Emerging Frontiers in Computer Vision: A Critical Analysis of Deep Learning Techniques and Their Real-World Applications

Dr. Ronith A. Velkar
Department of Artificial Intelligence, University of Tartu, Estonia

Dr. Elinora S. Kavesh
School of Computing, National University of Singapore (NUS), Singapore

V0LUME02 ISSUE02 (2025)

Published Date: 03 July 2025 // Page no.: - 01-11

ABSTRACT

Deep learning has become the cornerstone of modern computer vision, fundamentally transforming how machines perceive and interpret the visual world. This article presents a critical review of the key deep learning techniques that have driven this revolution. We trace the evolution of foundational concepts, from early neural networks to the sophisticated convolutional neural network (CNN) architectures that dominate the field today. The article is structured to provide a comprehensive overview, beginning with an introduction to the core concepts and historical context of deep learning in computer vision. We then delve into the methodologies, systematically examining influential architectures and techniques for major computer vision tasks, including image classification, object detection, semantic segmentation, and image restoration. Subsequently, we evaluate the performance and results of these methods, highlighting their groundbreaking impact on various application scenarios, from medical imaging to autonomous systems. Finally, we discuss the broader implications, current challenges such as the creation of deepfakes, and promising future directions for research and development. By synthesizing a wide array of seminal and contemporary works, this review offers a detailed landscape of the field, providing valuable insights for both new and experienced researchers.

Keywords: Deep Learning, Computer Vision, Convolutional Neural Networks (CNN), Object Detection, Semantic Segmentation, Image Restoration, Deepfakes, Artificial Intelligence.

INTRODUCTION

The ability to imbue machines with human-like vision has been a long-standing goal in the field of artificial intelligence. For decades, progress in computer vision (CV) was steady but incremental, relying on a variety of machine learning techniques that required meticulous, manual feature engineering. Methods such as the K-Nearest Neighbor (KNN) algorithm, Support Vector Machines (SVMs), and ensemble methods like Boosting and Random Forests were staples in the field [60]. In the realm of object detection, a landmark approach was the Viola-Jones detector, which used Haar-like features and a cascade of classifiers trained with Adaboost to achieve real-time face detection, a significant milestone at the time [85, 48]. However, these traditional methods were often brittle, struggling to generalize across the vast diversity of the visual world. Their performance was fundamentally limited by the quality and ingenuity of the handcrafted features they depended upon.

The paradigm shifted dramatically with the ascendancy of deep learning (DL), a subfield of machine learning characterized by neural networks with many layers (or "deep" architectures) [29, 6]. Unlike their predecessors, deep learning models, and particularly Convolutional Neural Networks (CNNs), possess the remarkable ability to automatically and hierarchically learn discriminative features directly from raw pixel data. This capability to bypass manual feature engineering and learn powerful data representations has been the single most important catalyst for the current revolution in computer vision [25]. The initial adoption of DL in CV was hampered by practical limitations, including insufficient computational power (CPU and GPU) and memory constraints, which made training deep networks on large datasets infeasible [58].

This all changed in 2012 with the stunning success of **AlexNet** in the ImageNet Large Scale Visual Recognition Challenge [38]. AlexNet, a deep CNN, outperformed all competing traditional methods by a staggering margin, signaling a definitive changing of the guard. This event ignited an explosion of research and development, solidifying CNNs as the dominant architecture for visual tasks [25]. In the years that followed, a series of increasingly sophisticated architectures were introduced, each pushing the boundaries of what was possible. These include the very deep **VGGNet** [76], the computationally efficient **GoogLeNet** with its novel Inception modules [80, 81], and the groundbreaking **ResNet**, which enabled the training of networks hundreds of layers deep by introducing residual connections [29].

The influence of these foundational models has permeated every corner of computer vision. This review provides a comprehensive analysis of this deep learning-driven landscape. We structure our analysis around eight emerging and foundational techniques: **AlexNet, VGGNet, GoogLeNet & Inception, ResNet, DenseNet, MobileNets, EfficientNet, and RegNet** [87]. We then explore how these core architectures and the principles they embody have been adapted to solve four key application scenarios: **(1) Recognition**, encompassing both image classification and object detection; **(2) Visual Tracking**; **(3) Semantic Segmentation**; and **(4) Image Restoration** [21, 88].

Furthermore, this article discusses the broader implications and future trajectory of the field. While the progress has been immense, it is not without its challenges. The "black box" nature of these models raises concerns about interpretability, and the malicious use of deep learning to create hyper-realistic "deepfakes" presents a significant societal threat [7, 2, 3, 33]. By critically examining the techniques, applications, challenges, and future trends, this review aims to provide a holistic and in-depth understanding of the state of the art in deep learning for computer vision.

2. Methodologies: Core Architectures and Techniques

The engine of progress in deep learning-powered computer vision is the continuous innovation in neural network architectures. This section details the foundational CNN models that established the modern paradigm and then explores how these concepts have been applied and adapted to key CV tasks.

2.1. Foundational CNN Architectures: The Pillars of Modern Vision

The evolution of CNN architectures can be seen as a quest for greater accuracy, efficiency, and depth. Eight models, in particular, represent key milestones in this journey.

- **AlexNet:** The model that started the revolution, AlexNet, consisted of five convolutional layers followed by three fully connected layers [38]. Its success was not just due to its depth but also its clever use of then-novel techniques. It was among the first to effectively use the Rectified Linear Unit (ReLU) activation function, which helped mitigate the vanishing gradient problem and sped up training compared to traditional sigmoid or tanh functions. It also employed overlapping pooling and data augmentation to improve performance and reduce overfitting. The model was so computationally demanding for its time that it had to be trained across two GPUs, a testament to the hardware limitations of the era [58].

- **VGGNet:** Following AlexNet, researchers at the Visual Geometry Group (VGG) at Oxford explored a simple yet powerful hypothesis: that depth is the critical component for performance [76]. VGGNets, most famously VGG-16 and VGG-19, are characterized by their extreme simplicity and uniformity, using only small 3x3 convolutional filters stacked on top of each other. By stacking these small filters, the network can achieve a larger receptive field with fewer parameters than a single large filter, while also incorporating more non-linearities. While VGGNets achieved state-of-the-art results, their depth came at a cost: VGG-16 has around 138 million parameters, leading to very large model sizes and high computational demands during training and inference.

- **GoogLeNet & the Inception Architecture:** While VGG pursued sheer depth, researchers at Google took a different approach, focusing on computational efficiency. Their key innovation was the **Inception module** [80, 81]. Instead of choosing a single filter size for a layer, the Inception module performs multiple convolutions (1x1, 3x3, 5x5) and a max-pooling operation in parallel and concatenates their outputs. This allows the network to capture features at multiple scales simultaneously. To manage the computational cost, they drew inspiration from the "Network in Network" (NIN) paper [50], strategically using 1x1 convolutions as bottleneck layers to reduce the feature map depth before the more expensive 3x5 and 5x5 convolutions. The resulting 22-layer network, dubbed GoogLeNet, achieved better performance than VGGNet with only a fraction of the parameters (around 5 million). The architecture was later refined in Inception-v2 and Inception-v3, which introduced ideas like factorizing larger convolutions into smaller ones and using Batch Normalization for more stable training [81]. Later, Inception-v4 and Inception-ResNet combined the Inception architecture with the residual connections from ResNet [79], and Xception further improved efficiency by replacing Inception modules with depthwise separable convolutions [14].

- **ResNet (Residual Network):** As networks got deeper, they encountered a new problem: degradation. Deeper networks would often perform worse than their shallower counterparts, not because of overfitting, but because they were harder to optimize. He et al. (2016) tackled this with a brilliantly simple idea: the **residual block** [29]. Instead of forcing a stack of layers to learn an underlying mapping H(x), ResNet reformulates it to learn a residual mapping F(x) = H(x) - x. The original mapping is then recast as F(x) + x. This is implemented via "shortcut" or "skip" connections that carry the identity mapping, skipping over one or more layers. This framework makes it easier for layers to learn an identity mapping if needed (by driving the weights to zero), allowing gradients to flow more freely and enabling the training of networks of unprecedented depth (e.g., 152 layers or more) while still achieving lower error rates.

- **DenseNet (Densely Connected Convolutional Network):** Taking the idea of shortcut connections to its logical extreme, DenseNet connects every layer to every other layer in a feed-forward fashion [36]. In a dense block, the input to any given layer is the concatenation of the feature maps from all preceding layers. This approach encourages massive feature reuse, strengthens feature propagation, and alleviates the vanishing gradient problem. Because features are reused so extensively, DenseNets can be very parameter-efficient, achieving state-of-the-art results with fewer parameters than comparable ResNets.

- **MobileNets:** As deep learning models began to move from servers to edge devices like smartphones, a new premium was placed on efficiency. MobileNets were designed specifically for mobile and embedded vision applications [32]. Their core building block is the **depthwise separable convolution**, which factors a standard convolution into two parts: a depthwise convolution that applies a single filter to each input channel, and a pointwise (1x1) convolution that combines the outputs of the depthwise convolution. This factorization dramatically reduces computation and model size. MobileNetV2 improved upon this by introducing inverted residuals and linear bottlenecks to enhance performance and memory efficiency [71]. MobileNetV3 further refined the architecture using Neural Architecture Search (NAS) to find optimal configurations [30].

- **EfficientNet:** The creators of EfficientNet observed that previous works typically scaled networks in an ad-hoc manner, increasing only one dimension: depth (more layers), width (more channels), or image resolution [82]. They proposed that a more principled approach is to balance all three dimensions using a fixed scaling factor. They developed a baseline network using NAS and then applied their "compound scaling" method to create a family of models (EfficientNet-B0 to B7) that achieve state-of-the-art accuracy with significantly fewer parameters and computations than previous models.

- **RegNet:** Rather than designing a single network or searching for one, the RegNet approach focuses on designing the *network design space* itself [65]. By analyzing what makes a good design space, they derived principles that led them to a simple, low-dimensional space of networks they call "RegNet." Models sampled from this space are simple and fast, outperforming efficient models like EfficientNet while being up to five times faster on GPUs, offering a new paradigm for network design.

2.2. Key Application Methodologies

The architectural principles described above form the "backbone" for a wide range of more specialized models designed for specific CV tasks.

2.2.1. Object Detection

Object detection goes beyond classification by also localizing objects with bounding boxes. Modern detectors can be understood as having three components: a backbone (a standard CNN like ResNet for feature extraction), a neck (which aggregates features from the backbone, like a Feature Pyramid Network or FPN [52]), and a head (which performs the final classification and bounding box regression). Detection methods largely fall into two categories:

- **Two-Stage Detectors:** These methods first generate a sparse set of candidate "regions of interest" (RoIs) and then run a classifier on these proposals. The pioneering **R-CNN** [24] was slow because it ran a CNN for every proposal. **Fast R-CNN** [23] improved speed by sharing computation, running the CNN once on the whole image and then pooling features for each RoI. **Faster R-CNN** [69] made the process end-to-end by replacing the slow external region proposal algorithm with a lightweight neural network called the Region Proposal Network (RPN). The pinnacle of this family is **Mask R-CNN** [27], which extends Faster R-CNN by adding a parallel branch that predicts a segmentation mask for each detected object, unifying object detection and instance segmentation.

- **One-Stage Detectors:** These methods skip the region proposal step and predict bounding boxes and classes directly in a single pass. The **YOLO (You Only Look Once)** family is famous for its incredible speed, making it ideal for real-time applications [67]. It divides the image into a grid and predicts bounding boxes and probabilities for each grid cell. It has seen numerous improvements, with YOLOv2 (YOLO9000) adding batch normalization and anchor boxes [68], YOLOv3 using a deeper Darknet-53 backbone [69], and community-driven efforts like YOLOv4 introducing a host of modern optimizations ("bag of freebies" and "bag of specials") to achieve an optimal balance of speed and accuracy [8]. Another key one-stage model is the **SSD (Single Shot MultiBox Detector)** [54], which uses feature maps from multiple layers of the backbone to detect objects at different scales. **RetinaNet** [53] addressed a major weakness of one-stage detectors—the extreme class imbalance between foreground and background—by introducing a novel "Focal Loss" function that focuses training on hard-to-classify examples.

2.2.2. Visual Tracking

Visual tracking involves following a specified object through a video sequence. This task is challenging due to factors like changes in object appearance, lighting, and occlusions [40]. Deep learning has largely replaced traditional methods. Early DL approaches used stacked autoencoders to learn feature representations offline [86]. Modern trackers often employ Siamese networks, like SiamRPN [47], which learn a similarity function to compare an initial template of the target with candidate regions in subsequent frames. Other advanced methods have incorporated reinforcement learning to train an agent to decide how to move the tracking box [93] or used adversarial learning to make the tracker more robust to distractors [77].

2.2.3. Semantic Segmentation

This task involves assigning a class label to every pixel in an image. The breakthrough came with Fully Convolutional Networks (FCNs) [56], which replaced the dense, fully connected layers of classification networks with 1x1 convolutions, allowing them to output a heatmap or segmentation map instead of a single class label. Building on this, the U-Net architecture introduced a symmetric encoder-decoder structure with "skip connections" linking the contracting path (encoder) to the expanding path (decoder) [70]. This allows the decoder to recover fine-grained spatial detail lost during downsampling, making U-Net and its variants like UNet++ [98] extremely popular, especially for biomedical image segmentation where precise boundaries are critical [61]. The DeepLab family of models [12, 13] introduced atrous (or dilated) convolutions, which allow the network to control the spatial resolution of feature maps and expand the field of view without increasing parameters, proving highly effective for general-purpose semantic segmentation.

2.2.4. Image Restoration

This broad category includes tasks like image denoising and super-resolution (SR). For SR, the SRCNN model was an early and influential deep learning approach that learned a direct end-to-end mapping from low-resolution to high-resolution images [18]. Later, SRGAN used a Generative Adversarial Network (GAN) to produce much more photo-realistic results that looked perceptually more convincing, even if they sometimes had lower peak signal-to-noise ratio (PSNR) scores [44]. In denoising, while early networks were trained on pairs of noisy and clean images [9], a major innovation was the concept of learning from corrupted data alone. Noise2Noise [46] showed that a network could learn to restore images by training on pairs of noisy images of the same underlying scene. Even more impressively, Noise2Void [39] demonstrated that a network can learn to denoise from single noisy images, a powerful technique for domains where clean data is unavailable.

3. Results and Applications

The application of the methods described above has led to transformative results across countless domains, moving computer vision from the research lab into the fabric of daily life and industry.

3.1. Performance Breakthroughs

On standard academic benchmarks, deep learning models have consistently shattered records.

- In **image classification**, the error rate on the ImageNet dataset fell from over 25% in the pre-deep learning era to under 3% with advanced models like EfficientNet and ResNet, surpassing human-level performance [82, 29].

- In **object detection**, one-stage detectors like YOLOv4 achieve real-time speeds (~65 FPS on a Tesla V100) with high accuracy (43.5% AP on the COCO dataset) [8], while two-stage detectors like Mask R-CNN provide state-of-the-art instance segmentation results (62.3% AP_{50}) [27]. The competition between these approaches has driven rapid progress, with YOLOv3 proving to be significantly faster (3.8x) than RetinaNet for similar accuracy levels on some metrics [53].

- In **semantic segmentation**, DeepLabv3+ achieved a mean Intersection over Union (mIoU) of 89% on the PASCAL VOC 2012 benchmark, showcasing its powerful ability to delineate object boundaries [13].

- In **visual tracking**, methods like D3S have pushed performance on tracking benchmarks to over 72% AUC [54], while adversarial learning in VITAL trackers has improved robustness to challenging scenarios [77].

3.2. Proliferation of Real-World Applications

These performance gains have unlocked a vast array of real-world applications:

- **Healthcare and Medical Imaging:** Deep learning is revolutionizing medical diagnostics. CNNs, particularly U-Net and its derivatives like V-Net for 3D data and Lu-Net, are routinely used to segment tumors in brain MRIs, identify nodules in lung CT scans, and analyze retinal fundus images for diabetic retinopathy [70, 61, 66, 60]. Transfer learning approaches are also highly effective for tasks like classifying different types of brain tumors from limited medical data [60].

- **Autonomous Systems:** The perception systems of self-driving cars, drones, and robots rely heavily on deep learning. Real-time object detection (YOLO, SSD) [67, 54], semantic segmentation (for understanding lanes, sidewalks, and drivable areas) [56], and visual tracking [40] are all critical components for safe navigation in complex and dynamic environments.

- **Security, Forensics, and Content Moderation:** Deep learning models are used for everything from biometric authentication (face recognition) to large-scale content moderation on social media. A significant emerging challenge is the detection of **deepfakes**. These AI-generated videos can be used to spread misinformation or for malicious impersonation. This has spurred a new field of research focused on creating detectors that can spot subtle artifacts, such as inconsistencies between spoken phonemes and visual visemes [3] or unnatural behavioral patterns [2]. High-profile initiatives like the Deepfake Detection Challenge (DFDC) on Kaggle have been instrumental in advancing the state of the art in this area [36].

- **Creative Industries and Content Enhancement:** Generative Adversarial Networks (GANs) have unlocked incredible creative potential. They are used for artistic style transfer, generating novel imagery, and enhancing existing content. Models like SRGAN can perform "photo-realistic" super-resolution, restoring old photographs or upscaling low-resolution video [44]. JSI-GAN can perform joint super-resolution and inverse tone-mapping, converting standard video to stunning High Dynamic Range (HDR) [37]. EventSR even demonstrates the ability to reconstruct high-quality images from the sparse data of event-based cameras [83].

- **Scientific and Historical Discovery:** The power of computer vision is being applied in increasingly diverse domains. In archaeology and history, deep learning is being used to virtually "unfold" and read sealed historical documents that are too fragile to be opened physically, using data from X-ray microtomography [17]. In construction, CV techniques are used to monitor on-site progress and improve safety [94].

4. Discussion, Challenges, and Future Directions

The trajectory of deep learning in computer vision has been one of explosive growth and undeniable success. By analyzing the evolution of the field, we can identify key trends, persistent challenges, and exciting directions for future research.

4.1. The Three Stages of Development

The past decade of progress can be broadly categorized into three overlapping stages:

1. **The Early Stage (c. 2012–2016):** This era was defined by the validation of deep CNNs as the superior paradigm for vision tasks. It began with AlexNet's victory and was characterized by a focus on improving raw accuracy in image classification through architectural innovation. Models like VGGNet demonstrated the power of depth [76], while GoogLeNet showed the benefits of efficient, multi-scale design [80]. This stage culminated with ResNet, which solved the deep network optimization problem and surpassed human performance on ImageNet, solidifying CNNs as the backbone for nearly all subsequent work [29].

2. **The Middle Stage (c. 2016–2019):** With classification largely considered a solved problem, research began to branch out and mature in two key directions. First was the penetration of deep learning into more complex application scenarios like object detection (with the maturation of Faster R-CNN and YOLO) [69, 68], semantic segmentation, and visual tracking. Second was the rise of **efficiency**. As models were deployed on resource-constrained devices, architectures like MobileNet, ShuffleNet, and SqueezeNet prioritized low latency and small model size, making on-device AI a reality [32].

3. **The Latest Stage (c. 2019–Present):** This current stage is characterized by refinement, automation, and broadening horizons. Models like EfficientNet and RegNet have introduced more principled methods for network design and scaling [82, 65]. We are also seeing a trend towards semi-supervised and unsupervised learning, tackling the reliance on massive labeled datasets [46, 39]. Furthermore, the field is becoming more specialized and combinatorial, tackling niche problems and combining vision with

other domains like natural language processing (NLP) [11].

4.2. Future Research Trends

Looking forward, several key trends are poised to shape the future of computer vision:

- **Enhanced Model Design and Scalability:** Manually designing optimal neural networks is challenging. **Neural Architecture Search (NAS)**, which uses algorithms to automatically search for the best network structure, will continue to be a major area of research [30]. Concurrently, developing scalable models like EfficientNet and RegNet that can be easily adapted to different computational budgets without redesign is crucial for practical deployment [82, 65].

- **Model Visualization and Interpretability:** A major criticism of deep learning is its "black box" nature [90]. For high-stakes applications like medical diagnosis or autonomous driving, understanding *why* a model made a particular decision is essential. Research into visualization techniques and interpretability methods like Layer-Wise Relevance Propagation (LRP) [42] is critical for building trust and enabling debugging.

- **Cross-Domain and Multi-Modal Integration:** The most sophisticated AI systems will not operate in a vacuum. The future lies in combining vision with other modalities. For example, integrating vision with NLP can lead to advanced chatbots that understand visual context [1] or systems that can generate rich textual descriptions of images and videos. Combining vision with audio can power applications that diagnose medical conditions from both visual cues and speech patterns [62].

- **Expansion into Broader Application Domains:** While CV is established in many areas, its application is still expanding into new fields. We are seeing promising work in using deep learning for industrial forecasting (e.g., predicting petroleum production) [4], archaeological analysis [17], agriculture (e.g., crop monitoring), and environmental science (e.g., climate change modeling from satellite imagery).

4.3. Ethical Considerations and Societal Impact

The immense power of computer vision technology brings with it significant ethical responsibilities. The proliferation of surveillance systems raises profound questions about privacy and civil liberties. The potential for bias in AI systems, where models trained on non-representative data perpetuate and even amplify societal inequalities, is a major concern that requires careful attention to data collection and algorithmic fairness.

Perhaps the most prominent and pressing ethical challenge today is the rise of **deepfakes** [7]. The ability to create highly realistic, AI-generated videos has created a powerful tool for misinformation, propaganda, and malicious personal attacks. The technology is rapidly improving, making it increasingly difficult for humans to distinguish real from fake content, a concern highlighted by mainstream media [7]. This has sparked an arms race between generative and detection technologies. Researchers are actively developing methods to detect deepfakes [2, 3, 33], and industry leaders have launched initiatives like the Deepfake Detection Challenge to spur innovation [36]. However, the problem is not purely technical. It highlights the responsibility of the research community, as articulated by the creators of YOLO, who ceased their work on the project due to concerns about its potential misuse [69]. Addressing the threat of deepfakes will require a multi-faceted approach that combines technical solutions, media literacy education, and sensible public policy.

5. Conclusion

In just over a decade, deep learning has redefined the field of computer vision, turning long-held aspirations into tangible realities. This review has charted the remarkable journey of this transformation, starting from the foundational CNN architectures like AlexNet, VGGNet, GoogLeNet, and ResNet that established the deep learning paradigm. We have detailed how the principles from these core models were extended and adapted to create a powerful suite of techniques for tackling the most important tasks in computer vision: recognition, visual tracking, semantic segmentation, and image restoration.

The resulting real-world applications are already widespread and impactful, ranging from life-saving medical diagnostic tools and the perception systems of autonomous vehicles to the algorithms that enhance and moderate our digital content. However, as the field matures, it faces new and complex challenges. The drive for greater efficiency, the need for model interpretability, and the profound ethical questions raised by technologies like deepfakes are now at the forefront of the research agenda. The future of computer vision will be shaped not only by further technical innovation but also by our collective ability to guide its development in a responsible and beneficial direction. The progress continues at a blistering pace, promising to further erode the boundaries between machine and human perception in the years to come.

References

[1] Adamopoulou, E., & Moussiades, L. (2020). Chatbots:

History, technology, and applications. *Machine Learning with Applications, 2*, 100006.

[2] Agarwal, S., Farid, H., El-Gaaly, T., & Lim, S. N. (2020). Detecting deep-fake videos from appearance and behavior. *Proceedings of the 2020 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–6.

[3] Agarwal, S., Farid, H., Fried, O., & Agrawala, M. (2020). Detecting deep-fake videos from phoneme-viseme mismatches. *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2814–2822.

[4] Al-Shabandar, R., Jaddoa, A., Liatsis, P., & Hussain, A. J. (2021). A deep gated recurrent neural network for petroleum production forecasting. *Machine Learning with Applications, 3*, 100013.

[5] Altan, A., Karasu, S., & Zio, E. (2021). A new hybrid model for wind speed forecasting combining long short-term memory neural network, decomposition methods and grey wolf optimizer. *Applied Soft Computing, 100*, 106996.

[6] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data, 8*(1), 53.

[7] Bloomberg Quicktake (2018). *It's getting harder to spot a deep fake video*. Retrieved from https://www.youtube.com/watch?v=gLoI9hAX9dw

[8] Bochkovskiy, A., Wang, C. Y., & Liao, H.-Y. M. (2020). YOLOV4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

[9] Burger, H. C., Schuler, C. J., & Harmeling, S. (2012). Image denoising: Can plain neural networks compete with BM3D?. *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2392–2399.

[10] Cai, Z., Fan, Q., Feris, R. S., & Vasconcelos, N. (2016). A unified multiscale deep convolutional neural network for fast object detection. *Proceedings of the 14th European Conference on Computer Vision (ECCV)*, 354–370.

[11] Chai, J., & Li, A. (2019). Deep learning in natural language processing: A state-of-the-art survey. *Proceedings of the 2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, 1–6.

[12] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2015). Semantic image segmentation with deep convolutional nets and fully connected CRFs. *arXiv preprint arXiv:1412.7062*.

[13] Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder–decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the 15th European Conference on Computer Vision (ECCV)*, 801–818.

[14] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1251–1258.

[15] Dabov, K., Foi, A., Katkovnik, V., & Egiazarian, K. (2007). Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Transactions on Image Processing, 16*(8), 2080–2095.

[16] Dai, J., Li, Y., He, K., & Sun, J. (2016). R-FCN: Object detection via region-based fully convolutional networks. *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)*, 379–387.

[17] Dambrogio, J., Ghassaei, A., Smith, D. S., Jackson, H., Demaine, M. L., Davis, G., Mills, D., Ahrendt, R., Akkerman, N., van der Linden, D., & Demaine, E. D. (2021). Unlocking history through automated virtual unfolding of sealed documents imaged by X-ray microtomography. *Nature Communications, 12*(1), 1184.

[18] Dong, C., Loy, C. C., He, K., & Tang, X. (2016). Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(2), 295–307.

[19] Dosovitskiy, A., & Brox, T. (2016). Inverting visual representations with convolutional networks. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4829–4837.

[20] Elad, M., & Aharon, M. (2006). Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing, 15*(12), 3736–3745.

[21] Fu, C.-Y., Liu, W., Ranga, A., Tyagi, A., & Berg, A. C. (2017). DSSD: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*.

[22] Gando, G., Yamada, T., Sato, H., Oyama, S., & Kurihara, M. (2016). Fine-tuning deep convolutional neural networks for distinguishing illustrations from photographs. *Expert Systems with Applications, 66*, 295–301.

[23] Girshick, R. (2015). Fast r-cnn. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1440–1448.

[24] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, 580–587.

[25] Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing, 187*, 27–48.

[26] Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., & Xu, C. (2020). GhostNet: More features from cheap operations. *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1577–1586.

[27] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, 2961–2969.

[28] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(9), 1904–1916.

[29] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition resnet. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

[30] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). Searching for MobileNetV3. *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1314–1324.

[31] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

[32] Hsu, C.-C., Zhuang, Y.-X., & Lee, C.-Y. (2020). Deep fake image detection based on pairwise learning. *Applied Sciences, 10*(1), 370.

[33] Huang, J., Chai, J., & Cho, S. (2020). Deep learning in finance and banking: A literature review and classification. *Frontiers of Business Research in China, 14*, 1–24.

[34] Huang, Y., Cheng, Y., Bapna, A., Firat, O., Chen, D., Chen, M., Lee, H., Ngiam, J., Le, Q. V., & Wu, Y. (2019). Gpipe: Efficient training of giant neural networks using pipeline parallelism. *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS), 32*, 103-112.

[35] Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269.

[36] Kaggle (2019). *Deepfake detection challenge | kaggle*. Retrieved from https://www.kaggle.com/c/deepfake-detection-challenge

[37] Kim, S. Y., Oh, J., & Kim, M. (2020). Jsi-gan: Gan-based joint super-resolution and inverse tone-mapping with pixel-wise task-specific filters for uhd hdr video. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(07), 11287–11295.

[38] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS), 1*, 1097-1105.

[39] Krull, A., Buchholz, T.-O., & Jug, F. (2019). Noise2Void—LEarning denoising from single noisy images. *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2124–2132.

[40] Kumar, A., Walia, G. S., & Sharma, K. (2020). Recent trends in multicue based visual tracking: A review. *Expert Systems with Applications, 162*, 113711.

[41] Lai, W.-S., Huang, J. B., Ahuja, N., & Yang, M. H. (2017). Deep laplacian pyramid networks for fast and accurate super-resolution. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5835–5843.

[42] Lapuschkin, S., Binder, A., Montavon, G., Muller, K.-R., & Samek, W. (2016). The LRP toolbox for artificial neural networks. *Journal of Machine Learning Research, 17*(114), 1–5.

[43] Law, H., & Deng, J. (2018). CornerNet: Detecting objects as paired keypoints. *Proceedings of the 15th European Conference on Computer Vision (ECCV), 11218*, 734–750.

[44] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., & Wang, Z. (2017). Photo-realistic single image super-resolution using a generative adversarial network. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 105–114.

[45] Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., & Aila, T. (2018). Noise2Noise: Learning image restoration without clean data. *Proceedings of the 35th International Conference on Machine Learning (ICML), 80*, 2965-2974.

[46] Li, B., Yan, J., Wu, W., Zhu, Z., & Hu, X. (2018). High

performance visual tracking with siamese region proposal network. *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8971–8980.

[47] Lienhart, R., & Maydt, J. (2002). An extended set of haar-like features for rapid object detection. *Proceedings of the International Conference on Image Processing, 1*, I–900–I–903.

[48] Lin, M., Chen, Q., & Yan, S. (2014). Network in network. *Proceedings of the 2014 International Conference on Learning Representations (ICLR)*.

[49] Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 936–944.

[50] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, 2999–3007.

[51] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. *Proceedings of the European Conference on Computer Vision (ECCV), 9905*, 21–37.

[52] Long, X., Deng, K., Wang, G., Zhang, Y., Dang, Q., Gao, Y., Shen, H., Ren, J., Han, S., Ding, E., & Wen, S. (2020). PP-YOLO: An effective and efficient implementation of object detector. *arXiv preprint arXiv:2007.12099*.

[53] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431–3440.

[54] Lukezic, A., Matas, J., & Kristan, M. (2020). D3S – a discriminative single shot segmentation tracker. *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7131–7140.

[55] Mehrotra, R., Ansari, M. A., Agrawal, R., & Anand, R. S. (2020). A transfer learning approach for AI-based classification of brain tumors. *Machine Learning with Applications, 2*, 100003.

[56] Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-NEt: Fully convolutional neural networks for volumetric medical image segmentation. *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV)*, 565–571.

[57] Muzammel, M., Salam, H., Hoffmann, Y., Chetouani, M., & Othmani, A. (2020). AudVowelConsNet: A phoneme-level based deep CNN architecture for clinical depression diagnosis. *Machine Learning with Applications, 2*, 100005.

[58] Noh, H., Hong, S., & Han, B. (2015). Learning deconvolution network for semantic segmentation. *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, 1520–1528.

[59] Pinheiro, P. O., Lin, T.-Y., Collobert, R., & Dollàr, P. (2016). Learning to refine object segments. *ECCV 2016*, 75–91.

[60] Qi, Y., Zhang, S., Qin, L., Yao, H., Huang, Q., Lim, J., & Yang, M. H. (2016). Hedged deep tracking. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4303–4311.

[61] Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., & Dollár, P. (2020). Designing network design spaces. *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10425–10433.

[62] Rai, H. M., & Chatterjee, K. (2020). Detection of brain abnormality by a novel lu-net deep neural CNN model from MR images. *Machine Learning with Applications, 2*, 100004.

[63] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788.

[64] Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6517–6525.

[65] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

[66] Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(6), 1137-1149.

[67] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241.

[68] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4510–4520.

[69] Simonyan, K., & Zisserman, A. (2015). Very deep

convolutional networks for large-scale image recognition. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

[70] Song, Y., Ma, C., Wu, X., Gong, L., Bao, L., Zuo, W., Shen, C., Lau, R. W., & Yang, M. H. (2018). Vital: Visual tracking via adversarial learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8990–8999.

[71] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 4278–4284.

[72] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9.

[73] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2826.

[74] Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning (ICML), 97*, 6105-6114.

[75] Tian, Z., Shen, C., Chen, H., & He, T. (2019). Fcos: Fully convolutional one-stage object detection. *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9626–9635.

[76] Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2018). Deep image prior. *International Journal of Computer Vision, 128*(7), 1867–1888.

[77] Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*.

[78] Walia, G. S., & Kapoor, R. (2016). Recent advances on multicue object tracking: A survey. *The Artificial Intelligence Review, 46*(1), 1–39.

[79] Wang, L., Kim, T. K., & Yoon, K. J. (2020). Eventsr: From asynchronous events to image reconstruction, restoration, and super-resolution via end-to-end adversarial learning. *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8312–8322.

[80] Wang, N., Li, S., Gupta, A., & Yeung, D. Y. (2015). Transferring rich feature hierarchies for robust visual tracking. *arXiv preprint arXiv:1501.04587*.

[81] Wang, L., Liu, T., Wang, G., Chan, K. L., & Yang, Q. (2015). Video tracking using learned hierarchical features. *IEEE Transactions on Image Processing, 24*(4), 1424–1435.

[82] Wang, N., & Yeung, D. Y. (2013). Learning a deep compact image representation for visual Tracking. *Proceedings of the 27th Annual Conference on Neural Information Processing Systems, 1*, 809-817.

[83] Xie, J., Xu, L., & Chen, E. (2012). Image denoising and inpainting with deep neural networks. *Advances in Neural Information Processing Systems, 25*, 341–349.

[84] Xu, T., Feng, Z.-H., Wu, X.-J., & Kittler, J. (2019). Joint group feature selection and discriminative filter learning for robust visual object tracking. *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 7949–7959.

[85] Xu, S., Wang, J., Shou, W., Ngo, T., Sadick, A.-M., & Wang, X. (2020). Computer vision techniques in construction: A critical review. *Archives of Computational Methods in Engineering*.

[86] Yang, Z., Liu, S., Hu, H., Wang, L., & Lin, S. (2019). RepPoints: Point set representation for object detection. *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9656–9665.

[87] Ye, L., Liu, Z., & Wang, Y. (2020). Dual convolutional LSTM network for referring image segmentation. *IEEE Transactions on Multimedia, 22*(12), 3224–3235.

[88] Young, A. L., & Quan-Haase, A. (2013). Privacy protection strategies on facebook: The internet privacy paradox revisited. *Information, Communication & Society, 16*(4), 479–500.

[89] Yun, S., Choi, J., Yoo, Y., Yun, K., & Choi, J. Y. (2017). Action-decision networks for visual tracking with deep reinforcement learning. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1349–1358.

[90] Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *Proceedings of the 13th European Conference on Computer Vision (ECCV), 8689*, 818–833.

[91] Zhang, C., Lin, G., Liu, F., Yao, R., & Shen, C. (2019). Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5212–5221.

[92] Zhang, K., Liu, Q., Wu, Y., & Yang, M. H. (2016). Robust visual tracking via convolutional networks without training. *IEEE Transactions on Image Processing, 25*(4), 1779–1792.

[93] Zhao, Z., Jiao, L., Zhao, J., Gu, J., & Zhao, J. (2017). Discriminant deep belief network for high-resolution SAR image classification. *Pattern Recognition, 61*, 686–701.

[94] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2020). Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging, 39*(6), 1856–1867.

[95] Zhou, X., Wang, D., & Krähenbühl, P. (2019). Objects as points. *arXiv preprint arXiv:1904.07850*.