0.71 on the ShapeNet dataset. The creation of a 3D photograph from 2D images is also possible via tools such as SketchUp[72] and Smoothie-3D.[73] Very recently (Feb 2020), Facebook allowed users to add a 3D effect to all 2D images.[74] They trained a CNN on millions of pairs of public 3D images with their associating depth maps. Their Mesh R-CNN (Gkioxari et al. 2019) leverages the Mask R-CNN approach (He et al. 2017) for object recognition and segmentation to help estimate depth cues. A common limitation when converting a single 2D image to a 3D representation is associated with occluded areas that require spatial interpolation.

AI has also been used to increase the dimensionality of audio signals. Humans have an ability to spatially locate a sound as our brain can sense the differences between arrival times of sounds at the left and the right ears, and between the volumes (interaural level) that the left and the right ears hear. Moreover, our ear flaps distort the sound telling us whether the sound emanates in front of or behind the head. With this knowledge, Gao and Grauman (2019) created binaural audio from a mono signal driven by the subject's visual environment to enrich the perceptual experience of the scene. This framework exploits U-Net to extract audio features, merged with visual features extracted from ResNet to predict the sound for the left and the right channels. Subjective tests indicate that this method can improve realism and the sensation being in a 3D space. Morgado et al. (2018) expand mono audio, recorded using a 360° video camera, to the sound over the full viewing surface of sphere. The process extracts semantic environments from the video with CNNs and then high-level features of vision and audio are combined to generate the sound corresponding to different viewpoints. Vasudevan et al. (2020) also include depth estimation to improve realistic quality of super-resolution sound.

[72] https://www.sketchup.com/plans-and-pricing/sketchup-free.

[73] https://smoothie-3d.com/.

[74] https://ai.facebook.com/blog/powered-by-ai-turning-any-2d-photo-into-3d-using-convolutional-neural-nets/.

3.5 Data compression

Visual information is the primary consumer of communications bandwidth across broadcasting and internet communications. The demand for increased qualities and quantities of visual content is particularly driven by the creative media sector, with increased numbers of users expecting increased quality and new experiences. Cisco predict, in their Video Network Index report, (Barnett et al. 2018) that there will be 4.8 zettabytes (4.8×10^{21} bytes) of global annual internet traffic by 2022—equivalent to all movies ever made crossing global IP networks in 53 seconds. Video will account for 82 percent of all internet traffic by 2022. This will be driven by increased demands for new formats and more immersive experiences with multiple viewpoints, greater interactivity, higher spatial resolutions, frame rates and dynamic range and wider color gamut. This is creating a major tension between available network capacity and required video bit rate. Network operators, content creators and service providers all need to transmit the highest quality video at the lowest bit rate and this can only be achieved through the exploitation of content awareness and perceptual redundancy to enable better video compression.

Traditional image encoding systems (e.g., JPEG) encode a picture without reference to any other frames. This is normally achieved by exploiting spatial redundancy through transform-based decorrelation followed by variable length, quantization and symbol encoding. While video can also be encoded as a series of still images, significantly higher coding gains can be achieved if temporal redundancies are also exploited. This is achieved using inter-frame motion prediction and compensation. In this case the encoder processes the low energy residual signal remaining after prediction, rather than the original frame. A thorough coverage of image and video compression methods is provided by Bull and Zhang (2021).

Deep neural networks have gained popularity for image and video compression in recent years and can achieve consistently greater coding gain than conventional approaches. Deep compression methods are also now starting to be considered as components in mainstream video coding standards such as VVC and AV2. They have been applied to optimize a range of coding tools including intra prediction (Li et al. 2018; Schiopu et al. 2019), motion estimation (Zhao et al. 2019b), transforms (Liu et al. 2018), quantization (Liu et al. 2019), entropy coding (Zhao et al. 2019a) and loop filtering (Lu et al. 2019). Post processing is also commonly applied at the video decoder to reduce various coding artefacts and enhance the visual quality of the reconstructed frames [e.g., (Xue and Su 2019; Zhang et al. 2020)]. Other work has implemented a complete coding framework based on neural networks using end-to-end training and optimisation (Lu et al. 2020). This approach presents a radical departure from conventional coding strategies and, while it is not yet competitive with state-of-the-art conventional video codecs, it holds significant promise for the future.

Perceptually based resampling methods based on SR methods using CNNs and GANs have been introduced recently. Disney Research proposed a deep generative video compression system (Han et al. 2019) that involves downscaling using a VAE and entropy coding via a deep sequential model. ViSTRA2 (Zhang et al. 2019b), exploits adaptation of spatial resolution and effective bit depth, downsampling these parameters at the encoder based on perceptual criteria, and up-sampling at the decoder using a deep convolutional neural network. ViSTRA2 has been integrated with the reference software of both the HEVC (HM 16.20) and VVC (VTM 4.01), and evaluated under the Joint Video

Exploration Team Common Test Conditions using the Random Access configuration. Results show consistent and significant compression gains against HM and VVC based on Bjønegaard Delta measurements, with average BD-rate savings of 12.6% (PSNR) and 19.5% (VMAF) over HM and 5.5% and 8.6% over VTM. This work has been extended to a GAN architecture by Ma et al. (2020a). Recently, Mentzer et al. (2020) optimize a neural compression scheme with a GAN, yielding reconstructions with high perceptual fidelity. Ma et al. (2021) combine several quantitative losses to achieve maximal perceptual video quality when training a relativistic sphere GAN.

Like all deep learning applications, training data is a key factor in compression performance. Research by Ma et al. (2020) has demonstrated the importance of large and diverse datasets when developing CNN-based coding tools. Their BVI-DVC database is publicly available and produces significant improvements in coding gain across a wide range of deep learning networks for coding tools such as loop filtering and post-decoder enhancement. An extensive review of AI for compression can be found in Bull and Zhang (2021) and Ma et al. (2020b).

4 Future challenges for AI in the creative sector

There will always be philosophical and ethical questions relating to the creative capacity, ideas and thought processes, particularly where computers or AI are involved. The debate often focuses on the fundamental difference between humans and machines. In this section we will briefly explore some of these issues and comment on their relevance to and impact on the use of AI in the creative sector.

4.1 Ethical issues, fakes and bias

An AI-based machine can work 'intelligently', providing an impression of understanding but nonetheless performing without 'awareness' of wider context. It can however offer probabilities or predictions of what could happen in the future from several candidates, based on the trained model from an available database. With current technology, AI cannot truly offer broad context, emotion or social relationship. However, it can affect modern human life culturally and societally. UNESCO has specifically commented on the potential impact of AI on culture, education, scientific knowledge, communication and information provision particularly relating to the problems of the digital divide.[75] AI seems to amplify the gap between those who can and those who cannot use new digital technologies, leading to increasing inequality of information access. In the context of the creative industries, UNESCO mentions that collaboration between intelligent algorithms and human creativity may eventually bring important challenges for the rights of artists.

One would expect that the authorship of AI creations resides with those who develop the algorithms that drive the art work. Issues of piracy and originality thus need special attention and careful definition, and deliberate and perhaps unintentional exploitation needs to be addressed. We must be cognizant of how easy AI technologies can be accessed and used in the wrong hands. AI systems are now becoming very competent at creating fake images, videos, conversations, and all manner of content. Against this, as reported in

[75] https://ircai.org/project/preliminary-study-on-the-ethics-of-ai/.

Sect. 3.1.7, there are also other AI-based methods under development that can, with some success, detect these fakes.

The primary learning algorithms for AI are data-driven. This means that, if the data used for training are unevenly distributed or unrepresentative due to human selection criteria or labeling, the results after learning can equally be biased and ultimately judgemental. For example, streaming media services suggest movies that the users may enjoy and these suggestions must not privilege specific works over others. Similarly face recognition or autofocus methods must be trained on a broad range of skin types and facial features to avoid failure for certain ethnic groups or genders. Bias in algorithmic decision-making is also a concern of governments across the world.[76] Well-designed AI systems can not only increase the speed and accuracy with which decisions are made, but they can also reduce human bias in decision-making processes. However, throughout the lifetime of a trained AI system, the complexity of data it processes is likely to grow, so even a network originally trained with balanced data may consequently establish some bias. Periodic retraining may therefore be needed. A review of various sources of bias in ML is provided in Ntoutsi et al. (2020).

Dignum (2018) provide a useful classification of the relationships between ethics and AI, defining three categories: (i) Ethics by Design, methods that ensure ethical behaviour in autonomous systems, (ii) Ethics in Design, methods that support the analysis of the ethical implications of AI systems, and (iii) Ethics for Design, codes and protocols to ensure the integrity of developers and users. A discussion of ethics associated with AI in general can be found in Bostrom and Yudkowsky (2014).

AI can, of course, also be used to help identify and resolve ethical issues. For example, Instagram uses an anti-bullying AI[77] to identify negative comments before they are published and asks users to confirm if they really want to post such messages.

4.2 The human in the loop: AI and creativity

Throughout this review we have recognized and reported on the successes of AI in supporting and enhancing processes within constrained domains where there is good availability of data as a basis for ML. We have seen that AI-based techniques work very well when they are used as tools for information extraction, analysis and enhancement. Deep learning methods that characterize data from low-level features and connect these to extract semantic meaning are well suited to these applications. AI can thus be used with success, to perform tasks that are too difficult for humans or are too time-consuming, such as searching through a large database and examining its data to draw conclusions. Post production workflows will therefore see increased use of AI, including enhanced tools for denoising, colorization, segmentation, rendering and tracking. Motion and volumetric capture methods will benefit from enhanced parameter selection and rendering tools. Virtual production methods and games technologies will see greater convergence and increased reliance on AI methodologies.

In all the above examples, AI tools will not be used in isolation as a simple black box solution. Instead, they must be designed as part of the associated workflow and incorporate

[76] https://www.gov.uk/government/publications/interim-reports-from-the-centre-for-data-ethics-and-innovation/interim-report-review-into-bias-in-algorithmic-decision-making.

[77] https://about.instagram.com/blog/announcements/instagrams-commitment-to-lead-fight-against-online-bullying.

a feedback framework with the human in the loop. For the foreseeable future, humans will need to check the outputs from AI systems, make critical decisions, and feedback 'faults' that will be used to adjust the model. In addition, the interactions between audiences or users and machines are likely to become increasingly common. For example, AI could help to create characters that learn context in location-based storytelling and begin to understand the audience and adapt according to interactions.

Currently, the most effective AI algorithms still rely on supervised learning, where ground truth data readily exist or where humans have labeled the dataset prior to using it for training the model (as described in Sect. 2.3.1). In contrast, truly creative processes do not have pre-defined outcomes that can simply be classed as good or bad. Although many may follow contemporary trends or be in some way derivative, based on known audience preferences, there is no obvious way of measuring the quality of the result in advance. Creativity almost always involves combining ideas, often in an abstract yet coherent way, from different domains or multiple experiences, driven by curiosity and experimentation. Hence, labeling of data for these applications is not straightforward or even possible in many cases. This leads to difficulties in using current ML technologies.

In the context of creating a new artwork, generating low-level features from semantics is a one-to-many relationship, leading to inconsistencies between outputs. For example, when asking a group of artists to draw a cat, the results will all differ in color, shape, size, context and pose. Results of the creative process are thus unlikely to be structured, and hence may not be suitable for use with ML methods. We have previously referred to the potential of generative models, such as GANs, in this respect, but these are not yet sufficiently robust to consistently create results that are realistic or valuable. Also, most GAN-based methods are currently limited to the generation of relatively small images and are prone to artefacts at transitions between foreground and background content. It is clear that significant additional work is needed to extract significant value from AI in this area.

4.3 The future of AI technologies

Research into, and development of, AI-based solutions continue apace. AI is attracting major investments from governments and large international organisations alongside venture capital investments in start-up enterprises. ML algorithms will be the primary driver for most AI systems in the future and AI solutions will, in turn, impact an even wider range of sectors. The pace of AI research has been predicated, not just on innovative algorithms (the basics are not too dissimilar to those published in the 1980s), but also on our ability to generate, access and store massive amounts of data, and on advances in graphics processing architectures and parallel hardware to process these massive amounts of data. New computational solutions such as quantum computing, will likely play an increasing role in this respect (Welser et al. 2018).

In order to produce an original work, such as music or abstract art, it would be beneficial to support increased diversity and context when training AI systems. The quality of the solution in such cases is difficult to define and will inevitably depend on audience preferences and popular contemporary trends. High-dimensional datasets that can represent some of these characteristics will therefore be needed. Furthermore, the loss functions that drive the convergence of the network's internal weights must reflect perceptions rather than simple mathematical differences. Research into such loss functions that better reflect human perception of performance or quality is therefore an area for further research.

ML-based AI algorithms are data-driven; hence how to select and prepare data for creative applications will be key to future developments. Defining, cleaning and organizing bias-free data for creative applications are not straightforward tasks. Because the task of data collection and labeling can be highly resource intensive,labeling services are expected to become more popular in the future. Amazon currently offers a cloud management tool, SageMaker,[78] that uses ML to determine which data in a dataset needs to be labeled by humans, and consequently sends this data to human annotators through its Mechanical Turk system or via third party vendors. This can reduce the resources needed by developers during the key data preparation process. In this or other contexts, AI may converge with blockchain technologies. Blockchains create decentralized, distributed, secure and transparent networks that can be accessed by anyone in public (or private) blockchain networks. Such systems may be a means of trading trusted AI assets, or alternatively AI agents may be trusted to trade other assets (e.g., financial (or creative) across blockchain networks. Recently, Microsoft has tried to improve small ML models hosted on public blockchains and plan to expand to more complex models in the future.[79] Blockchains make it possible to reward participants who help to improve models, while providing a level of trust and security.

As the amount of unlabeled data grows dramatically, unsupervised or self-supervised ML algorithms are prime candidates for underpinning future advancements in the next generation of ML. There exist techniques that employ neural networks to learn statistical distributions of input data and then transfer this to the distribution of the output data (Damodaran et al. 2018; Xu et al. 2019; Zhu et al. 2017). These techniques do not require a precise matching pair between the input and the ground truth, reducing the limitations for a range of applications.

It is clear that current AI methods do not mimic the human brain, or even parts of it, particularly closely. The data driven learning approach with error backpropagation is not apparent in human learning. Humans learn in complex ways that combine genetics, experience and prediction-failure reinforcement. A nice example is provided by Yan LeCun of NYU and Facebook[80] who describes a 4–6 month old baby being shown a picture of a toy floating in space; the baby shows little surprise that this object defies gravity. Showing the same image to the same child at around 9 months produces a very different result, despite the fact that it is very unlikely that the child has been explicitly trained about gravity. It has instead learnt by experience and is capable of transferring its knowledge across a wide range of scenarios never previously experienced. This form of reinforcement and transfer learning holds significant potential for the next generation of ML algorithms, providing much greater generalization and scope for innovation.

Reinforcement Learning generally refers to a goal-oriented approach, which learns how to achieve a complex objective through reinforcement via penalties and rewards based on its decisions over time. Deep Reinforcement Learning (DRL) integrates this approach into a deep network which, with little initialisation and through self-supervision, can achieve extraordinary performance in certain domains. Rather than depend on manual labeling, DRL automatically extracts weak annotation information from the input data, reinforced over several steps. It thus learns the semantic features of the data, which can be transferred

[78] https://docs.aws.amazon.com/sagemaker/latest/dg/sms.html.

[79] https://www.microsoft.com/en-us/research/blog/leveraging-blockchain-to-make-machine-learning-models-more-accessible/.

[80] LeCun credits Emmanuel Dupoux for this example.

to other tasks. DRL algorithms can beat human experts playing video games and the world champions of Go. The state of the art in this area is progressing rapidly and the potential for strong AI, even with ambiguous data in the creative sector is significant. However, this will require major research effort as the human processes that underpin this are not well understood.

5 Concluding remarks

This paper has presented a comprehensive review of current AI technologies and their applications, specifically in the context of the creative industries. We have seen that ML-based AI has advanced the state of the art across a range of creative applications including content creation, information analysis, content enhancement, information extraction, information enhancement and data compression. ML–AI methods are data driven and benefit from recent advances in computational hardware and the availability of huge amounts of data for training—particularly image and video data.

We have differentiated throughout between the use of ML–AI as a creative tool and its potential as a creator in its own right. We foresee, in the near future, that AI will be adopted much more widely as a tool or collaborative assistant for creativity, supporting acquisition, production, post-production, delivery and interactivity. The concurrent advances in computing power, storage capacities and communication technologies (such as 5G) will support the embedding of AI processing within and at the edge of the network. In contrast, we observe that, despite recent advances, significant challenges remain for AI as the sole generator of original work. ML–AI works well when there are clearly defined problems that do not depend on external context or require long chains of inference or reasoning in decision making. It also benefits significantly from large amounts of diverse and unbiased data for training. Hence, the likelihood of AI (or its developers) winning awards for creative works in competition with human creatives may be some way off. We therefore conclude that, for creative applications, technological developments will, for some time yet, remain human-centric—designed to augment, rather than replace, human creativity. As AI methods begin to pervade the creative sector, developers and deployers must however continue to build trust; technological advances must go hand-in-hand with a greater understanding of ethical issues, data bias and wider social impact.

Acknowledgements This work has been funded by Bristol+Bath Creative R+D under AHRC grant AH/S002936/1. The Creative Industries Clusters Programme is managed by the Arts and Humanities Research Council as part of the Industrial Strategy Challenge Fund. The authors would like to acknowledge the following people who provided valuable contributions that enabled us to improve the quality and accuracy of this review: Ben Trewhella (Opposable Games), Darren Cosker (University of Bath), Fan Zhang (University of Bristol), and Paul Hill (University of Bristol).

References

Abdelhamed A, Afifi M, Timofte R, Brown MS (2020) NTIRE 2020 challenge on real image denoising: dataset, methods and results. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops

Adithya V, Rajesh R (2020) A deep convolutional neural network approach for static hand gesture recognition. Proced Comput Sci 171:2353–2361. https://doi.org/10.1016/j.procs.2020.04.255

Agostinelli F, Hoffman M, Sadowski P, Baldi P (2015) Learning activation functions to improve deep neural networks. In: Proceedings of international conference on learning representations, pp 1–9

Alsaih K, Lemaitre G, Rastgoo M, Sidibé D, Meriaudeau F (2017) Machine learning techniques for diabetic macular EDEMA (DME) classification on SD-OCT images. BioMed Eng 16(1):1–12. https://doi.org/10.1186/s12938-017-0352-9

Amato G, Falchi F, Gennaro C, Rabitti F (2017) Searching and annotating 100M images with YFCC100M-HNfc6 and MI-File. In: Proceedings of the 15th international workshop on content-based multimedia indexing https://doi.org/10.1145/3095713.3095740

Anantrasirichai N, Bull D (2019) DefectNet: multi-class fault detection on highly-imbalanced datasets. In: IEEE international conference on image processing (ICIP), pp 2481–2485

Anantrasirichai N, Bull D (2021) Contextual colorization and denoising for low-light ultra high resolution sequences. In: IEEE international conference on image processing (ICIP)

Anantrasirichai N, Achim A, Kingsbury N, Bull D (2013) Atmospheric turbulence mitigation using complex wavelet-based fusion. Image Process, IEEE Trans 22(6):2398–2408

Anantrasirichai N, Gilchrist ID, Bull DR (2016) Fixation identification for low-sample-rate mobile eye trackers. In: IEEE international conference on image processing (ICIP), pp 3126–3130. https://doi.org/10.1109/ICIP.2016.7532935

Anantrasirichai N, Achim A, Bull D (2018) Atmospheric turbulence mitigation for sequences with moving objects using recursive image fusion. In: 2018 25th IEEE international conference on image processing (ICIP), pp 2895–2899

Anantrasirichai N, Biggs J, Albino F, Hill P, Bull D (2018) Application of machine learning to classification of volcanic deformation in routinely-generated InSAR data. J Geophys Res: Solid Earth 123:1–15. https://doi.org/10.1029/2018JB015911

Anantrasirichai N, Daniels KAJ, Burn JF, Gilchrist ID, Bull DR (2018) Fixation prediction and visual priority maps for biped locomotion. IEEE Trans Cybern 48(8):2294–2306. https://doi.org/10.1109/TCYB.2017.2734946

Anantrasirichai N, Biggs J, Albino F, Bull D (2019) A deep learning approach to detecting volcano deformation from satellite imagery using synthetic datasets. Remote Sensing Environ 230:111179

Anantrasirichai N, Zhang F, Malyugina A, Hill P, Katsenou A (2020a) Encoding in the dark grand challenge: an overview. In: IEEE international conference on multimedia and Expo (ICME)

Anantrasirichai N, Zheng R, Selesnick I, Achim A (2020b) Image fusion via sparse regularization with non-convex penalties. Pattern Recogn Lett 131:355–360. https://doi.org/10.1016/j.patrec.2020.01.020

Anantrasirichai N, Geravand M, Braendler D, Bull DR (2021) Fast depth estimation for view synthesis. In: 2020 28th European signal processing conference (EUSIPCO), pp 575–579. https://doi.org/10.23919/Eusipco47968.2020.9287371

Anthony T, Eccles T, Tacchetti A, Kramár J, Gemp I, Hudson TC, Porcel N, Lanctot M, Pérolat J, Everett R, Singh S, Graepel T, Bachrach Y (2020) Learning to play no-press diplomacy with best response policy iteration. In: 34th Conference on neural information processing systems

Antic J (2020) DeOldify image colorization on DeepAI. https://github.com/jantic/DeOldify/. Accessed 10 Apr 2020

Arjovsky M, Chintala S, Bottou L (2017) Wasserstein GAN. In: Proceedings of machine learning research, vol 70

Asgari Taghanaki S, Abhishek K, Cohen J, Hamarneh G (2021) Deep semantic segmentation of natural and medical images: a review. Artif Intell Rev 54(1):137–178. https://doi.org/10.1007/s10462-020-09854-1

Azam N, Yao J (2012) Comparison of term frequency and document frequency based feature selection metrics in text categorization. Expert Syst Appl 39(5):4760–4768. https://doi.org/10.1016/j.eswa.2011.09.160

Barber A, Cosker D, James O, Waine T, Patel R (2016) Camera tracking in visual effects an industry perspective of structure from motion. In: Proceedings of the 2016 symposium on digital production, association for computing machinery, New York, DigiPro '16, pp 45–54. https://doi.org/10.1145/2947688.2947697

Barnett JT, Jain S, Andra U, Khurana T (2018) Cisco visual networking index (VNI): complete forecast update, pp 2017–2022. https://www.cisco.com/c/dam/m/en_us/network-intelligence/service-provider/digital-transformation/knowledge-network-webinars/pdfs/1211_BUSINESS_SERVICES_CKN_PDF.pdf
Bastug E, Bennis M, Medard M, Debbah M (2017) Toward interconnected virtual reality: opportunities, challenges, and enablers. IEEE Commun Maga 55(6):110–117
Batmaz Z, Yurekli A, Bilge A, Kaleli C (2019) A review on deep learning for recommender systems: challenges and remedies. Artif Intell Rev 52:1–37. https://doi.org/10.1007/s10462-018-9654-y
Berman D, treibitz T, Avidan S (2016) Non-local image dehazing. In: The IEEE conference on computer vision and pattern recognition (CVPR)
Bhattacharyya A, Fritz M, Schiele B (2019) "Best-of-many-samples" distribution matching. In: Workshop on Bayesian deep learning
Biemond J, Lagendijk RL, Mersereau RM (1990) Iterative methods for image deblurring. Proc IEEE 78(5):856–883
Black S, Keshavarz S, Souvenir R (2020) Evaluation of image inpainting for classification and retrieval. In: IEEE winter conference on applications of computer vision (WACV), pp 1049–1058
Bochkovskiy A, Wang CY, Liao HYM (2020) YOLOv4: optimal speed and accuracy of object detection. arXiv:abs/2004.10934
Borji A, Cheng M, Hou Q, Li J (2019) Salient object detection: a survey. Comput Vis Media 5:117–150. https://doi.org/10.1007/s41095-019-0149-9
Borysenko D, Mykheievskyi D, Porokhonskyy V (2020) Odesa: object descriptor that is smooth appearance-wise for object tracking task. In: To be submitted to ECCV'20
Bostrom N (2014) Superintelligence. Oxford University Press, Oxford
Bostrom N, Yudkowsky E (2014) The ethics of artificial intelligence. In: In Cambridge handbook of artificial intelligence
Bragg D, Koller O, Bellard M, Berke L, Boudreault P, Braffort A, Caselli N, Huenerfauth M, Kacorri H, Verhoef T, Vogler C, Ringel Morris M (2019) Sign language recognition, generation, and translation: An interdisciplinary perspective. In: International ACM SIGACCESS conference on computers and accessibility, pp 16–31. https://doi.org/10.1145/3308561.3353774
Briot JP, Hadjeres G, Pachet FD (2020) Deep learning techniques for music generation. Springer, Cham. https://doi.org/10.1007/978-3-319-70163-9
Brock A, Donahue J, Simonyan K (2019) Large scale GAN training for high fidelity natural image synthesis. In: International conference on learning representations (ICLR)
Brooks T, Mildenhall B, Xue T, Chen J, Sharlet D, Barron JT (2019) Unprocessing images for learned raw denoising. In: The IEEE conference on computer vision and pattern recognition (CVPR)
Buades A, Duran J (2019) CFA video denoising and demosaicking chain via spatio-temporal patch-based filtering. IEEE Trans Circ Syst Video Tech 30(11):1. https://doi.org/10.1109/TCSVT.2019.2956691
Bulat A, Tzimiropoulos G (2017) How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In: The IEEE international conference on computer vision (ICCV)
Bull D, Zhang F (2021) Intelligent image and video compression: communicating pictures, 2nd edn. Elsevier, New York
Caballero J, Ledig C, Aitken A, Acosta A, Totz J, Wang Z, Shi W (2017) Real-time video super-resolution with spatio-temporal networks and motion compensation. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 2848–2857. https://doi.org/10.1109/CVPR.2017.304
Cai B, Xu X, Jia K, Qing C, Tao D (2016) DehazeNet: an end-to-end system for single image haze removal. IEEE Trans Image Process 25(11):5187–5198
Cai X, Pu Y (2019) Flattenet: a simple and versatile framework for dense pixelwise prediction. IEEE Access 7:179985–179996
Caramiaux B, Lotte F, Geurts J, Amato G, Behrmann M, Falchi F, Bimbot F, Garcia A, Gibert J, Gravier G, Hadmut Holken HK, Lefebvre S, Liutkus A, Perkis A, Redondo R, Turrin E, Vieville T, Vincent E (2019) AI in the media and creative industries. In: New European media (NEM), hal-02125504f
Chak WH, Lau CP, Lui LM (2018) Subsampled turbulence removal network. arXiv:1807.04418v2
Chan C, Ginosar S, Zhou T, Efros A (2019) Everybody dance now. In: IEEE/CVF international conference on computer vision (ICCV), pp 5932–5941
Chang AX, Funkhouser T, Guibas L, Hanrahan P, Huang Q, Li Z, Savarese S, Savva M, Song S, Su H, Xiao J, Yi L, Yu F (2015) ShapeNet: an information-rich 3D model repository. arXiv:1512.03012
Chang J, Chen Y (2018) Pyramid stereo matching network. In: IEEE/CVF conference on computer vision and pattern recognition, pp 5410–5418. https://doi.org/10.1109/CVPR.2018.00567

Chang Y, Liu ZY, Lee K, Hsu W (2019) Free-form video inpainting with 3d gated convolution and temporal patchgan. In: IEEE/CVF international conference on computer vision (ICCV), pp 9065–9074
Chaplot DS, Salakhutdinov R, Gupta A, Gupta S (2020) Neural topological slam for visual navigation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)
Chen C, Chen Q, Xu J, Koltun V (2018a) Learning to see in the dark. In: IEEE/CVF conference on computer vision and pattern recognition, pp 3291–3300
Chen C, Jain U, Schissler C, Gari SVA, Al-Halah Z, Ithapu VK, Robinson P, Grauman K (2020) Soundspaces: audio-visual navigation in 3D environments. In: European Conference on Computer Vision (ECCV)
Chen F, De Vleeschouwer C, Cavallaro A (2014) Resource allocation for personalized video summarization. IEEE Trans Multimed 16(2):455–469. https://doi.org/10.1109/TMM.2013.2291967
Chen G, Ye D, Xing Z, Chen J, Cambria E (2017) Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. In: 2017 international joint conference on neural networks (IJCNN), pp 2377–2383. https://doi.org/10.1109/IJCNN.2017.7966144
Chen J, Chen J, Chao H, Yang M (2018b) Image blind denoising with generative adversarial network based noise modeling. In: IEEE/CVF conference on computer vision and pattern recognition, pp 3155–3164
Chen H, Ding G, Zhao S, Han J (2018) Temporal-difference learning with sampling baseline for image captioning. In: 32nd AAAI conference on artificial intelligence
Chen K, Wang J, Pang J, Cao Y, Xiong Y, Li X, Sun S, Feng W, Liu Z, Xu J, Zhang Z, Cheng D, Zhu C, Cheng T, Zhao Q, Li B, Lu X, Zhu R, Wu Y, Dai J, Wang J, Shi J, Ouyang W, Loy CC, Lin D (2019) MMDetection: open mmlab detection toolbox and benchmark. arXiv preprint arXiv:190607155
Chen SF, Chen YC, Yeh CK, Wang YCF (2018) Order-free rnn with visual attention for multi-label classification. In: AAAI conference on artificial intelligence
Chen Z, Wei X, Wang P, Guo Y (2019) Multi-label image recognition with graph convolutional networks. In: 2019 IEEE/CVF conference on computer vision and pattern recognition, pp 5172–5181. https://doi.org/10.1109/CVPR.2019.00532
Cheng MM, Zhang FL, Mitra NJ, Huang X, Hu SM (2010) Repfinder: finding approximately repeated scene elements for image editing 29(4), 1-8. https://doi.org/10.1145/1778765.1778820
Cheng X, Wang P, Yang R (2019) Learning depth with convolutional spatial propagation network. IEEE Trans Pattern Anal Mach Intell 42(10):1
Cheng Z, Yang Q, Sheng B (2015) Deep colorization. In: The IEEE international conference on computer vision (ICCV)
Chuah SHW (2018) Why and who will adopt extended reality technology? Literature review, synthesis, and future research agenda. SSRN. https://doi.org/10.2139/ssrn.3300469
Claus M, van Gemert J (2019) ViDeNN: deep blind video denoising. In: CVPR workshop
Cohen NS (2015) From pink slips to pink slime: transforming media labor in a digital age. Commun Rev 18(2):98–122. https://doi.org/10.1080/10714421.2015.1031996
Dabov K, Foi A, Katkovnik V, Egiazarian K (2007) Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE Trans Image Process 16(8):2080–2095
Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H, Wei Y (2017) Deformable convolutional networks. In: IEEE international conference on computer vision (ICCV), pp 764–773. https://doi.org/10.1109/ICCV.2017.89
Dai T, Cai J, Zhang Y, Xia S, Zhang L (2019) Second-order attention network for single image super-resolution. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 11057–11066
Damen D, Doughty H, Farinella GM, Fidler S, Furnari A, Kazakos E, Moltisanti D, Munro J, Perrett T, Price W, Wray M (2018) Scaling egocentric vision: the epic-kitchens dataset. In: European conference on computer vision
Damodaran BB, Kellenberger B, Flamary R, Tuia D, Courty N (2018) DeepJDOT: deep joint distribution optimal transport for unsupervised domain adaptation. In: The European conference on computer vision (ECCV)
Davies J, Klinger J, Mateos-Garcia J, Stathoulopoulos K (2020) The art in the artificial AI and the creative industries. Creat Ind Policy Evid Centre 1–38
Davy A, Ehret T, Morel J, Arias P, Facciolo G (2019) A non-local cnn for video denoising. In: IEEE international conference on image processing (ICIP), pp 2409–2413. https://doi.org/10.1109/ICIP.2019.8803314
Deldjoo Y, Constantin MG, Eghbal-Zadeh H, Ionescu B, Schedl M, Cremonesi P (2018) Audio-visual encoding of multimedia content for enhancing movie recommendations. In: Proceedings of the 12th ACM conference on recommender systems, association for computing machinery, New York, NY, USA, RecSys '18, pp 455–459. https://doi.org/10.1145/3240323.3240407

Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Conference of the North American chapter of the association for computational linguistics: human language technologies, vol 1

Dignum V (2018) Ethics in artificial intelligence: introduction to the special issue. Ethics Inf Technol, 20:1–3

Dodds L (2020) The ai that unerringly predicts hollywood's hits and flops. https://www.telegraph.co.uk/technology/2020/01/20/ai-unerringly-predicts-hollywoods-hits-flops/. Accessed 10 Apr 2020

Doetsch P, Kozielski M, Ney H (2014) Fast and robust training of recurrent neural networks for offline handwriting recognition. In: 2014 14th international conference on frontiers in handwriting recognition, pp 279–284

Donahue C, McAuley J, Puckette M (2019) Adversarial audio synthesis. In: International conference on learning representations (ICLR)

Dong C, Loy CC, He K, Tang X (2014) Learning a deep convolutional network for image super-resolution. In: The European conference on computer vision (ECCV), pp 184–199

Dörr KN (2016) Mapping the field of algorithmic journalism. Digit J 4(6):700–722. https://doi.org/10.1080/21670811.2015.1096748

Dzmitry Bahdanau YB Kyunghyun Cho (2015) Neural machine translation by jointly learning to align and translate. In: International conference on learning representations

Elgammal A, Liu B, Elhoseiny M, Mazzone M (2017) CAN: creative adversarial networks, generating "art" by learning about styles and deviating from style norms. arXiv:1706.07068

Engel J, Agrawal KK, Chen S, Gulrajani I, Donahue C, Roberts A (2019) GANSynth: adversarial neural audio synthesis. In: International conference on learning representations

Engin D, Genc A, Kemal Ekenel H (2018) Cycle-Dehaze: enhanced CycleGAN for single image dehazing. In: The IEEE conference on computer vision and pattern recognition (CVPR) workshops

Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2012) The PASCAL visual object classes challenge 2012 (VOC2012) results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html

Fan D, Wang W, Cheng M, Shen J (2019) Shifting more attention to video salient object detection. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 8546–8556. https://doi.org/10.1109/CVPR.2019.00875

Fan DP, Lin Z, Ji GP, Zhang D, Fu H, Cheng MM (2020) Taking a deeper look at co-salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)

Fang K (2016) Track-RNN: Joint detection and tracking using recurrent neural networks. In: Conference on neural information processing systems

Flynn J, Broxton M, Debevec P, DuVall M, Fyffe G, Overbeck R, Snavely N, Tucker R (2019) DeepView: view synthesis with learned gradient descent. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 2362–2371

Foster D (2019) Generative deep learning: teaching machines to paint, write, compose, and play. O'Reilly Media Inc

Frogner C, Zhang C, Mobahi H, Araya-Polo M, Poggio T (2015) Learning with a wasserstein loss. In: Proceedings of the 28th international conference on neural information processing systems, NIPS'15, vol 2. MIT Press, Cambridge, pp 2053–2061

Fukushima K (1980) Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybern 36:193–202. https://doi.org/10.1007/BF00344251

Gao H, Tao X, Shen X, Jia J (2019) Dynamic scene deblurring with parameter selective sharing and nested skip connections. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 3843–3851

Gao J, Anantrasirichai N, Bull D (2019) Atmospheric turbulence removal using convolutional neural network. arXiv:1912.11350

Gao R, Grauman K (2019) 2.5D visual sound. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 324–333

Gatys L, Ecker A, Bethge M (2016) A neural algorithm of artistic style. J Vis. https://doi.org/10.1167/16.12.326

Geiger A, Lenz P, Urtasun R (2012) Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Conference on computer vision and pattern recognition (CVPR)

Ghani NA, Hamid S, Hashem IA, Ahmed E (2019) Social media big data analytics: a survey. Comput Hum Behav 101:417–428. https://doi.org/10.1016/j.chb.2018.08.039

Gkioxari G, Johnson J, Malik J (2019) Mesh r-CNN. In: IEEE/CVF international conference on computer vision (ICCV), pp 9784–9794
Golbeck J, Robles C, Turner K (2011) Predicting personality with social media. In: CHI '11 extended abstracts on human factors in computing systems, pp 253–262. https://doi.org/10.1145/1979742.1979614
Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (eds) Advances in neural information processing systems, vol 27. Curran Associates, Inc., pp 2672–2680. http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf
Gordo A, Almazán J, Revaud J, Larlus D (2016) Deep image retrieval: learning global representations for image search. In: The European conference on computer vision (ECCV). Springer, pp 241–257
Gordon D, Farhadi A, Fox D (2018) Re^3: real-time recurrent regression networks for visual tracking of generic objects. IEEE Robot Autom Lett 3(2):788–795
Goyal M, Tatwawadi K, Chandak S, Ochoa I (2019) DeepZip: lossless data compression using recurrent neural networks. In: 2019 data compression conference (DCC), pp 575–575
Graves A, Mohamed A, Hinton G (2013) Speech recognition with deep recurrent neural networks. In: IEEE international conference on acoustics, speech and signal processing, pp 6645–6649
Gregor K, Papamakarios G, Besse F, Buesing L, Weber T (2019) Temporal difference variational auto-encoder. In: International conference on learning representations
Güera D, Delp EJ (2018) Deepfake video detection using recurrent neural networks. In: 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS), pp 1–6
Gunasekara I, Nejadgholi I (2018) A review of standard text classification practices for multi-label toxicity identification of online content. In: Proceedings of the 2nd workshop on abusive language online (ALW2). Association for Computational Linguistics, Brussels, Belgium, pp 21–25. https://doi.org/10.18653/v1/W18-5103. https://www.aclweb.org/anthology/W18-5103
Guo K, Lincoln P, Davidson P, Busch J, Yu X, Whalen M, Harvey G, Orts-Escolano S, Pandey R, Dourgarian J, DuVall M, Tang D, Tkach A, Kowdle A, Cooper E, Dou M, Fanello S, Fyffe G, Rhemann C, Taylor J, Debevec P, Izadi S (2019) The relightables: volumetric performance capture of humans with realistic relighting. In: ACM SIGGRAPH Asia
Gupta R, Thapar Khanna M, Chaudhury S (2013) Visual saliency guided video compression algorithm. Signal Process: Image Commun 28(9):1006–1022. https://doi.org/10.1016/j.image.2013.07.003
Ha D, Eck D (2018) A neural representation of sketch drawings. In: International conference on learning representations
Hall DW, Pesenti J (2018) Growing the artificial intelligence industry in the UK. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artificial_intelligence_industry_in_the_UK.pdf
Han J, Lombardo S, Schroers C, Mandt S (2019) Deep generative video compression. In: Conference on neural information processing systems 32:1–12
Han X, Laga H, Bennamoun M (2021) Image-based 3D object reconstruction: state-of-the-art and trends in the deep learning era. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(5):1578–1604
Haris M, Shakhnarovich G, Ukita N (2019) Recurrent back-projection network for video super-resolution. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 3892–3901
Hasan HR, Salah K (2019) Combating deepfake videos using blockchain and smart contracts. IEEE Access 7:41596–41606
Haugeland J (1985) Artificial intelligence: the very idea. MIT Press, New York
He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21(9):1263–1284. https://doi.org/10.1109/TKDE.2008.239
He K, Sun J, Tang X (2011) Single image haze removal using dark channel prior. IEEE Trans Pattern Anal Mach Intell 33(12):2341–2353
He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 37(9):1904–1916
He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 770–778
He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-CNN. In: IEEE international conference on computer vision (ICCV), pp 2980–2988
He Z, Zuo W, Kan M, Shan S, Chen X (2019) AttGAN: facial attribute editing by only changing what you want. IEEE Trans Image Process 28(11):5464–5478. https://doi.org/10.1109/TIP.2019.2916751
Héctor R (2014) MADE—massive artificial drama engine for non-player characters. FOSDEM VZW. https://doi.org/10.5446/32569. Accessed 26 May 2020

Hessel M, Modayil J, van Hasselt H, Schaul T, Ostrovski G, Dabney W, Horgan D, Piot B, Azar M, Silver D (2018) Rainbow: combining improvements in deep reinforcement learning. In: 32nd AAAI conference on artificial intelligence

Hildebrand HA (1999) Pitch detection and intonation correction apparatus and method. US Patent 5973252A

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

Holden D, Saito J, Komura T, Joyce T (2015) Learning motion manifolds with convolutional autoencoders. In: SIGGRAPH Asia 2015 technical briefs. Association for Computing Machinery,SA '15, New York. https://doi.org/10.1145/2820903.2820918

Honavar V (1995) Symbolic artificial intelligence and numeric artificial neural networks: towards a resolution of the dichotomy. Springer, Boston, pp 351–388. https://doi.org/10.1007/978-0-585-29599-2_11

Hong X, Xiong P, Ji R, Fan H (2019) Deep fusion network for image completion. In: Proceedings of the 27th ACM international conference on multimedia, pp 2033–2042. https://doi.org/10.1145/3343031.3351002

Hossain MS, Muhammad G (2019) Emotion recognition using deep learning approach from audio-visual emotional big data. Inf Fusion 49:69–78. https://doi.org/10.1016/j.inffus.2018.09.008

Hou Q, Cheng M, Hu X, Borji A, Tu Z, Torr PHS (2019) Deeply supervised salient object detection with short connections. IEEE Trans Pattern Anal Mach Intell 41(4):815–828. https://doi.org/10.1109/TPAMI.2018.2815688

Hradis M, Kotera J, Zemcik P, Sroubek F (2015) Convolutional neural networks for direct text deblurring. In: Proceedings of the British machine vision conference (BMVC), pp 6.1–6.13. https://doi.org/10.5244/C.29.6

Hu L, Saito S, Wei L, Nagano K, Seo J, Fursund J, Sadeghi I, Sun C, Chen YC, Li H (2017) Avatar digitization from a single image for real-time rendering. ACM Trans Graph 36(6):1–4. https://doi.org/10.1145/3130800.31310887

Hu Y, Wang K, Zhao X, Wang H, Li Y (2018) Underwater image restoration based on convolutional neural network. In: Proceedings of the 10th Asian conference on machine learning, PMLR, proceedings of machine learning research, vol 95, pp 296–311. http://proceedings.mlr.press/v95/hu18a.html

Huang G, Liu Z, v d Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 2261–2269. https://doi.org/10.1109/CVPR.2017.243

Huang SW, Lin CT, Chen SP, Wu YY, Hsu PH, Lai SH (2018) AugGAN: cross domain adaptation with GAN-based data augmentation. In: The European conference on computer vision (ECCV)

Huang Y, Wang W, Wang L (2015) Bidirectional recurrent convolutional networks for multi-frame super-resolution. In: Advances in neural information processing systems, vol 28. Curran Associates, Inc., pp 235–243. http://papers.nips.cc/paper/5778-bidirectional-recurrent-convolutional-networks-for-multi-frame-super-resolution.pdf

Huang Z, Zhou S, Heng W (2019) Learning to paint with model-based deep reinforcement learning. In: IEEE/CVF international conference on computer vision (ICCV), pp 8708–8717

Hyun Kim T, Mu Lee K, Scholkopf B, Hirsch M (2017) Online video deblurring via dynamic temporal blending network. In: The IEEE international conference on computer vision (ICCV)

Iqbal T, Qureshi S (2020) The survey: text generation models in deep learning. J King Saud Univ-Comput Inf Sci. https://doi.org/10.1016/j.jksuci.2020.04.001

Isola P, Zhu J, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 5967–5976. https://doi.org/10.1109/CVPR.2017.632

Jabeen S, Khan G, Naveed H, Khan Z, Khan UG (2018) Video retrieval system using parallel multi-class recurrent neural network based on video description. In: 2018 14th international conference on emerging technologies (ICET), pp 1–6

Jackson AS, Bulat A, Argyriou V, Tzimiropoulos G (2017) Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In: International conference on computer vision

Jalal MA, Chen R, Moore RK, Mihaylova L (2018) American sign language posture understanding with deep neural networks. In: International conference on information fusion (FUSION), pp 573–579. https://doi.org/10.23919/ICIF.2018.8455725

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning. Springer, New York

Jeon J, Lavrenko V, Manmatha R (2003) Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval, pp 119–126. https://doi.org/10.1145/860435.860459
Ji S, Xu W, Yang M, Yu K (2013) 3d convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231. https://doi.org/10.1109/TPAMI.2012.59
Jia J (2007) Single image motion deblurring using transparency. In: IEEE conference on computer vision and pattern recognition, pp 1–8
Jiang B, Zhou Z, Wang X, Tang J, Luo B (2020) CMSALGAN: RGB-D salient object detection with cross-view generative adversarial networks. IEEE Trans Multimed. https://doi.org/10.1109/TMM.2020.2997184
Jiang F, Tao W, Liu S, Ren J, Guo X, Zhao D (2018) An end-to-end compression framework based on convolutional neural networks. IEEE Trans Circuits Syst Video Technol 28(10):3007–3018
Jiang L, Shi S, Qi X, Jia J (2018) GAL: geometric adversarial loss for single-view 3D-object reconstruction. In: The European conference on computer vision (ECCV). Springer, Cham, pp 820–834
Jiang Y, Zhou T, Ji GP, Fu K, jun Zhao Q, Fan DP (2020) Light field salient object detection: a review and benchmark. arXiv:abs/2010.04968
Jiang Y, Gong X, Liu D, Cheng Y, Fang C, Shen X, Yang J, Zhou P, Wang Z (2021) Enlightengan: deep light enhancement without paired supervision. IEEE Trans Image Process 30:2340–2349. https://doi.org/10.1109/TIP.2021.3051462
Jin Y, Zhang J, Li M, Tian Y, Zhu H, Fang Z (2017) Towards the automatic anime characters creation with generative adversarial networks. arXiv:1708.05509
Johnson J, Alahi A, Fei-Fei L (2016) Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision
Johnson R, Zhang T (2015) Effective use of word order for text categorization with convolutional neural networks. In: Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies, association for computational linguistics, pp 103–112. https://doi.org/10.3115/v1/N15-1011. https://www.aclweb.org/anthology/N15-1011
Justesen N, Bontrager P, Togelius J, Risi S (2020) Deep learning for video game playing. IEEE Trans Games 12(1):1–20
Kaminskas M, Ricci F (2012) Contextual music information retrieval and recommendation: State of the art and challenges. Comput Sci Rev 6(2):89–119. https://doi.org/10.1016/j.cosrev.2012.04.002
Kanazawa A, Black MJ, Jacobs DW, Malik J (2018) End-to-end recovery of human shape and pose. In: IEEE/CVF conference on computer vision and pattern recognition, pp 7122–7131
Kaneko H, Goto J, Kawai Y, Mochizuki T, Sato S, Imai A, Yamanouchi Y (2020) AI-driven smart production. SMPTE Motion Imaging J 129(2):27–35
Kappeler A, Yoo S, Dai Q, Katsaggelos AK (2016) Video super-resolution with convolutional neural networks. IEEE Trans Comput Imaging 2(2):109–122
Karras T, Aila T, Laine S, Lehtinen J (2018) Progressive growing of GANs for improved quality, stability, and variation. In: International conference on learning representations (ICLR)
Kartynnik Y, Ablavatski A, Grishchenko I, Grundmann M (2019) Real-time facial surface geometry from monocular video on mobile GPUs. In: CVPR workshop on computer vision for augmented and virtual reality
Kazakos E, Nagrani A, Zisserman A, Damen D (2019) EPIC-Fusion: audio-visual temporal binding for egocentric action recognition. In: IEEE/CVF international conference on computer vision (ICCV), pp 5491–5500
Keswani B, Mohapatra AG, Mishra TC, Keswani P, Mohapatra PCG, Akhtar MM, Vijay P (2020) World of virtual reality (VR) in healthcare. Springer, pp 1–23. https://doi.org/10.1007/978-3-030-35252-3_1
Kietzmann J, Lee LW, McCarthy IP, Kietzmann TC (2020) Deepfakes: trick or treat? Bus Horiz 63(2):135–146. https://doi.org/10.1016/j.bushor.2019.11.006
Kim D, Woo S, Lee J, Kweon IS (2019) Deep video inpainting. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 5785–5794. https://doi.org/10.1109/CVPR.2019.00594
Kim J, Lee JK, Lee KM (2016) Accurate image super-resolution using very deep convolutional networks. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 1646–1654
Kim N, Lee D, Oh S (2020a) Learning instance-aware object detection using determinantal point processes. Comput Vis Image Underst 201:103061. https://doi.org/10.1016/j.cviu.2020.103061
Kim SW, Zhou Y, Philion J, Torralba A, Fidler S (2020b) Learning to Simulate Dynamic Environments with GameGAN. In: IEEE conference on computer vision and pattern recognition (CVPR)
Kirillov A, Wu Y, He K, Girshick R (2020) Pointrend: image segmentation as rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)

Ko B (2018) A brief review of facial emotion recognition based on visual information. Sensors 18:401
Kopf J, Neubert B, Chen B, Cohen M, Cohen-Or D, Deussen O, Uyttendaele M, Lischinski D (2008) Deep photo: model-based photograph enhancement and viewing. ACM Trans Graph 27(5):1–10. https://doi.org/10.1145/1409060.1409069
Kowsari K, Jafari Meimandi K, Heidarysafa M, Mendu S, Barnes L, Brown D (2019) Text classification algorithms: a survey. Information 10(4):150. https://doi.org/10.3390/info10040150
Kratimenos A, Pavlakos G, Maragos P (2020) 3D hands, face and body extraction for sign language recognition. In: European conference on computer vision workshop
Krishnan D, Tay T, Fergus R (2011) Blind deconvolution using a normalized sparsity measure. CVPR 2011:233–240
Kristan M, Matas J, Leonardis A, Vojir T, Pflugfelder R, Fernandez G, Nebehay G, Porikli F, Čehovin L (2016) A novel performance evaluation methodology for single-target trackers. IEEE Trans Pattern Anal Mach Intell 38(11):2137–2155. https://doi.org/10.1109/TPAMI.2016.2516982
Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th international conference on neural information processing systems, vol 1. Curran Associates Inc., USA, pp 1097–1105
Krull A, Buchholz T, Jug F (2019) Noise2Void—learning denoising from single noisy images. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 2124–2132
Kuang X, Sui X, Liu Y, Chen Q, Gu G (2019) Single infrared image enhancement using a deep convolutional neural network. Neurocomputing 332:119–128. https://doi.org/10.1016/j.neucom.2018.11.081
Kuang X, Zhu J, Sui X, Liu Y, Liu C, Chen Q, Gu G (2020) Thermal infrared colorization via conditional generative adversarial network. Infrared Phys Technol 107:103338. https://doi.org/10.1016/j.infrared.2020.103338
Kupyn O, Budzan V, Mykhailych M, Mishkin D, Matas J (2018) DeblurGAN: Blind motion deblurring using conditional adversarial networks. In: The IEEE conference on computer vision and pattern recognition (CVPR)
Kwon OW, Chan K, Hao J, Lee TW (2003) Emotion recognition by speech signals. In: EUROSPEECH-2003, pp 125–128
Lacerda A, Cristo M, Gonçalves MA, Fan W, Ziviani N, Ribeiro-Neto B (2006) Learning to advertise. In: Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, association for computing machinery, New York, NY, USA, SIGIR '06, pp 549–556. https://doi.org/10.1145/1148170.1148265
Laver KE, Lange B, George S, Deutsch JE, Saposnik G, Crotty M (2017) Virtual reality for stroke rehabilitation. Cochrane Database Syst Rev 11(11):1–183
LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. Neural Comput 1(4):541–551
Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
Ledig C, Theis L, Huszar F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, Shi W (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 105–114
Lee K, Lee S, Lee J (2018) Interactive character animation by learning multi-objective control. ACM Trans Graph 37(6):1–10
Lehtinen J, Munkberg J, Hasselgren J, Laine S, Karras T, Aittala M, Aila T (2018) Noise2Noise: learning image restoration without clean data. In: Proceedings of the 35th international conference on machine learning, vol 80, pp 2965–2974
Lempitsky V, Vedaldi A, Ulyanov D (2018) Deep image prior. In: IEEE/CVF conference on computer vision and pattern recognition, pp 9446–9454
Leppänen L, Munezero M, Granroth-Wilding M, Toivonen H (2017) Data-driven news generation for automated journalism. In: Proceedings of the 10th international conference on natural language generation, association for computational linguistics, Santiago de Compostela, Spain, pp 188–197. https://doi.org/10.18653/v1/W17-3528
Lewis JJ, O'Callaghan RJ, Nikolov SG, Bull DR, Canagarajah N (2007) Pixel- and region-based image fusion with complex wavelets. Info Fusion 8(2):119–130 Special Issue on Image Fusion: Advances in the State of the Art
Li B, Peng X, Wang Z, Xu J, Feng D (2017) AOD-Net: all-in-one dehazing network. In: IEEE international conference on computer vision (ICCV), pp 4780–4788
Li B, Yan J, Wu W, Zhu Z, Hu X (2018) High performance visual tracking with siamese region proposal network. In: The IEEE conference on computer vision and pattern recognition (CVPR)

Li B, Ren W, Fu D, Tao D, Feng D, Zeng W, Wang Z (2019) Benchmarking single-image dehazing and beyond. IEEE Trans Image Process 28(1):492–505
Li J, Li B, Xu J, Xiong R, Gao W (2018) Fully connected network-based intra prediction for image coding. IEEE Trans Image Process 27(7):3236–3247
Li S, Kang X, Hu J (2013) Image fusion with guided filtering. IEEE Trans Image Process 22(7):2864–2875
Li J, Li H, Zong C (2019a) Towards personalized review summarization via user-aware sequence network. Proceed AAAI Conf Artif Intell 33(01):6690–6697. https://doi.org/10.1609/aaai.v33i01.33016690
Li S, Jang S, Sung Y (2019b) Automatic melody composition using enhanced GAN. Mathematics 7:883
Li W, Zhang P, Zhang L, Huang Q, He X, Lyu S, Gao J (2019c) Object-driven text-to-image synthesis via adversarial training. In: The IEEE conference on computer vision and pattern recognition (CVPR)
Li Z, Ma Y, Chen Y, Zhang X, Sun J (2019d) Joint COCO and mapillary workshop at ICCV 2019: Coco instance segmentation challenge track Technical report: MegDetV2. In: IEEE international conference on computer vision workshop
Li X, Liu M, Ye Y, Zuo W, Lin L, Yang R (2018a) Learning warped guidance for blind face restoration. In: The European conference on computer vision (ECCV), pp 278–296
Li Y, Lyu S (2019) Exposing deepfake videos by detecting face warping artifacts. In: IEEE conference on computer vision and pattern recognition workshops (CVPRW)
Li Y, Lu H, Li J, Li X, Li Y, Serikawa S (2016) Underwater image de-scattering and classification by deep neural network. Comput Electr Eng 54:68–77. https://doi.org/10.1016/j.compeleceng.2016.08.008
Li Y, Pan Q, Wang S, Yang T, Cambria E (2018b) A generative model for category text generation. Inf Sci 450:301–315. https://doi.org/10.1016/j.ins.2018.03.050
Limmer M, Lensch HPA (2016) Infrared colorization using deep convolutional neural networks. In: 15th IEEE international conference on machine learning and applications (ICMLA), pp 61–68
Lin T, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 936–944
Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: common objects in context, pp 740–755
Liu D, Ma H, Xiong Z, Wu F (2018) CNN-based DCT-like transform for image compression. In: MultiMedia modeling, pp 61–72
Liu D, Wang Z, Fan Y, Liu X, Wang Z, Chang S, Wang X, Huang TS (2018a) Learning temporal dynamics for video super-resolution: a deep learning approach. IEEE Trans Image Process 27(7):3432–3445
Liu J, Xia S, Yang W, Li M, Liu D (2019) One-for-All: grouped variation network-based fractional interpolation in video coding. IEEE Trans Image Process 28(5):2140–2151
Liu L, Ouyang W, Wang X, Fieguth P, Chen J, Liu X, Pietikainen M (2020) Deep learning for generic object detection: a survey. Int J Comput Vis 128:261–318. https://doi.org/10.1007/s11263-019-01247-4
Liu P, Zhang H, Zhang K, Lin L, Zuo W (2018b) Multi-level wavelet-CNN for image restoration. In: IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), pp 886–88609
Liu Y, Chen X, Peng H, Wang Z (2017) Multi-focus image fusion with a deep convolutional neural network. Inf Fusion 36:191–207. https://doi.org/10.1016/j.inffus.2016.12.001
Liu Y, Chen X, Wang Z, Wang ZJ, Ward RK, Wang X (2018) Deep learning for pixel-level image fusion: recent advances and future prospects. Inf Fusion 42:158–173. https://doi.org/10.1016/j.inffus.2017.10.007
Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965
Lore KG, Akintayo A, Sarkar S (2017) Llnet: a deep autoencoder approach to natural low-light image enhancement. Pattern Recogn 61:650–662. https://doi.org/10.1016/j.patcog.2016.06.008
Lu C, Uchiyama H, Thomas D, Shimada A, Ichiro Taniguchi R, (2018) Sparse cost volume for efficient stereo matching. Remote sensing 10(11):1–12
Lu G, Ouyang W, Xu D, Zhang X, Cai C, Gao Z (2019) DVC: an end-to-end deep video compression framework. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 10998–11007
Lu G, Zhang X, Ouyang W, Chen L, Gao Z, Xu D (2020) An end-to-end learning framework for video compression. IEEE Trans Pattern Anal Mach Intell 1
Lucas A, Iliadis M, Molina R, Katsaggelos AK (2018) Using deep neural networks for inverse problems in imaging: beyond analytical methods. IEEE Signal Process Maga 35(1):20–36
Lundervold AS, Lundervold A (2019) An overview of deep learning in medical imaging focusing on MRI. Z Med Phys 29(2):102–127. https://doi.org/10.1016/j.zemedi.2018.11.002. Special Issue: Deep Learning in Medical Physics

Ma D, Afonso M, Zhang F, Bull D (2019a) Perceptually-inspired super-resolution of compressed videos. In: Proc. SPIE 11137, applications of digital image processing XLII, vol 1113717, pp 310–318
Ma D, Zhang F, Bull DR (2020) BVI-DVC: a training database for deep video compression. arXiv:2003.13552
Ma D, Zhang F, Bull DR (2020a) Gan-based effective bit depth adaptation for perceptual video compression. In: IEEE international conference on multimedia and expo (ICME), pp 1–6
Ma D, Zhang F, Bull DR (2021) CVEGAN: a perceptually-inspired gan for compressed video enhancement. arXiv:2011.09190v2
Ma J, Ma Y, Li C (2019b) Infrared and visible image fusion methods and applications: a survey. Inf Fusion 45:153–178. https://doi.org/10.1016/j.inffus.2018.02.004
Ma J, Yu W, Liang P, Li C, Jiang J (2019c) FusionGAN: a generative adversarial network for infrared and visible image fusion. Inf Fusion 48:11–26. https://doi.org/10.1016/j.inffus.2018.09.004
Ma S, Zhang X, Jia C, Zhao Z, Wang S, Wang S (2020b) Image and video compression with neural networks: a review. IEEE Trans Circuits Syst Video Technol 30(6):1683–1698
Maas A, Le QV, O'Neil TM, Vinyals O, Nguyen P, Ng AY (2012) Recurrent neural networks for noise reduction in robust ASR. In: INTERSPEECH
Maggioni M, Katkovnik V, Egiazarian K, Foi A (2012) Nonlocal transform-domain filter for volumetric data denoising and reconstruction. IEEE Trans Image Process 22(1):119–133
Maier R, Kim K, Cremers D, Kautz J, Nießner M (2017) Intrinsic3D: high-quality 3D reconstruction by joint appearance and geometry optimization with spatially-varying lighting. In: IEEE international conference on computer vision (ICCV), pp 3133–3141
Malleson C, Guillemaut JY, Hilton A (2019) 3D reconstruction from RGB-D data. Springer, pp 87–115. https://doi.org/10.1007/978-3-030-28603-3_5
Malm H, Oskarsson M, Warrant E, Clarberg P, Hasselgren J, Lejdfors C (2007) Adaptive enhancement and noise reduction in very low light-level video. In: IEEE ICCV, pp 1–8. https://doi.org/10.1109/ICCV.2007.4409007
Mansimov E, Parisotto E, Ba JL, Salakhutdinov R (2016) Generating images from captions with attention. In: International conference on learning representations
Mao HH, Shin T, Cottrell G (2018) DeepJ: style-specific music generation. In: IEEE 12th international conference on semantic computing (ICSC), pp 377–382
Mariani G, Scheidegger F, Istrate R, Bekas C, Malossi C (2018) BAGAN: Data augmentation with balancing GAN. arXiv:1803.09655v2
Matsugu M, Mori K, Mitari Y, Kaneda Y (2003) Subject independent facial expression recognition with robust face detection using a convolutional neural network. Neural Netw 16(5–6):555–559. https://doi.org/10.1016/S0893-6080(03)00115-1
McCulloch W, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 5:115–133. https://doi.org/10.1007/BF02478259
Mejjati Y, Gomez C, Kim K, Shechtman E, Bylinskii Z (2020) Look here! a parametric learning based approach to redirect visual attention. In: European conference on computer vision. https://doi.org/10.1007/978-3-030-58592-1_21
Mentzer F, Toderici GD, Tschannen M, Agustsson E (2020) High-fidelity generative image compression. Adv Neural Inf Process Syst 33:1–12
Mescheder L, Oechsle M, Niemeyer M, Nowozin S, Geiger A (2019) Occupancy networks: learning 3D reconstruction in function space. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 4455–4465
Milan A, Rezatofighi SH, Dick A, Reid I, Schindler K (2017) Online multi-target tracking using recurrent neural networks. In: Proceedings of the 31st AAAI conference on artificial intelligence. AAAI Press, AAAI'17, pp 4225–4232
Milgram P, Kishino F (1994) A taxonomy of mixed reality visual displays. IEICE Trans Inf Syst 77(12):1–15
Milgram P, Takemura H, Utsumi A, Kishino F (1995) Augmented reality: a class of displays on the reality-virtuality continuum. Telemanipulator Telepresence Technol, SPIE 2351:282–292. https://doi.org/10.1117/12.197321
Mirza M, Osindero S (2014) Conditional generative adversarial nets. arXiv:1411.1784v1
Mitchell TM (1997) Machine learning. McGraw Hill Education
Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, Riedmiller M (2013) Playing atari with deep reinforcement learning. In: NIPS deep learning workshop
Morgado P, Nvasconcelos N, Langlois T, Wang O (2018) Self-supervised generation of spatial audio for 360° video. In: Advances in neural information processing systems, vol 11. pp 362–372

Nagano K, Seo J, Xing J, Wei L, Li Z, Saito S, Agarwal A, Fursund J, Li H (2018) PaGAN: real-time avatars using dynamic textures. ACM Trans Graph 37(6):1–12. https://doi.org/10.1145/3272127.3275075

Nah S, Hyun Kim T, Mu Lee K (2017) Deep multi-scale convolutional neural network for dynamic scene deblurring. In: The IEEE conference on computer vision and pattern recognition (CVPR)

Nah S, Son S, Lee KM (2019) Recurrent neural networks with intra-frame iterations for video deblurring. In: The IEEE conference on computer vision and pattern recognition (CVPR)

Nah S, Timofte R, Zhang R, Suin M, Purohit K, Rajagopalan AN, S AN, Pinjari JB, Xiong Z, Shi Z, Chen C, Liu D, Sharma M, Makwana M, Badhwar A, Singh AP, Upadhyay A, Trivedi A, Saini A, Chaudhury S, Sharma PK, Jain P, Sur A, Özbulak G (2019) NTIRE 2019 challenge on image colorization: report. In: IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), pp 2233–2240

Nalbach O, Arabadzhiyska E, Mehta D, Seidel HP, Ritschel T (2017) Deep shading: convolutional neural networks for screen space shading. Comput Graph Forum 36(4):65–78. https://doi.org/10.1111/cgf.13225

Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. In: The European conference on computer vision (ECCV). Springer, Cham, pp 483–499

Ng AK, Chan LK, Lau HY (2020) A study of cybersickness and sensory conflict theory using a motion-coupled virtual reality system. Displays 61:101922. https://doi.org/10.1016/j.displa.2019.08.004

Nguyen TT, Nguyen ND, Nahavandi S (2020) Deep reinforcement learning for multiagent systems: a review of challenges, solutions, and applications. IEEE Trans Cybern 50(9):1–14

Nieuwenhuizen R, Schutte K (2019) Deep learning for software-based turbulence mitigation in long-range imaging. Artif Intell Mach Learn Def Appl, Int Soc Opt Photon, SPIE 11169:153–162. https://doi.org/10.1117/12.2532603

Noh H, Hong S, Han B (2015) Learning deconvolution network for semantic segmentation. In: The IEEE international conference on computer vision (ICCV)

NSTC (2016) Preparing for the future of artificial intelligence. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf. Accessed 10 Apr 2020

Ntoutsi E, Fafalios P, Gadiraju U, Iosifidis V, Nejdl W, Vidal ME, Ruggieri S, Turini F, Papadopoulos S, Krasanakis E, Kompatsiaris I, Kinder-Kurlanda K, Wagner C, Karimi F, Fernandez M, Alani H, Berendt B, Kruegel T, Heinze C, Broelemann K, Kasneci G, Tiropanis T, Staab S (2020) Bias in data-driven artificial intelligence systems—an introductory survey. WIREs Data Mining Knowl Discov 10(3):e1356. https://doi.org/10.1002/widm.1356

Oh BT, Lei S, Kuo CJ (2009) Advanced film grain noise extraction and synthesis for high-definition video coding. IEEE Trans Circ Syst Video Tech 19(12):1717–1729. https://doi.org/10.1109/TCSVT.2009.2026974

Ozcinar C, Smolic A (2018) Visual attention in omnidirectional video for virtual reality applications. In: 2018 10th international conference on quality of multimedia experience (QoMEX), pp 1–6. https://doi.org/10.1109/QoMEX.2018.8463418

Palmarini R, Erkoyuncu JA, Roy R, Torabmostaedi H (2018) A systematic review of augmented reality applications in maintenance. Robot Comput-Integr Manuf 49:215–228. https://doi.org/10.1016/j.rcim.2017.06.002

Panphattarasap P, Calway A (2018) Automated map reading: image based localisation in 2-D maps using binary semantic descriptors. In: IEEE/RSJ international conference on intelligent robots and systems (IROS), pp 6341–6348

Pawar PY, Gawande SH (2012) A comparative study on different types of approaches to text categorization. Int J Mach Learn Comput 2(4):423

Peng C, Xiao T, Li Z, Jiang Y, Zhang X, Jia K, Yu G, Sun J (2018) Megdet: A large mini-batch object detector. In: IEEE/CVF conference on computer vision and pattern recognition, pp 6181–6189

Perov I, Gao D, Chervoniy N, Liu K, Marangonda S, Umé C, Dpfks M, Facenheim CS, RP L, Jiang J, Zhang S, Wu P, Zhou B, Zhang W (2020) Deepfacelab: a simple, flexible and extensible face swapping framework. arXiv preprint arXiv:200505535v4

Pizer SM, Amburn EP, Austin JD, Cromartie R, Geselowitz A, Greer T, [ter Haar Romeny] B, Zimmerman JB, Zuiderveld K, (1987) Adaptive histogram equalization and its variations. Comput Vis, Graph, Image Process 39(3):355–368. https://doi.org/10.1016/S0734-189X(87)80186-X

Prabhakar KR, Srikar VS, Babu RV (2017) DeepFuse: a deep unsupervised approach for exposure fusion with extreme exposure image pairs. In: IEEE international conference on computer vision (ICCV), pp 4724–4732

Pu Y, Gan Z, Henao R, Yuan X, Li C, Stevens A, Carin L (2016) Variational autoencoder for deep learning of images, labels and captions. In: Advances in neural information processing systems, vol 29. Curran Associates, Inc., pp 2352–2360. http://papers.nips.cc/paper/6528-variational-autoencoder-for-deep-learning-of-images-labels-and-captions.pdf

Qi CR, Su H, Mo K, Guibas LJ (2017) Pointnet: deep learning on point sets for 3D classification and segmentation. In: The IEEE conference on computer vision and pattern recognition (CVPR)

Quesnel D, DiPaola S, Riecke B (2018) Deep learning for classification of peak emotions within virtual reality systems. In: International SERIES on information systems and management in creative media, pp 6–11

Radford A, Metz L, Chintala S (2016) Unsupervised representation learning with deep convolutional generative adversarial networks. In: International conference on learning representations

Razavi A, van den Oord A, Vinyals O (2019) Generating diverse high-resolution images with VQ-VAE. In: ICLR 2019 workshop DeepGenStruct

Redmon J, Farhadi A (2018) YOLOv3: an incremental improvement. arXiv:abs/1804.02767

Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 779–788

Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149

Rezaei-Ravari M, Eftekhari M, Saberi-Movahed F (2021) Regularizing extreme learning machine by dual locally linear embedding manifold learning for training multi-label neural network classifiers. Eng Appl Artif Intell 97:104062. https://doi.org/10.1016/j.engappai.2020.104062

Riedl M, Bulitko V (2012) Interactive narrative: a novel application of artificial intelligence for computer games. In: 16th AAAI conference on artificial intelligence

Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 234–241

Rosca M, Lakshminarayanan B, Mohamed S (2019) Distribution matching in variational inference. arXiv:1802.06847v4

Rowe J, Partridge D (1993) Creativity: a survey of AI approaches. Artif Intell Rev 7:43–70. https://doi.org/10.1007/BF00849197

Rumelhart D, Hinton G, Williams R (1986) Learning representations by back-propagating errors. Nature 323:533–536. https://doi.org/10.1038/323533a0

Rush AM, Chopra S, Weston J (2015) A neural attention model for abstractive sentence summarization. In: Proceedings of the 2015 conference on empirical methods in natural language processing, association for computational linguistics, Lisbon, Portugal, pp 379–389. https://doi.org/10.18653/v1/D15-1044

Russell S, Norvig P (2020) Artificial intelligence: a modern approach, 4th edn. Pearson

Rutishauser U, Walther D, Koch C, Perona P (2004) Is bottom-up attention useful for object recognition? In: IEEE computer society conference on computer vision and pattern recognition, vol 2, p II. https://doi.org/10.1109/CVPR.2004.1315142

Sabour S, Frosst N, Hinton GE (2017) Dynamic routing between capsules. In: Proceedings of the 31st international conference on neural information processing systems, pp 3859–3869

Sajjadi MSM, Schölkopf B, Hirsch M (2017) EnhanceNet: single image super-resolution through automated texture synthesis. In: IEEE international conference on computer vision (ICCV), pp 4501–4510

Sandfort V, Yan K, Pickhardt P, Summers R (2019) Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. Sci Rep 9(16884):1–9. https://doi.org/10.1038/s41598-019-52737-x

Sautoy MD (2019) The creativity code: art and innovation in the age of AI. Harvard University Press

Schiopu I, Huang H, Munteanu A (2020) CNN-based intra-prediction for lossless HEVC. IEEE Trans Circuits Syst Video Technol 30(7):1816–1828

Schuler CJ, Hirsch M, Harmeling S, Schölkopf B (2016) Learning to deblur. IEEE Trans Pattern Anal Mach Intell 38(7):1439–1451

See A, Liu PJ, Manning CD (2017) Get to the point: summarization with pointer-generator networks. In: Association for computational linguistics, 1073–1083

Shi J, Jiang X, Guillemot C (2020) Learning fused pixel and feature-based view reconstructions for light fields. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)

Shi W, Caballero J, Huszár F, Totz J, Aitken AP, Bishop R, Rueckert D, Wang Z (2016) Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 1874–1883

Shi X, Chen Z, Wang H, Yeung DY, Wong Wk, Woo Wc (2015) Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Proceedings of the 28th international conference on neural information processing systems, vol 1, p 802–810
Shillingford B, Assael Y, Hoffman MW, Paine T, Hughes C, Prabhu U, Liao H, Sak H, Rao K, Bennett L, Mulville M, Coppin B, Laurie B, Senior A, de Freitas N (2019) Large-scale visual speech recognition. In: INTERSPEECH
Shimada S, Golyanik V, Theobalt C, Stricker D (2019) ISMO-gan: Adversarial learning for monocular non-rigid 3d reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops
Shin Y, Cho Y, Pandey G, Kim A (2016) Estimation of ambient light and transmission map with common convolutional architecture. In: OCEANS 2016 MTS/IEEE Monterey, pp 1–7
Short T, Adams T (2017) Procedural generation in game design. Taylor & Francis Inc
Shorten C, Khoshgoftaar T (2019) A survey on image data augmentation for deep learning. J Big Data 6(60):1–48. https://doi.org/10.1186/s40537-019-0197-0
Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: International conference on learning representations
Siyao L, Zhao S, Yu W, Sun W, Metaxas DN, Loy CC, Liu Z (2021) Deep animation video interpolation in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)
Soccini AM (2017) Gaze estimation based on head movements in virtual reality applications using deep learning. In: IEEE virtual reality (VR), pp 413–414
Soltani AA, Huang H, Wu J, Kulkarni TD, Tenenbaum JB (2017) Synthesizing 3D shapes via modeling multi-view depth maps and silhouettes with deep generative networks. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 2511–2519
Song J, He T, Gao L, Xu X, Hanjalic A, Shen HT (2018a) Binary generative adversarial networks for image retrieval. In: 32nd AAAI conference on artificial intelligence
Song J, Zhang J, Gao L, Liu X, Shen HT (2018b) Dual conditional gans for face aging and rejuvenation. In: Proceedings of the 27th international joint conference on artificial intelligence, pp 899–905
Stankiewicz O (2019) Video coding technique with a parametric modelling of noise. Opto-Electron Rev 27(3):241–251. https://doi.org/10.1016/j.opelre.2019.05.006
Stanley KO, D'Ambrosio DB, Gauci J (2009) A hypercube-based encoding for evolving large-scale neural networks. Artif Life 15(2):185–212
Starke S, Zhang H, Komura T, Saito J (2019) Neural state machine for character-scene interactions. ACM Trans Graph 38(6):209. https://doi.org/10.1145/3355089.3356505
Starke S, Zhao Y, Komura T, Zaman K (2020) Local motion phases for learning multi-contact character movements. In: ACM SIGGRAPH
Sturm B, Santos JF, Ben-Tal O, Korshunova I (2016) Music transcription modelling and composition using deep learning. In: 1st conference on computer simulation of musical creativity
Su S, Delbracio M, Wang J, Sapiro G, Heidrich W, Wang O (2017) Deep video deblurring for hand-held cameras. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 237–246
Suarez PL, Sappa AD, Vintimilla BX (2017) Infrared image colorization based on a triplet DCGAN architecture. In: The IEEE conference on computer vision and pattern recognition (CVPR) workshops
Subramanian S, Rajeswar S, Sordoni A, Trischler A, Courville A, Pal C (2018) Towards text generation with adversarially learned neural outlines. In: NeurIPS 2018
Sun S, Pang J, Shi J, Yi S, Ouyang W (2018) Fishnet: A versatile backbone for image, region, and pixel level prediction. In: Advances in neural information processing systems, pp 760–770
Suwajanakorn S, Seitz SM, Kemelmacher-Shlizerman I (2017) Synthesizing Obama: learning lip sync from audio. ACM Trans Graph 36(4):1–13. https://doi.org/10.1145/3072959.3073640
Tai Y, Yang J, Liu X (2017) Image super-resolution via deep recursive residual network. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 2790–2798
Tang G, Zhao L, Jiang R, Zhang X (2019) Single image dehazing via lightweight multi-scale networks. In: IEEE international conference on big data (big data), pp 5062–5069
Tao L, Zhu C, Xiang G, Li Y, Jia H, Xie X (2017) Llcnn: a convolutional neural network for low-light image enhancement. In: IEEE visual communications and image processing (VCIP), pp 1–4
Tao X, Gao H, Shen X, Wang J, Jia J (2018) Scale-recurrent network for deep image deblurring. In: IEEE/CVF conference on computer vision and pattern recognition, pp 8174–8182
Tesfaldet M, Brubaker MA, Derpanis KG (2018) Two-stream convolutional networks for dynamic texture synthesis. In: The IEEE conference on computer vision and pattern recognition (CVPR)

Tewari A, Zollhöfer M, Bernard F, Garrido P, Kim H, Pérez P, Theobalt C (2020) High-fidelity monocular face reconstruction based on an unsupervised model-based face autoencoder. IEEE Trans Pattern Anal Mach Intell 42(2):357–370
Theis L, Korshunova I, Tejani A, Huszár F (2018) Faster gaze prediction with dense networks and fisher pruning. arXiv:1801.05787v2
Tian C, Fei L, Zheng W, Xu Y, Zuo W, Lin CW (2020) Deep learning on image denoising: an overview. Neural Netw 131:251–275. https://doi.org/10.1016/j.neunet.2020.07.025
Tian Y, Peng X, Zhao L, Zhang S, Metaxas DN (2018) Cr-gan: Learning complete representations for multi-view generation. In: International joint conference on artificial intelligence
Torrejon OE, Peretti N, Figueroa R (2020) Rotoscope automation with deep learning. SMPTE Mot Imaging J 129(2):16–26
Truşcă M, Wassenberg D, Frasincar F, Dekker R (2020) A hybrid approach for aspect-based sentiment analysis using deep contextual word embeddings and hierarchical attention. In: International conference on web engineering, vol 12128. https://doi.org/10.1007/978-3-030-50578-3_25
Ummenhofer B, Zhou H, Uhrig J, Mayer N, Ilg E, Dosovitskiy A, Brox T (2017) DeMoN: depth and motion network for learning monocular stereo. In: The IEEE conference on computer vision and pattern recognition (CVPR)
Vasudevan AB, Dai D, Gool LV (2020) Semantic object prediction and spatial sound super-resolution with binaural sounds. In: European conference on computer vision
Venugopalan S, Xu H, Donahue J, Rohrbach M, Mooney R, Saenko K (2015) Translating videos to natural language using deep recurrent neural networks. In: Conference of the North American chapter of the association for computational linguistics—human language technologies
Vesperini F, Gabrielli L, Principi E, Squartini S (2019) Polyphonic sound event detection by using capsule neural networks. IEEE J Sel Top Signal Process 13(2):310–322. https://doi.org/10.1109/JSTSP.2019.2902305
Wan C, Probst T, Van Gool L, Yao A (2017) Crossing nets: combining GANs and VAEs with a shared latent space for hand pose estimation. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 1196–1205
Wan J, Wang D, Hoi SCH, Wu P, Zhu J, Zhang Y, Li J (2014) Deep learning for content-based image retrieval: A comprehensive study. In: Proceedings of the 22nd ACM international conference on multimedia, association for computing machinery, New York, NY, USA, MM '14, pp 157–166. https://doi.org/10.1145/2647868.2654948
Wang C, Dong S, Zhao X, Papanastasiou G, Zhang H, Yang G (2020a) Saliencygan: deep learning semisupervised salient object detection in the fog of iot. IEEE Trans Ind Inf 16(4):2667–2676. https://doi.org/10.1109/TII.2019.2945362
Wang H, Su D, Liu C, Jin L, Sun X, Peng X (2019a) Deformable non-local network for video super-resolution. IEEE Access 7:177734–177744
Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, Van Gool L (2016) Temporal segment networks: Towards good practices for deep action recognition. In: The European conference on computer vision (ECCV), pp 20–36
Wang P, Rowe J, Min W, Mott B, Lester J (2017) Interactive narrative personalization with deep reinforcement learning. In: International joint conference on artificial intelligence
Wang Q, Zhang L, Bertinetto L, Hu W, Torr PHS (2019b) Fast online object tracking and segmentation: A unifying approach. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 1328–1338. https://doi.org/10.1109/CVPR.2019.00142
Wang TC, Liu MY, Zhu JY, Liu G, Tao A, Kautz J, Catanzaro B (2018) Video-to-video synthesis. In: Advances in neural information processing systems (NeurIPS)
Wang W, Lai Q, Fu H, Shen J, Ling H, Yang R (2021) Salient object detection in the deep learning era: an in-depth survey. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2021.3051099
Wang X, Chan KC, Yu K, Dong C, Loy CC (2019) EDVR: video restoration with enhanced deformable convolutional networks. In: The IEEE conference on computer vision and pattern recognition (CVPR) workshops
Wang Y, Perazzi F, McWilliams B, Sorkine-Hornung A, Sorkine-Hornung O, Schroers C (2018) A fully progressive approach to single-image super-resolution. In: IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), pp 977–97709. https://doi.org/10.1109/CVPRW.2018.00131
Wang Z, Chen J, Hoi SCH (2020b) Deep learning for image super-resolution: a survey. IEEE Trans Pattern Anal Mach Intell 1

Wei SE, Saragih J, Simon T, Harley AW, Lombardi S, Perdoch M, Hypes A, Wang D, Badino H, Sheikh Y (2019) Vr facial animation via multiview image translation. ACM Trans Graph 38(4):1–16. https://doi.org/10.1145/3306346.3323030

Welser J, Pitera JW, Goldberg C (2018) Future computing hardware for AI. In: IEEE international electron devices meeting (IEDM), pp 1.3.1–1.3.6

Woo S, Park J, Lee JY, Kweon IS (2018) CBAM: convolutional block attention module. In: The European conference on computer vision (ECCV), pp 3–19

Wright C, Allnutt J, Campbell R, Evans M, Forman R, Gibson J, Jolly S, Kerlin L, Lechelt S, Phillipson G, Shotton M (2020) AI in production: video analysis and machine learning for expanded live events coverage. SMPTE Mot Imaging J 129(2):36–45

Wu H, Zheng S, Zhang J, Huang K (2019) GP-GAN: towards realistic high-resolution image blending. In: ACM international conference on multimedia

Wu J, Yu Y, Huang C, Yu K (2015) Deep multiple instance learning for image classification and auto-annotation. In: The IEEE conference on computer vision and pattern recognition (CVPR)

Wu J, Wang Y, Xue T, Sun X, Freeman B, Tenenbaum J (2017) Marrnet: 3D shape reconstruction via 2.5d sketches. In: Advances in Neural Information Processing Systems, vol 30, pp 540–550. https://proceedings.neurips.cc/paper/2017/file/ad972f10e0800b49d76fed33a21f6698-Paper.pdf

Xia Y, Wang J (2005) A recurrent neural network for solving nonlinear convex programs subject to linear constraints. IEEE Trans Neural Netw 16(2):379–386

Xiangyu Xu WS Muchen Li (2019) Learning deformable kernels for image and video denoising. arXiv:1904.06903

Xie H, Yao H, Sun X, Zhou S, Zhang S (2019) Pix2Vox: context-aware 3D reconstruction from single and multi-view images. In: IEEE/CVF international conference on computer vision (ICCV), pp 2690–2698

Xie J, Xu L, Chen E (2012) Image denoising and inpainting with deep neural networks. In: Advances in neural information processing systems, vol 25. Curran Associates, Inc., pp 341–349. http://papers.nips.cc/paper/4686-image-denoising-and-inpainting-with-deep-neural-networks.pdf

Xie J, Girshick R, Farhadi A (2016) Deep3D: fully automatic 2D-to-3D video conversion with deep convolutional neural networks. In: The European conference on computer vision (ECCV). Springer, Cham, pp 842–857

Xie Y, Zhang W, Tao D, Hu W, Qu Y, Wang H (2016) Removing turbulence effect via hybrid total variation and deformation-guided kernel regression. IEEE Trans Image Process 25(10):4943–4958

Xu A, Liu Z, Guo Y, Sinha V, Akkiraju R (2017a) A new chatbot for customer service on social media. In: Proceedings of the 2017 CHI conference on human factors in computing systems, association for computing machinery, New York, NY, USA, CHI '17, pp 3506–3510. https://doi.org/10.1145/3025453.3025496

Xu J, Yao T, Zhang Y, Mei T (2017b) Learning multimodal attention LSTM networks for video captioning. In: Proceedings of the 25th ACM international conference on multimedia, association for computing machinery, New York, NY, USA, MM '17, p 537–545. https://doi.org/10.1145/3123266.3123448

Xu L, Sun H, Liu Y (2019) Learning with batch-wise optimal transport loss for 3D shape recognition. In: The IEEE conference on computer vision and pattern recognition (CVPR)

Xu M, Li C, Zhang S, Callet PL (2020) State-of-the-art in 360° video/image processing: perception, assessment and compression. IEEE J Sel Top Signal Process 14(1):5–26. https://doi.org/10.1109/JSTSP.2020.2966864

Xu Z, Wang T, Fang F, Sheng Y, Zhang G (2020) Stylization-based architecture for fast deep exemplar colorization. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 9360–9369. https://doi.org/10.1109/CVPR42600.2020.00938

Xue T, Chen B, Wu J, Wei D, Freeman WT (2019) Video enhancement with task-oriented flow. Int J Comput Vis 127:1106–1125

Xue Y, Su J (2019) Attention based image compression post-processing convolutional neural network. In: IEEE/CVF conference on computer vision and pattern recognition workshop (CVPRW)

Yahya AA, Tan J, Su B, Liu K (2016) Video denoising based on spatial-temporal filtering. In: 6th intern. conf. on digital home, pp 34–37. https://doi.org/10.1109/ICDH.2016.017

Yang B, Wen H, Wang S, Clark R, Markham A, Trigoni N (2017) 3D object reconstruction from a single depth view with adversarial learning. In: Proceedings of the IEEE international conference on computer vision (ICCV) workshops

Yang D, Sun J (2018) Proximal Dehaze-Net: a prior learning-based deep network for single image dehazing. In: The European conference on computer vision (ECCV)

Yang F, Chang X, Dang C, Zheng Z, Sakti S, SN, Wu Y (2020a) ReMOTS: self-supervised refining multi-object tracking and segmentation. arXiv:2007.03200v2
Yang J, Hong Z, Qu X, Wang J, Xiao J (2020b) NAS-YODO. http://host.robots.ox.ac.uk:8080/leaderboard/displaylb_main.php?challengeid=11&compid=3#KEY_NAS%20Yolo
Yang Q, Yan P, Zhang Y, Yu H, Shi Y, Mou X, Kalra MK, Zhang Y, Sun L, Wang G (2018) Low-dose ct image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. IEEE Trans Med Imaging 37(6):1348–1357
Yang W, Zhang X, Tian Y, Wang W, Xue J, Liao Q (2019) Deep learning for single image super-resolution: a brief review. IEEE Trans Multimed 21(12):3106–3121
Yao G, Lei T, Zhong J (2019) A review of convolutional-neural-network-based action recognition. Pattern Recogn Lett 118:14–22. https://doi.org/10.1016/j.patrec.2018.05.018. Cooperative and Social Robots: Understanding Human Activities and Intentions
Yi K, Guo Y, Wang Z, Sun L, Zhu W (2020) Personalized text summarization based on gaze patterns. In: 2020 IEEE conference on multimedia information processing and retrieval (MIPR), pp 307–313. https://doi.org/10.1109/MIPR49039.2020.00070
Yi Z, Zhang H, Tan P, Gong M (2017) DualGAN: unsupervised dual learning for image-to-image translation. In: IEEE international conference on computer vision (ICCV), pp 2868–2876. https://doi.org/10.1109/ICCV.2017.310
Young T, Hazarika D, Poria S, Cambria E (2018) Recent trends in deep learning based natural language processing [review article]. IEEE Comput Intell Maga 13(3):55–75
Yu F, Koltun V (2016) Multi-scale context aggregation by dilated convolutions. In: International conference on learning representations
Yu J, Lin Z, Yang J, Shen X, Lu X, Huang TS (2018) Generative image inpainting with contextual attention. In: The IEEE conference on computer vision and pattern recognition (CVPR)
Yu J, Lin Z, Yang J, Shen X, Lu X, Huang T (2019) Free-form image inpainting with gated convolution. In: IEEE/CVF international conference on computer vision (ICCV), pp 4470–4479. https://doi.org/10.1109/ICCV.2019.00457
Zakharov E, Shysheya A, Burkov E, Lempitsky V (2019) Few-shot adversarial learning of realistic neural talking head models. In: 2019 IEEE/CVF international conference on computer vision (ICCV), pp 9458–9467. https://doi.org/10.1109/ICCV.2019.00955
Zhang C, Li Y, Du N, Fan W, Yu P (2019a) Joint slot filling and intent detection via capsule neural networks. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 5259–5267. https://doi.org/10.18653/v1/P19-1519
Zhang F, Afonso M, Bull D (2019b) ViSTRA2: video coding using spatial resolution and effective bit depth adaptation. arXiv:1911.02833
Zhang F, Prisacariu V, Yang R, Torr PHS (2019) GA-Net: guided aggregation net for end-to-end stereo matching. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 185–194. https://doi.org/10.1109/CVPR.2019.00027
Zhang F, Chen F, Bull DR (2020) Enhancing VVC through CNN-based Post-Processing. In: IEEE ICME
Zhang G (2020) Design of virtual reality augmented reality mobile platform and game user behavior monitoring using deep learning. Int J Electr Eng Edu. https://doi.org/10.1177/0020720920931079
Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X, Metaxas D (2017) StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: IEEE international conference on computer vision (ICCV), pp 5908–5916
Zhang H, Goodfellow I, Metaxas D, Odena A (2019) Self-attention generative adversarial networks. In: Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, CA, USA, Proceedings of machine learning research, vol 97, pp 7354–7363
Zhang J, Pan J, Ren J, Song Y, Bao L, Lau RW, Yang MH (2018) Dynamic scene deblurring using spatially variant recurrent neural networks. In: The IEEE conference on computer vision and pattern recognition (CVPR)
Zhang K, Zuo W, Chen Y, Meng D, Zhang L (2017) Beyond a gaussian denoiser: residual learning of deep CNN for image denoising. IEEE Trans Image Process 26(7):3142–3155
Zhang K, Zuo W, Zhang L (2018) FFDNet: toward a fast and flexible solution for cnn-based image denoising. IEEE Trans Image Process 27(9):4608–4622
Zhang R, Isola P, Efros AA (2016) Colorful image colorization. In: The European conference on computer vision (ECCV), pp 649–666
Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y (2018a) Image super-resolution using very deep residual channel attention networks. In: The European conference on computer vision (ECCV). Springer, Cham, pp 294–310

Zhang Z, Geiger J, Pohjalainen J, Mousa AED, Jin W, Schuller B (2018b) Deep learning for environmentally robust speech recognition: an overview of recent developments. ACM Trans Intell Syst Technol 9(5):1–26. https://doi.org/10.1145/3178115

Zhao H, Shao W, Bao B, Li H (2019a) A simple and robust deep convolutional approach to blind image denoising. In: IEEE/CVF international conference on computer vision workshop (ICCVW), pp 3943–3951

Zhao L, Wang S, Zhang X, Wang S, Ma S, Gao W (2019b) Enhanced motion-compensated video coding with deep virtual reference frame generation. IEEE Trans Image Process 28(10):4832–4844

Zhao W, Peng H, Eger S, Cambria E, Yang M (2019) Towards scalable and reliable capsule networks for challenging NLP applications. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 1549–1559. https://doi.org/10.18653/v1/P19-1150

Zhao Z, Wang S, Wang S, Zhang X, Ma S, Yang J (2019a) Enhanced bi-prediction with convolutional neural network for high-efficiency video coding. IEEE Trans Circuits Syst Video Technol 29(11):3291–3301

Zhao Z, Zheng P, Xu S, Wu X (2019b) Object detection with deep learning: a review. IEEE Trans Neural Netw Learn Syst 30(11):3212–3232

Zhen M, Wang J, Zhou L, Fang T, Quan L (2019) Learning fully dense neural networks for image semantic segmentation. In: 33rd AAAI conference on artificial intelligence (AAAI-19)

Zhou S, Zhang J, Pan J, Zuo W, Xie H, Ren J (2019) Spatio-temporal filter adaptive network for video deblurring. In: IEEE/CVF international conference on computer vision (ICCV), pp 2482–2491

Zhou T, Fan D, Cheng M, Shen J, Shao L (2021) RGB-D salient object detection: a survey. Comput Vis Media. https://doi.org/10.1007/s41095-020-0199-z

Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: The IEEE international conference on computer vision (ICCV)

Zhu X, Milanfar P (2013) Removing atmospheric turbulence via space-invariant deconvolution. IEEE Trans Pattern Anal Mach Intell 35(1):157–170

Zhu X, Liu Y, Li J, Wan T, Qin Z (2018) Emotion classification with data augmentation using generative adversarial networks. In: Advances in knowledge discovery and data mining. Springer, Cham, pp 349–360

Zollhöfer M, Stotko P, Görlitz A, Theobalt C, Nießner M, Klein R, Kolb A (2018) State of the art on 3D reconstruction with RGB-D cameras. Eurographics 37(2):625–652. https://doi.org/10.1111/cgf.13386

Zuo C, Liu Y, Tan X, Wang W, Zhang M (2013) Video denoising based on a spatiotemporal Kalman-bilateral mixture model. Sci World J. https://doi.org/10.1155/2013/438147