IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 February 2023, accepted 26 February 2023, date of publication 6 March 2023, date of current version 6 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3253206

RESEARCH ARTICLE

Pareto Optimized Large Mask Approach for Efficient and Background Humanoid Shape Removal

RYTIS MASKELIUNAS[1,2], (Member, IEEE), ROBERTAS DAMAŠEVIČIUS[2,3], (Member, IEEE), DAIVA VITKUTE-ADZGAUSKIENE[3], AND SANJAY MISRA[4]

[1]Center of Excellence Forest 4.0, Faculty of Informatics, Kaunas University of Technology, 50186 Kaunas, Lithuania
[2]Faculty of Applied Mathematics, Silesian University of Technology, 44-100 Gliwice, Poland
[3]Center of Excellence Forest 4.0, Department of Applied Informatics, Vytautas Magnus University, 44044 Kaunas, Lithuania
[4]Department of Computer Science and Communication, Ostfold University College, 3001 Halden, Norway

Corresponding author: Robertas Damaševičius (robertas.damasevicius@polsl.pl)

ABSTRACT The purpose of automated video object removal is to not only detect and remove the object of interest automatically, but also to utilize background context to inpaint the foreground area. Video inpainting requires to fill spatiotemporal gaps in a video with convincing material, necessitating both temporal and spatial consistency; the inpainted part must seamlessly integrate into the background in a variety of scenes, and it must maintain a consistent appearance in subsequent frames even if its surroundings change noticeably. We introduce deep learning-based methodology for removing unwanted human-like shapes in videos. The method uses Pareto-optimized Generative Adversarial Networks (GANs) technology, which is a novel contribution. The system automatically selects the Region of Interest (ROI) for each humanoid shape and uses a skeleton detection module to determine which humanoid shape to retain. The semantic masks of human like shapes are created using a semantic-aware occlusion-robust model that has four primary components: feature extraction, and local, global, and semantic branches. The global branch encodes occlusion-aware information to make the extracted features resistant to occlusion, while the local branch retrieves fine-grained local characteristics. A modified big mask inpainting approach is employed to eliminate a person from the image, leveraging Fast Fourier convolutions and utilizing polygonal chains and rectangles with unpredictable aspect ratios. The inpainter network takes the input image and the mask to create an output image excluding the background humanoid shapes. The generator uses an encoder-decoder structure with included skip connections to recover spatial information and dilated convolution and squeeze and excitation blocks to make the regions behind the humanoid shapes consistent with their surroundings. The discriminator avoids dissimilar structure at the patch scale, and the refiner network catches features around the boundaries of each background humanoid shape. The efficiency was assessed using the Structural Learned Perceptual Image Patch Similarity, Frechet Inception Distance, and Similarity Index Measure metrics and showed promising results in fully automated background person removal task. The method is evaluated on two video object segmentation datasets (DAVIS indicating respective values of 0.02, FID of 5.01 and SSIM of 0.79 and YouTube-VOS, resulting in 0.03, 6.22, 0.78 respectively) as well a database of 66 distinct video sequences of people behind a desk in an office environment (0.02, 4.01, and 0.78 respectively).

INDEX TERMS Semantic segmentation, occlusion-robust network, human shape extraction, background person removal, image inpainting.

I. INTRODUCTION

The worldwide intellectual community has discussed the prospects for continuous human expansion in recent years,

The associate editor coordinating the review of this manuscript and approving it for publication was Yu-Da Lin.

using the metaphors of the fourth industrial revolution and the next age of social interaction, produced by prominent technology revolution missionaries [1], provides a new social ontology in which modern technology plays a significant (if not the most significant) role [2]. Modern visual communication technologies introduce a radically new formulation of the question of relationship in the system of 'human-technology, ''metaverse,' and other virtual forms of interaction, ushering in an era of collaboration between advanced visualization technology, artificial intelligence systems, and people - all essential components of modern interactive systems [3]. Background removal is a function shared by virtual communication technologies such as Zoom, Teams, and the Big Blue Button. More advanced technology allowing selecting zones to be removed may now be found in applications like as Photoshop, Pxilemator, and others. Unfortunately, the approaches used are not always smart in their object choices and require human input to correct some peculiar areas. Removing unwanted objects from movies and live video feeds is even more critical in many applications, including film post-production and video editing [4], ranging from automatic detection of specific subjects [5], [6] to semantic pre-segmentation [7]. Background subject of interest removal is also very important in remote work communications to protect the privacy and confidentiality of individuals. When personal information is shared in a remote work setting, it can easily be intercepted by unauthorized parties, potentially leading to identity theft, financial fraud, or other forms of exploitation. By removing human subjects from communications, organizations can reduce the risk of sensitive information being disclosed and help maintain the privacy and security of their employees, customers, and partners. Schwab has also stressed the need for companies to adopt a proactive and responsible approach to data protection, including implementing appropriate security measures and being transparent about their data collection and usage practices [1].

Object removal implies excluding the given object from the image and then inpainting it with appropriate content. While manually eliminating objects from a video involves significant human work, automated video object removal has the potential to save a significant amount of time. The aim of automated video object exclusion is to inpaint the foreground region using background information and produce a video without the target item given the foreground object's location in each frame. Traditional video inpainting frequently seeks to fill spatiotemporal gaps in a video with convincing material, which necessitates both temporal and spatial consistency; the inpainted part must blend seamlessly into the background in a variety of scenes, and it must maintain a consistent look in subsequent frames while its neighborhood may change notable. Despite significant advances in deep learning models for image inpainting, extending these methodologies to the video domain remains challenging because of the added dimension of time, which necessitates image-based encoder-decoder models to gather and modify data from adjacent frames and generate the unknown areas [8].

Graphical object removal can be treated as a subtask of image inpainting. The earliest successful examples used patch inpainting algorithms [9] that divided images into small patches and recovered the masked region by inserting the most comparable patch anywhere in the frame. These approaches may produce genuine results, but these are typically time-consuming due to the intricacy of neighbor-finding algorithms [10]. Furthermore, patch-based approaches presume the missing section has a reference and frequently fail to restore non-repetitive and complicated regions in real-world [11]. Deep learning (DL) is clearly a tool of choice in computer vision, and image inpainting is included [12]. Based on the training data, such models may predict the missing pieces and produce innovative outcomes of high quality. Unfortunately, when applied to movies, most DL image-based algorithms provide temporally erratic results, resulting in flickering or dismorphed images, because they ignore the temporal relationship between frames and process it individually [13], requiring dedicated methods such as foreground and spatial awareness to combat this issue.

We present a novel deep learning based methods for removing of unwanted objects in videos. The method is employs Generative Adversarial Networks (GAN) technology. The main novelty and contribution of this paper is the use of Pareto-optimized GANs for the task of inpainting, which has not been done before, and a novel approach for creating semantic masks in images, which is robust to occlusion and generalizes well to higher resolutions, while being efficient in terms of computational complexity and required parameters.

Paper is structured as follows. Section II presents and discusses the state-of-the-art in inpainting. Section III explains materials and methods. Section IV details the experiments and their results. Finally, Section V presents a discussion of the findings and concludes.

II. STATE OF THE ART REVIEW IN INPAINTING

This section is aimed at presenting the reader with some of the most common existing image inpainting algorithms, which have been divided into two and a half common categories: sequential, generic image processing oriented deep learning, and their subset called adversarial networks, separating object removal into its own subsection. Furthermore, for every group, a summary of techniques for various forms of visual distortion is offered. An objective comparison is offered in the discussion part of the paper.

A. SEQUENTIAL APPROACHES

Patch-based solutions rely on methods that refill in the required (cut object) section patch by patch, by looking for well-matching potential replacement patches surrounding the part of the image we wish to trim and then copying them to corresponding spots. The problem of image inpainting may be approached from the standpoint of consecutive partial signal recovery by assuming that each image patch permits a sparse expression over a redundant vocabulary [14]. Li's

approach employs the super-wavelet transformation to predict the multi-direction features of an affected image, which is then combined with color data to calculate the weighted color-direction distance of two patches [15]. Traditional inpainting approaches based on low-rank priors often require an iterative singular value shrinking procedure to solve a convex optimization problem in order to restore the distorted pixels [16]. This can be solved with low rank approximation, which saves time on repeated shrinking [10]. Xu et al. developed patch-level sparsity ideas for modeling patch representation and priorrity, i.e., two critical phases for patch propagation in the example-oriented inpainting technique [9]. Others suggested a use numerous pyramids [17], [18], localized patch stats, and geometrical feature-based sparse representation to eliminate new items in images while keeping texture integrity and structural consistency [19]. Study [20] highlighted that post-processing reduces the resemblance of block pairings while also disrupting the correlations between neighboring pixels to a certain degree. Zhang et al. suggested an approach based on prior data of surface fitting areas and angle-aware patch comparison [21]. The top-down splitting approach proposed by Ruzit was able to separate the image into variable - sized blocks based on their context, limiting the search for potential patches to non-local image areas with fitting frame of reference [22]. Li and Wozniak [23], based on bilateral filtering, proposed a hole in-filling and optimization approach for remote sensing pictures. The compensation function of similarity determination and the Thiele continuous fraction approximation exponential function are used to augment the standard bilateral filtering process. The picture is binarized to construct a hole mask after the histogram sets the threshold. The constraint term is incorporated to the modified bilateral filtering method for further improvement. The remote sensing image's hole region is filled based on the features that pixels in a given range have the same gray value.

More recent techniques are often hybrids of the other two kinds discussed in this section. Yang et al. propose a multi-dimensional neural patch synthesis approach based on collaborative enhancement of image content and texture constraints that not only keeps context - specific structures but also generates higher details by matching and adapting patches to the DL's most comparable mid-layer feature correlations [24], or exploring the semantic relevance [25], while Zhu suggests using a convolutional neural network (CNN) to detect patch-based inpainting operation [26]. Meanhile, another hybrid example is the PGGAN method which suggests using a discriminator network that blends a global GAN model with a patchGAN technique [27] with a variation using spectral-normalized discriminator on dense frame patches [28].

B. GENERAL INPAINTING ORIENTED DEEP LEARNING ARCHITECTURES

Deep convolutional networks show great promise in multiple computer vision problems, including image inpainting [29]. CNNs in particular are the most common architecture used for this task, used to enhance predicted results, often utilizing huge training data [30]. CNN-based image inpainting has several drawbacks [31]. For starters, convolution improves the repair network's performance on rectangular hole image repair assignments, leading the network to look over-fitting and unsuitable for general scenarios. Also, convolution increases the reliance of the image restoration result on the original value of the hole region, achieving bad repair outcomes such as artifacts. To compensate for these repair flaws, costly post-processing is required. Alternatively, Zheng et al., suggested a deep multi-resolution mutual learning approach capable of fully exploring data from multiple resolutions [32]. Liu et al. [33] suggeste using partial convolutions, in which the convolution is masked and renormalized so that it is only conditioned on valid pixels. Wang et al. created a concurrent multi-resolution inpainting network using cross partial convolution, whereas low-resolution sections focus on overall structure and high-resolution sections concentrate on local texture details, in contrast to the standard image inpainting architecture [34]. Xu et al. presented an universal deep learning algorithm for high-resolution image inpainting that produces a semantically consistent blurred outcome using low-resolution inpainting while suppressing computing cost [35]. Given that multi-task image recognition network CNNs cannot completely utilize all image scale features extracted during the feature extraction process, the authors of [36] recommend using Feature Pyramid Networks and back connections to acquire more features.

Baupame et al. [37] introduced a U-NET-based technique for portraying metro traffic by creating an image that displays the geographical data of trains traveling on a metro line while accounting for the sporadic temporal sampling of train loads for improved precision and reliability. Lee et al. [38] presented the Copy-and-Paste Networks architecture for video inpainting, which takes advantage of extra data in other frames of the video. The network was taught to copy and paste relevant data from reference frames to fill gaps in the target frame. Li et al. suggested recurrent feature reasoning, which is similar to how people solve puzzles, initially aiming at easier locations, then more difficult ones [39]. This method frequently insinuates the unit borders of the convolutional feature maps and afterwards utilizes these as cues for any further inferences. To combat same issue, Zhou proposed a multi-homography transformed fusion approach that refers to a different source image having the scene with the target image [40]. The model aligns the source and target images by predicting several homographies guided by various depth levels.

Some of these methods result in deformed architecture or hazy, uneven texturing. The issue stems from the encoder layers' inability to construct a comprehensive and accurate embedding of the missing areas from start. Suin et al. [41] suggested a distillation-based strategy for inpainting training in which they gave direct feature-level supervision. They

used cross and self-distillation methods, as well as a specialized completion-block in the encoder, to generate more precise hole encoding. Splitting inpainting task into two steps, namely structure reconstruction and texture generation, is another common approach [42]. As a viariant, Guo et al. [43] suggested a two-stream architecture for image inpainting that couples structure-constrained image synthesis with image-guided structure restoration to better utilize each other for more believable production. Liu employed CNN characteristics from the encoder's shallow and deep layers to describe the structures and texture of an input image, correspondingly. The shallow layer characteristics were routed to a texture side, whereas the deep layer characteristics were routed to a structure side [44]. Although modern inpainting algorithms using deep neural networks have shown great progress, they still suffer from other artifacts, such as harsh shapes and abrupt colors when filling in the empty regions. Wang et al. [45] presented an external-internal inpainting approach with a monochromic bottleneck to enable image inpainting models eliminate these artifacts to solve these concerns. The authors of [46] presented a dynamic selection network to address image inpainting challenges in which stochastically corrupted patches in the input images likely to mislead the inpainting procedure and provide illogical content. Reference [47] created the Mask-Aware Dynamic Filtering module to successfully train multi-scale features for missing parts during the encoding process. Filters for each convolution window were created using characteristics from the mask's corresponding area.

1) GENERATIVE ADVERSARIAL NETWORKS

Generative adversarial networks has strong data generating capabilities, making it ideal for image inpainting [48]. The Context Encoder network [49] was the first to integrate CNN and GAN for image inpainting, while a later variant featured pyramid style attention transfer to further enhance the effectiveness [50]. The authors of [51] offered a method for image inpainting that improves the reproduction of filled regions with fine features by employing a two-stage adversarial model called EdgeConnect [52], which consists of an edge generator accompanied by an image completion network.

Cai et al. [53] proposed devising a consistency loss to lead the produced image toward an approximate representation of the ground truth. After several rounds, their generator learned the mapping of styles to numerous sets of vectors. The suggested model might produce a vast variety of solutions that are consistent with the image's context semantics. The authors of [54] advocated dividing the hole filling process into various phases, with each phase aiming to complete a course of the full curriculum. The use of a two-discrimination network was proposed in the [55] paper. The suggested technique creates a fusion network by combining an image inpainting, global discrimination, and local discrimination networks to apply computational images. The suggested algorithm's training technique employs a comparable patch method to fill the damaged region in the image and use them for input training objects, which dramatically enhances the speed and reliability of image inpainting. Following that, a Long short-term memory (LSTM) framework was utilized to connect all of the stages. By using this learning mechanism, their approach was able to employ a progressive GAN to gradually decrease huge corrupted regions in natural images, yielding encouraging inpainting outcomes.

Lui proposed changing the GAN during image formation to adjust deep characteristics of input noisy data from coarse-to-fine by inserting an originally recovered image and the hole areas at numerous scales [56]. Wu et al. proposed a coarse-to-fine generative model for reducing artifacts by merging a local binary pattern learning architecture within an actual inpainting architecture [57]. Unsupervised three-part Cross-space Translation Generative Adversarial Network was introduced by Zhao et al. [58]. The manifold projection and generation modules were merged to learn unsupervised one-to-one image mapping among two spaces by trying to project instance image space and contingent completion image space into a prevalent low-dimensional manifold space, which might significantly improve the variety of the reconditioned samples. Yang et al. [59] proposed training a shared generator to generate the corrupted image and accompanying structures — edge and gradient — at the same time, implicitly encouraging the generator to use pertinent structural information during inpainting. Zeng tackled the problem of significant computational overhead by suggesting to train a patch-borrowing procedure to an attention-free generator through joint training of a supplementary contextual restoration task, which incentivizes generated outcome to be feasible even when rebuilt by neighboring regions [60].

C. OBJECT REMOVAL

Object removal is the process of eliminating a specific object from an image and replacing it with appropriate content [61], and applying numerous intelligent methods to maintain the realistic recreation of the surrounding image areas [62]. The segmentation technique may be used to select the region to be inpainted first [63]. Le et al. [64] proposed a semi-automated (manual object marking) approach in which segmentation masks were improved initially and then automatically transmitted over the video. Using video inpainting methods, missing sections were then reconstituted. Tuperware et al. [65] suggested a video inpainting approach employing region segmentation employing the robust exemplar-based inpainting procedure. To overcome the dropping effect, they employed a strong priority function and region segmentation to establish the adaptable patch size and search region. Others suggested using spatial contextual dependencies [63]. In comparison to de-fencing, the target point of object removal is fixed and focused, and the background info of the image is not affected across a wide range, resulting in a superior inpainted outcome. Object removal is often split into three steps: recognition, elimination, and inpainting [66]. Some deep

learning algorithms for image inpainting employ a regular convolutional network to regenerate the holes left by the removal of the item. However, the outcome of this process is subpar, since the resulting image is frequently damaged and unclear [67]. Multiple types of algorithms were used to solve this problem: texture synthesis methods were used for generating large image areas from sample textures or image recognition methods were used to find the neccessary object in the image, and ''inpainting'' approaches were used for filling in small image gaps [68]. Another difficulty is dealing with photos of busy settings containing objects with complicated details. The authors of [69] proposed using an image's depth map to discover the order of items in the target image and a collection of multi-views of objects to inpaint the gaps, while others suggested the spatial consistency of aligned frames by employing a region-based homography computing approach [70].

The assessment of the surroundings by the algorithms frequently hinders the execution of critical tasks including mapping and localization. Researchers partially solved this issue in study [71] by generating convincing texture, color and geometry in image parts hidden by dynamic objects. Authors suggested a geometry-aware architecture with a coarse-to-fine topology with gated reccurrent feedback method for adaptively fusing data from prior timesteps. Criminisi et al. [72] presented a best-first approach in which confidence in synthesized pixel values is conveyed similarly to information propagation in inpainting. Color values were calculated via exemplar-based synthesis. Pyo et al. [73] suggested utilizing GANs to remove the desired item from an image. They designed the network that combines two GANs: The first GAN removes the target item from the input frame, while the second GAN creates a frame with the backdrop filling up the empty area. The network may delete the desired item from the input frame and receive a frame with the erased region refilled with the background while avoiding employing any object detection approach.

Conditional random fields as RNNs can be employed to further segment the target in semantic sense, without the need for masking or artificial pre-processing. The representation characteristics may then be derived from the CNN feature maps of the missing region's neighbouring regions. The missing area can then be synthesized using large-scale bound-constrained optimization based on the CNN encoding properties of similar nearby regions [74]. Using such semantic segmentation to recognize different groups of objects, Paper [75] developed an image-based object elimination approach for automated object removal and inpainting with generative adversarial networks, automating the city-scape object detection and processing, while paper [76] applied a similar way to indoor-scape scenarios and paper [77] for autonomous driving data preparation. Li et al. [78] proposed a more universal approach, using three trainable GAN modules, including flow completion, feature propagation, and content hallucination, which gave improved outcomes in object removal accuracy.

Popular in object removal exemplar-based image inpainting strategy has two main components: estimating the fill priorities of patches in the vacant region and selecting the best matched patch. Wang et al. demonstrated a robust exemplar-based strategy that used a regularized variable to change the patch prioritization parameter [79]. Neleema et al. suggested a best-first approach in which confidence in synthesized pixel values is conveyed similarly to information diffusion in inpainting [80], with color values calculated via exemplar-based synthesis. Quite a few such exemplar-based image inpainting approaches have the following drawbacks: searching for abnormally similar patches is time-intense and imprecise, there is a high false detection rate, and there is a lack of resilience to many post-processing combination procedures [81]. In light of the aforementioned drawbacks, a detection approach for visual object removal based on LSTM-CNN hybrid was developed [82].

Pinjankar et al. proposed combining the examplar strategy with region segmentation to minimize the drop effect and to estimate the adaptable patch size and decreased search area [83]. Even though the exemplar-based classical object removal technique offers advantages such as the capacity to keep image texture and structure characteristics as well as image clarity, it fails to retain the graphical fidelity of inpainted images [84]. To address these shortcomings, the creators of [85] developed a two-stage approach. The dual-tree complex wavelet transform is used in the first stage to get sub-bands of wavelets from a low-resolution image. The image is then improved using the super-resolution approach in the second stage. Bandei developed a robust image refitting algorithm for both synthesized and genuine unshaded textured images, in which the local features from the generated patch is utilized to recreate the target area in the image pixel-by-pixel, thus promulgating structural and textural data concurrently [86].

Bonde et al. [87] set out to create a robust approach for image inpainting that could mend minute fissures as well as broad areas, such as those created by object removal. The image to be recovered is split into sub-bands using DWT, improving on [88] suggested use of regular Wavelet transform. In [89], a technique for calculating the structural sparsity of the targeted area and subsequently identifying the local features of the area where the targeted area is situated was proposed. Then, based on different geographical features, defined different search areas for the targeted areas. Next, in the search area, located the example area and restored the targeted area. Sum of Squared Differences is also a popular method for determining the level of resemblance between the exemplar patch and the target patch, which has a significant influence on restoration outcomes. Although the matching rule is straightforward, it is likely to result in a mismatch error. The authors of [90] presented a difference degree constraint-based inpainting approach for object

removal called Mean of Squared Differences. They utilized it to calculate the degree of separation between matching pixels in known places in the target patch and the exemplar patch. In addition, the authors of [91] proposed an adaptable two-round search approach. They employed the Differences Between Patches between both the target patch and the exemplar patch, and then utilized it to calculate the degree of distinction in the two patches.

III. MATERIALS AND METHODS

A. DATASETS AND IMPLEMENTATION

To train and assess our technique, we employed two prominent video object segmentation datasets: DAVIS [92], [93], [94] and YouTube-VOS [95], [96], [97]. We have used only the sequences related to humanoid masks in different contexts and environments. To replicated frequent vlogging scenarios in which only half of the body is visible, we utilized our own database of persons behind the desk in various positions, which included 66 distinct collected video sequence instances comprising 133 minutes of recording [98] in office environment. The proportion for training, validation and testing was 70/15/15. Pytorch v1.12.0 was used to build the project, which was trained on the above dataset in 320 × 180 frames per video. The warping temporal distance was set to 1, 3, and 5. The loss was handled by the relu layer. The patch was 12 × 12 in size. The loss weights were set at 0.01. A machine running Linux Mint, with AMD Ryzen 7 5700G CPU with 64 Gb of RAM and Nvidia Geforce 2060RTX Super GPU was used for training. The overall process of training took 122 hours.

B. ALGORITHM

Our approach is illustrated in Figure 1. Given an image with a multiple humans as input, the system automatically approximately assigns a Region of interest (ROI) block for each shape, sketches the contour of each humanoid shape region, and an internal binary mask is constructed for that region. A skeleton detection module is utilized to determine which humanoid shape is the main object to be retained. This is based on the analysis of quantity (visible number of body parts) and relevant size (distance) of the 32 monitored bones. This method also allows for the retention of two or more major characters, common in vloggins scenarios when a group of people is conversing behind a table, with a camera distance comparable to each individual. The inpainter network uses the fusion of the input image and the mask, to create a coarse output image free of bacground humanoid shapes. The generator, is an encoder-decoder architecture with skip connections to enable the recovery of spatial information lost due to network contracting and expansion by merging local and global information while upsampling. Encoder feature maps are supplied to the decoder through skip connections and concatenated with the matching decoder feature maps. GAN is configured using 4 layers of dilated convolution, and squeeze and excitation blocks between the encoder and decoder. Dilated convolution is used to record a broad field of vision with fewer parameters, making the regions behind each humanoid shape consistent with its surrounds. Squeeze and excitation improves a network's representational capability by learning channel weights based on their significance and re-calibrating feature maps. Except for the first and last layers, each layer is made up of ReLu, convolution, and instance normalization. Leaky ReLu was used as the activation function in the last layer as we empirically determined it to show the best performance. The decoder network gradually scales up the characteristics to image scale, using transpose convolution. The discriminator penalizes dissimilar structure at the patch scale. The refiner network design is similar to the inpainter network in that input is utilized at the refiner input as well as the inpainter outcomes. The refiner network catches features around the boundaries of each background humanoid area, which is a benefit. We receive the image without the background humanoid figures, but with an adequate amount of clarity in the excised region. A feature level reconstruction mistake is managed by a pre-trained loss network to improve quality.

1) ROI LOCALIZATION

The ROI (region of interest) is localized autonomously; the first stage in this procedure is to remove the background image. Non localized Gaussian Mixture Model was adapted from [99] to extract the background section of each frame of the video series (using previous frames) and computes the average value of each of these backgrounds images as the resultant background model. Inpainted video frames are utilized for the computation since they simplify the procedure by eliminating unwanted background people. The patch error function was selected to be Gaussian, that ensures that the weight of pixel defects nearer to the centre of the block is greater, while the weight of pixel defects nearer towards both sides of the block is less. Once the average background image for the image is retrieved, it is converted to grayscale and Fuzzy edge detection is applied [100] as it allows for a more dynamic content handling than a flood fill algorithm used by original authors of the approach. This enables values to be obtained from the x and y directions, resulting in finer edge detection by searching for similar or identical values as in the treated area center. As a result, all pixels till the edges are joined. Figure 2 shows the detected ROI. When ROI is identified, it is used to filter detections outside and to some extent) The generated bounding box is an axis-aligned rectangle containing all pixels discovered by our approach.

2) MASKING AND SORTING OF HUMANOID SHAPES

Blazepose [101] model was used to automatically designate a primary person (most frontal from camera perspective) to keep after semantic preprocessing by using the best accurate skeleton model and greatest bone area. We have manually adjusted the mediapipe to apply a two-step detector-tracker technique in which the tracker runs on a cropped region-

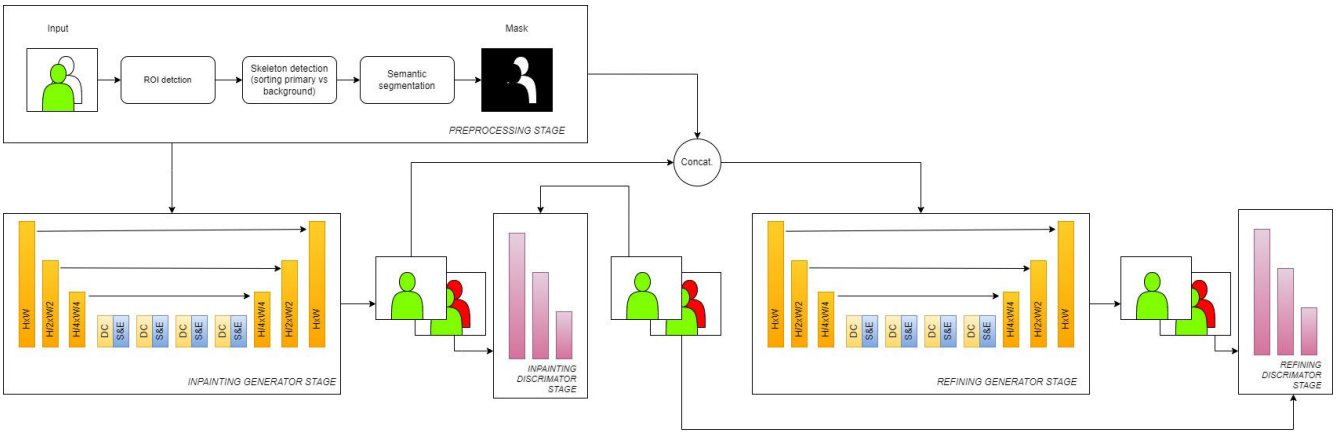


FIGURE 1. Architecture of DAIWA network. DC stands for Dilated Convolution. S - Squeeze. E - Excite.

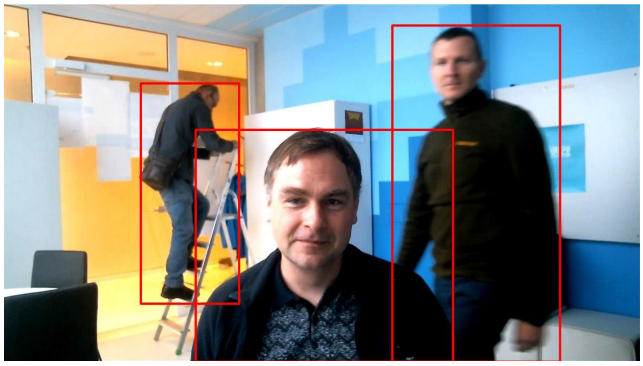


FIGURE 2. Detected ROI of the video frame.

FIGURE 3. Adjusted masks of the video frame.

of-interest including the person inside the original image suggested in the improved version called BlazePose GHUM Holistic [102] as the authors did not yet shared the source code for it. A bounding box, humanoid ID, detection probability are all part of each mask movement tracking operation. Because tracking may include erroneous class values, a specific class value must be specified for the entire processing stage. When the item was outside the ROI, the time should be any hundred or so miliseconds. It is essential to compute an array of time values in miliseconds and round down to the closest integer. Such time values are aggregated into distinct groups that correlate to correct seconds values. The location of the group with the greatest cumulative value is then resultant value in tracking.

The semantic masks for all human-like forms in the image were constructed using a method based by [103], which comprises of four basic components: a feature extraction block, and three branches: a local, a global, and a semantic one. The approach extracts global features for the global branch in order to encode occlusion-aware local data, resulting in occlusion-resistant extracted features. The fine-grained local properties of the local branch are obtained. To identify non-occluded parts, the semantic mask is built for the semantic branch. This implementation adapted to our purposes was able to include even ocluded human shapes as seen in Figure 3.

To eliminate a person, we employed a modified big mask inpainting approach [104] that leverages Fast Fourier convolutions (FFC) and generalizes well to resolutions greater than those visible at train time while requiring fewer parameters and taking less time. The use of dilation also improves the identified human masks. It helps to construct some higher exterior borders, and less ghosting artifacts appear in the frame. This approach employs samples from polygonal chains with a high random width (broad masks) and rectangles with unpredictable aspect ratios (box masks).

FFC employs a channel-wise FFT and has a receptive field that includes the whole picture. FFC separates channels into two parallel branches: one that uses classic convolutions and one that uses real FFT to account for the global environment. Real FFT can only be used with real valued signals, and inverse real FFT produces the real-valued output. In comparison to FFT, real FFT employs just half of the spectrum. FFC specifically performs the following actions:

- applies Real FFT2d on a tensor as an input

$$RealFFT2d : \mathbb{R}^{H\times W\times C} \rightarrow \mathbb{C}^{H\times \frac{W}{2}\times C} \quad (1)$$

- and combines actual and fictitious elements

$$ComplexToReal : \mathbb{C}^{H\times\frac{W}{2}\times C} \rightarrow \mathbb{R}^{H\times\frac{W}{2}\times C} \quad (2)$$

- applies a frequency-domain convolution block

$$ReLU \circ BN \circ Conv1 \times 1 : \mathbb{R}^{H\times\frac{W}{2}\times 2C} \rightarrow \mathbb{R}^{H\times\frac{W}{2}\times 2C} \quad (3)$$

- recovers a spatial structure using the inverse transform.

$$RealToComplex : \mathbb{R}^{H\times\frac{W}{2}\times 2C} \rightarrow \mathbb{C}^{H\times\frac{W}{2}\times C},$$
$$InverseRealFFT2d : \mathbb{C}^{H\times\frac{W}{2}\times C} \rightarrow \mathbb{R}^{H\times W\times C}. \quad (4)$$

3) HANDLING RECONSTRUCTION ERRORS USING PRETRAINED LOSS

Naive supervised losses need the generator correctly reconstructing the ground truth. But the viewable regions of the image frequently do not include enough information to recreate the masked area precisely. As a result of the average of many probable modes of the inpainted material, utilizing naïve supervision produces hazy results. A gap between characteristics retrieved from anticipated and target images by a base pre-trained model is measured as perceptual loss combined with Structural Similarity Index (SSIM) [105]. Adversarial loss can then be used to make that the inpainting model produces natural-looking local features. We may define a discriminator that distinguishes between ''genuine'' and ''fake'' patches at the local patch level. Only patches that connect with the veiled region are labeled ''fake.'' Because of the guided perceptual loss, the generator quickly learns to reproduce the known elements of the input image, hence the known parts of produced images are labeled as ''real.'' Finally, the quasi adversarial loss is applied.

The adversarial loss is paired with the difference between the real and synthetic pictures. The generator is then updated with the help of a

$$G^* = arg \min_G \max_D \lambda_{GAN}\mathcal{L}_{cGAN}(G, D) + \lambda_d \mathcal{L}_{L2}(G), \quad (5)$$

where

$$\mathcal{L}_{GAN}(G, D) = \mathbb{E}_{x\sim p_x(x)}\left[\log D(x, u)\right] + \mathbb{E}_{z\sim p_z(z)}\left[\log(1 - D(G(z,\ u))\right]. \quad (6)$$

and

$$\mathcal{L}_{L2} = \mathbb{E}_{x\sim p_x(x)}\left[(x - G(u))^2\right] \quad (7)$$

The final loss will be a weighted accumulation of all these losses, like in [104], yet in our approach we use dynamic weights, shifting in response to observed data and global image structure integrity.

On each sampled RoI during training, we define a multi-task loss $loss_{multi}$ as follows:

$$loss_{multi} = \alpha loss_{cls} + \beta loss_{box} + \gamma loss_{mask} + \eta loss_{part} \quad (8)$$

where α, β, γ and η are predetermined positive constants.

The definitions of the classification loss $loss_{cls}$, bounding-box loss $loss_{box}$, and mask loss $loss_{mask}$ are found in [106]. The part branch determined the binary cross-entropy loss $loss_{part}$ for each part box and then determined the loss of this RoI as the average of the part losses.

The following is the calculation's mathematical formula:

$$loss_{np} = -(y_n * \log(\delta(z_n)) + (1 - y_n) * \log(1 - \delta(z_n))) \quad (9)$$

$$loss_n = \frac{1}{10}\sum_{p=1}^{N}(l_{np}) \quad (10)$$

$$loss_{part}(z, y) = \frac{1}{N}\sum_{i=0}^{N-1}(l_i) \quad (11)$$

here z_n is the likelihood that the target object will be found in n samples, where y_n are sample labels and δ is the sigmoid function. Calculate the cross-entropy loss of each instance's parts on average, and then use the mean of all the cases in an image to determine the part loss value.

Our final loss includes GAN loss, multi-part loss, and $mathcalR_1 = E_x \parallel \nabla D_\xi(x) \parallel^2$ as:

$$loss_{final} = \kappa loss_{multi} + \nu loss_{adv} + \rho\mathcal{R}_1 \quad (12)$$

where κ, ν, ρ control the weight of each part.

IV. EXPERIMENTAL EVALUATION AND RESULTS

A. EVALUATION METRICS

We used the Learned Perceptual Image Patch Similarity (LPIPS) [107], [108], Frechet Inception Distance (FID) [109], [110] and Structural Similarity Index Measure (SSIM) [111] measurements, as these are standard procedure metrics to produce a valid comparison.

The LPIPS is computed done as:

$$d(x, \hat{x}) = \sum_l \frac{1}{H_l W_l}\sum_{h,w} \parallel w_l \odot \left(\hat{y}^l_{hw} - \hat{y}^l_{0hw}\right) \parallel_2^2 \quad (13)$$

where x denotes the true image, $\hat{x}$ denotes the synthesized image, and d denotes the distance between two images.

LPIPS is calculated as: x and $\hat{x}$ are fed into the VGG network to extract features, which activates the output of each layer and normalizes as $\hat{y}^l, \hat{y}^l_0 \in \mathbb{R}^{H_l\times W_l\times C_l}$.

Following the weight of the w layer, the L_2 distance is calculated. Finally, we compute the distance by taking the average.

The FID shows how similar two image datasets are. It can be used to assess the quality of samples from GANs. Lower scores are associated with higher-quality images, and smaller values indicate that the two sets of images have more in common. The FID is calculated by finding the Fréchet distance between two Gaussians fitted to the Inception network's feature representation. The following is the FID calculating expression:

$$FID(x, \hat{x}) = \parallel \mu_x - \mu_{\hat{x}} \parallel_2^2 + \mathrm{Tr}\left(A_x + A_{\hat{x}} - 2(A_x A_{\hat{x}})^{\frac{1}{2}}\right) \quad (14)$$

where Tr is the sum of the main diagonal elements (the matrix trace), μ is the mean, A is the matrix of covariance, x is real image, and x is synthetic image.

The SSIM index can compare the similarity of two images. SSIM considers image degradation to be a perceptual change in structural information, whereas taking into account some perceptual operations like luminance masking and contrast masking. SSIM is an appropriate verification metric for image completion. The following is the SSIM calculation:

$$\text{SSIM}\left(x, \hat{x}\right) = \frac{\left(2\mu_x\mu_{\hat{x}} + c_1\right)\left(2\sigma_{x\hat{x}} + c_2\right)}{\left(\mu_x^2 + \mu_{\hat{x}}^2 + c_1\right)\left(\sigma_x^2 + \sigma_{\hat{x}}^2 + c_2\right)} \tag{15}$$

where μx and $\mu\hat{x}$ are the mean of images x and $\hat{x}$, σx and $\sigma\hat{x}$ are the covariance of images x and $\hat{x}$, and $\sigma x\hat{x}$ is the covariance of images x and $\hat{x}$.

When many natural completion percentage are possible, LPIPS and FID are more appropriate for testing performance of big masks inpainting than pixel-level L1 and L2 distances. Lower values of LPIPS and FID indicate a better result, and on the contrary higher values of SSIM indicate a better result.

B. EXPERIMENTAL SETTINGS

Because of the nature and goal of our methodology, only photos with humans were examined on the DAVIS and Youtube-VOS datasets. Images of persons seated behind the table were included in our proprietary dataset. LaMa [104], MADF [47], Regionwise [112], and AOTGAN [113] were chosen for comparison since they were based on a large mask inpainting method, suitable for human-like forms and thus were similar in this regard to our approach (Figure 4). We have also included a state of the art method E2FGVI [78], operating on semantic masks, as it claims the best accuracy of most recent methods. Each of these methods was reproduced using the code available on their respective github repositories. Every one of these methods also required to manually select a backdrop humanoid form to be deleted, however our method was designed to pick the background humanoid shapes automatically, yet retaining a highly respectful outcome in LPIPS, FID and SSIM.

C. RESULTS WITH ORIGINAL IMAGES

PyTorch was used to build the metrics pipeline. We test performance with narrow, wide, and segmentation-based masks. Results are displayed in Tables 1, 2, 3.

Our model Daiwa performed well while doing a fully automated background person removal, yet the other models were not far behind in manual mask selection (automatic was not supported on either), with semantic masks producing the best result in almost all cases. Interestingly, we were unable to duplicate the high values shown in the E2FGVI [78] study, presumably due to the fact that we only employed a sample of persons from the DAVIS and Youtube-VOS datasets.

D. ROBUSTNESS EVALUATION

Real-world circumstances provide measurements and estimations that are far from perfect and frequently contain noise. In this part, we explore the impacts of noisy data inputs, such as semantic mask and depth, and report on how well an actual system performed when used to make inferences. In this part, we give an assessment of the robustness of our Daiwa approach. This is crucial in real-world inpainting settings since several post-processing processes, such as noise addition, scaling, and/or compression, may have an influence on inpainting accuracy. To that purpose, we used various post-processing techniques of varying kinds and magnitudes, and the results are shown in Figure 5.

For robustness evaluation, we follow the methodology outlined in [71]. In order to deform forms for depth noise, we first distort all semantic classes of pixels rather than just semantic masks. Furthermore, we employ the Sobel filter in the x-direction and set any pixels in the initial depth map that are greater than the Sobel threshold to 0. In our instance, the Sobel thresholds was set to 5. Then, we simulate pixel-level noise using the Kinetic depth noise model, ensuring that the offset does not exceed 5 meters and multiplying the standard deviation of the noise by the value of noise parameter $p_n \in [0, 1]$ that we use to alter the quantity of noise in order to compare the impacts of noise for each of the inputs. Each blob in a dynamic object binary mask is approximated with 20% of all the pixels in the relevant contour in order to mimic the semantic segmentation noise. We then shift $(\mathbf{p}_{ij} - \mathbf{c}_i)/\|\mathbf{p}_{ij} - \mathbf{c}_i\|_2 \cdot \varepsilon_i$ for each pixel $\mathbf{p}_{ij} = \begin{bmatrix} u_{ij} & v_{ij} \end{bmatrix}$ from the i-th contour, where c_i is the center pixel of the i-th blob and $\varepsilon_i \sim \mathcal{N}(0, (p_n\sigma_i)^2)$. The radius of the i-th contour, $r_i = \max_{j,k} \|\mathbf{p}_{ij} - \mathbf{p}_{ik}\|_2/2$, is used in this case as $\sigma_i = r_i/5$. Finally, with probability of p_n, all of the pixels with the highest depth value are likewise set to 0. The results of robustness evaluation are shown in Figure 5.

While the intensity of the perturbations is relatively low, the overall performance is good; for example, when executing H.264 compression with a quality factor of less than 1 Mbps (vs normal 3 Mbps), the performance is essentially unaffected. As the severity of the disturbance increases, so does the performance. According to the robustness evaluation findings, Daiwa has desirable resilience against disturbances of minor or medium scale. Of course, as the strength of the disturbance increases, the inpainting accuracy decreases, and significant perturbations result in badly deteriorated pictures, which defeats the objective of inpainting. This behavior is consistent with the findings from [114].

The results of the robustness study show that our proposed Daiwa model has a satisfactory level of resilience against disturbances of small to medium scale noise. Naturally, when the noise intensity increases, the inpainting evidence will be lost, leading to significant detection mistakes. However, significant noise amount can produce pictures that are highly damaged, which defeats the goal of applying inpainting.

E. PERFORMANCE EVALUATION

Picture inpainting is known to be directly proportional to the input image resolution and how distant the linked pixels

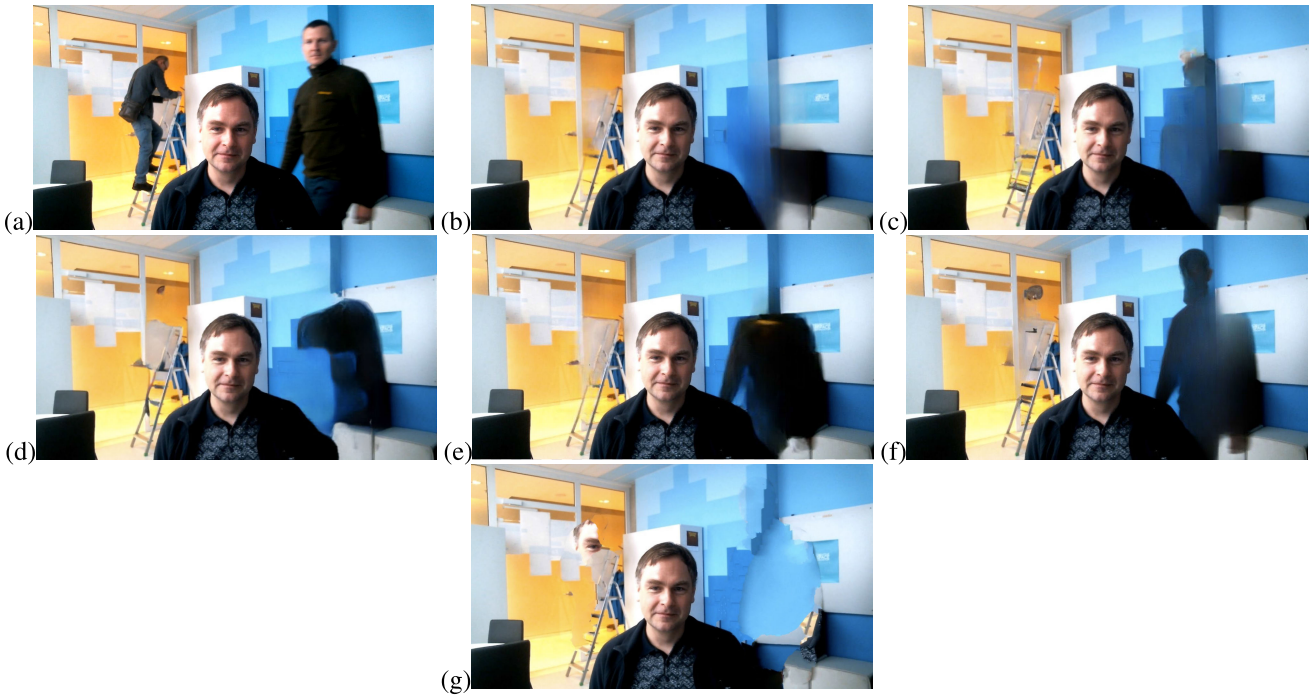


FIGURE 4. Comparison of background person removal (top to bottom): (a) original frame, (b) Daiwa, (c) E2FGVI, (d) LaMa, (e) MADF, (f) RegionWise, (g) AOTGAN.

TABLE 1. Results on DAVIS dataset: comparison with state-of-the-art method performance.

	Narrow masks			Wide masks			Segmentation masks		
Method	LPIPS	FID	SSIM	LPIPS	FID	SSIM	LPIPS	FID	SSIM
Daiwa	0.08	0.58	0.71	0.11	1.99	0.76	0.02	5.01	0.79
E2FGVI [78]	-	-	-	-	-	-	0.03	5.09	0.72
LaMa [104]	0.11	0.66	0.62	0.18	2.88	0.59	0.02	5.22	0.62
MADF [47]	0.28	0.79	0.63	0.37	2.49	0.61	0.22	6.22	0.61
RegionWise [112]	0.41	0.88	0.59	0.55	3.51	0.62	0.24	7.01	0.65
AOTGAN [113]	0.57	1.12	0.60	0.69	3.88	0.60	0.31	6.98	0.59

TABLE 2. Results on Youtube-VOS dataset: comparison with state-of-the-art methods.

	Narrow masks			Wide masks			Segmentation masks		
Method	LPIPS	FID	SSIM	LPIPS	FID	SSIM	LPIPS	FID	SSIM
Daiwa	0.12	1.22	0.68	0.58	4.12	0.69	0.03	6.22	0.78
E2FGVI [78]	-	-	-	-	-	-	0.03	6.02	0.71
LaMa [104]	0.18	1.41	0.62	0.71	4.88	0.63	0.03	6.91	0.64
MADF [47]	0.59	2.41	0.60	1.22	6.11	0.62	0.34	8.78	0.64
RegionWise [112]	1.31	2.99	0.59	1.87	7.52	0.61	0.38	8.61	0.60
AOTGAN [113]	1.67	2.82	0.54	2.22	6.99	0.60	0.29	7.58	0.59

TABLE 3. Results on our occluded person posture (sitting behind a table) dataset.

	Narrow masks			Wide masks			Segmentation masks		
Method	LPIPS	FID	SSIM	LPIPS	FID	SSIM	LPIPS	FID	SSIM
Daiwa	0.04	0.38	0.74	0.07	1.09	0.76	0.02	4.01	0.78
E2FGVI [78]	-	-	-	-	-	-	0.03	6.12	0.68
LaMa [104]	1.78	2.99	0.54	1.98	5.45	0.54	0.03	8.04	0.61
MADF [47]	1.99	4.54	0.53	2.37	7.41	0.54	0.44	9.45	0.61
RegionWise [112]	2.81	4.96	0.46	2.78	8.66	0.49	0.41	10.44	0.55
AOTGAN [113]	2.46	4.72	0.45	2.56	7.97	0.44	0.55	9.49	0.50

are positioned within the image from the missing regions. We conducted a dedicated experiment to focus on the computation time necessary to complete the inpainting. We have selected the DAVIS dataset in order to assess our suggested model in terms of computing cost, using 512×512 blocks of images (only the category with humans in the picture). Table 4 shows the model, platform and computational expenses (compute time, memory) for the evaluated techniques. As can be

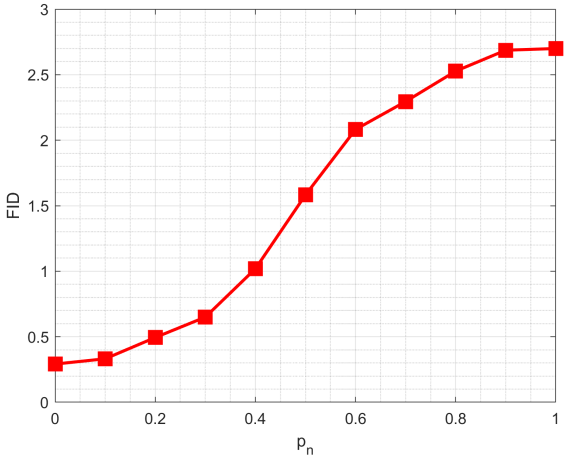


FIGURE 5. Evaluation of input noisyness on the performance of Daiwa model.

TABLE 4. Computational performance on davis dataset (categories with humans).

Model	Platform	Infill (ms)	RAM (Gb)
Daiwa	*CPU*	3141	**0.61**
Daiwa	*GPU*	274	**2.05**
E2FGVI [78]	*CPU*	2746	1.12
E2FGVI [78]	*GPU*	239	4.22
LaMa [104]	*CPU*	**2612**	0.78
LaMa [104]	*GPU*	**226**	3.21
MADF [47]	*CPU*	4208	1.03
MADF [47]	*GPU*	405	4.03
RegionWise [112]	*CPU*	4586	0.66
RegionWise [112]	*GPU*	312	3.66
AOTGAN [113]	*CPU*	6248	1.25
AOTGAN [113]	*GPU*	784	4.84

observed, our suggested technique has the lowest computing cost, demonstrating that our proposed model not only achieves strong efficacy in terms of inpainting quality, but also good efficiency in terms of processing time in milliseconds for both GPU (consumer card: Geforce 2060 RTX Super) and CPU (consumer cpu: AMD Ryzen 7 5700G). As it can be seen LaMa is about 15% more efficient on CPU/GPU, but consumes more memory as this approach was more of a "universal" design in comparison to our approach, with flow-based based E2FGVI trailing close behind. We believe the computational load could be reduced to acceptable 100 ms using a more powerfull GPU or running a few cards in parallel.

The performance of the methods is summarized in Figure 6. We analyze performance according to two criteria: infill time and memory consumption. We treat the problem of finding the best model as multi-objective optimization problem, ie. selecting a most preferred solution based on more than one criterion as follows.

$$\min_{x \in Q} \{F(x)\}, \tag{16}$$

where $Q \subset \mathbb{R}^n$ and F is a vector of the objective functions $F : Q \to \mathbb{R}^k$, $F(x) = (f_1(x), \ldots, f_k(x))$, and where each $f_i : Q \to \mathbb{R}$ is continuously differentiable.

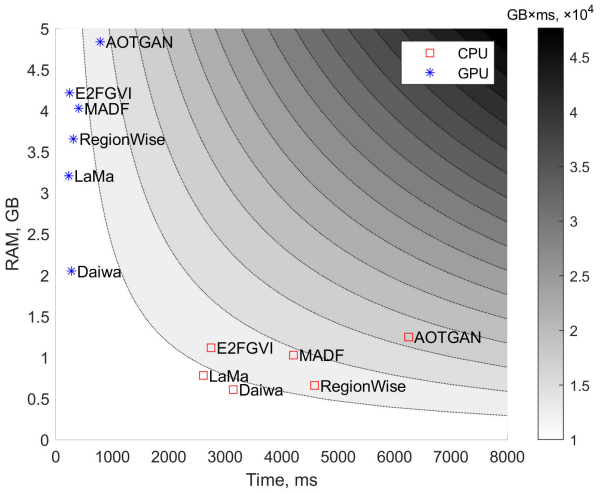


FIGURE 6. Summary of performance with Pareto fronts.

Let $v, w \in Q$. v is *smaller than* w ($v <_p w$), if $v_i < w_i$ for all $i \in \{1, \ldots, k\}$. The relation $\leq_p$ is set accordingly. $y \in \mathbb{R}^n$ is *dominated* by a point $x \in Q$ ($x \prec y$) with respect to (16) if $F(x) \leq_p F(y)$ and $F(x) \neq F(y)$, else y is nondominated by x. $x \in Q$ is a *Pareto point* if there is no $y \in Q$ which dominates x.

In this study we use the product of infill time and memory as the objective function and calculate the Pareto fronts for various values of the objective function. Note that the performance of Daiwa on both CPU and GPU is Pareto optimal, that is the models dominate over over models.

V. DISCUSSION AND CONCLUSION

We have presented a new method for removing the background from an image that contains human-like shapes using a semantic-aware occlusion-robust architecture. The method consists of four primary components: feature extraction and three branches (local, global branch, and semantic). By combining these components, we were able to extract global features for the global branch to encode occlusion-aware local information, and retrieve fine-grained local characteristics for the local branch. The semantic mask was constructed for the semantic branch to indicate the non-occluded sections. We then employed a modified big mask inpainting approach to eliminate the person, which leverages Fast Fourier convolutions (FFC) and generalizes well to higher resolutions while requiring fewer parameters and taking less time. The use of dilation also improved the identified human masks, resulting in higher exterior borders and fewer ghosting artifacts. The performance of our model, Daiwa, was evaluated using the LPIPS, FID, and SSIM measurements, which are standard procedure metrics for comparison. Our results showed that Daiwa performed well in fully automated background person removal, and other models were not far behind in manual mask selection. The semantic masks produced the best result in almost all cases.

However, there are some limitations to this study that need to be addressed. Firstly, the study only employs a sample of

persons from the DAVIS and Youtube-VOS datasets, which may not be representative of all human-like objects. This limits the generalizability of the results. Secondly, the study only compares the performance of the proposed method with a few other models, and more models need to be evaluated in order to establish the validity of the results. Additionally, the study does not explore the scalability of the method, and it is not known how well the method will perform on large images with many human-like objects. Furthermore, the study only focuses on removing human-like objects from images, and does not address the challenge of removing other types of objects, such as animals or inanimate objects. Finally, the study does not consider the computational efficiency of the method, and it is not known how well the method will perform in real-time applications.

In conclusion, our approach provides a new and effective solution for removing the background from an image that contains human-like shapes. The use of semantic-aware occlusion-robust network and the modified big mask inpainting approach enabled us to produce high-quality results that are resistant to occlusion and generalize well to higher resolutions. However, it should be noted that the high values shown in previous studies could not be duplicated in this study, likely due to the fact that we only used a sample of persons from the DAVIS and Youtube-VOS datasets. Further research is needed to validate the efficiency of our methodology on larger and more diverse datasets.

VI. ABBREVIATIONS

We used these abbreviations:

DL	Deep learning.
GAN	Generative Adversarial Network.
CNN	Convolutional Neural Network.
LSTM	Long short-term memory.
DAVIS	Dateset used.
YouTube-VOS	Dateset used.
Pytorch	Machine learning framework used.
GPU	Graphical Processing Unit.
CPU	Central Processing Unit.
ROI	Region of interest.
Leaky ReLu	Activation function.
Blazepose	Skeleton detection model.
FFC	Fast Fourier convolutions.
FFT	Fast Fourier transform.
SSIM	Structural Similarity Index.
LPIPS	Learned Perceptual Image Patch Similarity.
FID	Frechet inception distance.
LaMa	Competing approach.
MADF	Competing approach.
Regionwise	Competing approach.
AOTGAN	Competing approach.
E2FGVI	Competing approach.
H.264	Video codec.

REFERENCES

[1] K. Schwab, *Stakeholder Capitalism: A Global Economy That Works for Progress, People and Planet*. Nashville, TN, USA: Wiley, Jan. 2021.

[2] T. Malleret and K. Schwab, *Great Narrative (The Great Reset Book 2)*. 2021.

[3] J. Pearson, *Next Up*. Chicago, IL, USA: Moody Press, Jun. 2014.

[4] Y.-L. Chang, Z. Y. Liu, and W. Hsu, "VORNet: Spatio-temporally consistent video inpainting for object removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–10.

[5] V. Bartl, J. Šaňhel, and A. Herout, "PersonGONE: Image inpainting for automated checkout solution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 3115–3123.

[6] M. K. J. Khan, N. Ud Din, S. Bae, and J. Yi, "Interactive removal of microphone object in facial images," *Electronics*, vol. 8, no. 10, p. 1115, Oct. 2019.

[7] L. Liao, J. Xiao, Z. Wang, C.-W. Lin, and S. Satoh, "Image inpainting guided by coherence priors of semantics and textures," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 6539–6548.

[8] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon, "Deep video inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019.

[9] Z. Xu and J. Sun, "Image inpainting by patch propagation using patch sparsity," *IEEE Trans. Image Process.*, vol. 19, no. 5, pp. 1153–1165, May 2010.

[10] Q. Guo, S. Gao, X. Zhang, Y. Yin, and C. Zhang, "Patch-based image inpainting via two-stage low rank approximation," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 6, pp. 2023–2036, Jun. 2017.

[11] A. Bugeau, M. Bertalmío, V. Caselles, and G. Sapiro, "A comprehensive framework for image inpainting," *IEEE Trans. Image Process.*, vol. 19, no. 10, pp. 2634–2645, Oct. 2010.

[12] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, and Y. Akbari, "Image inpainting: A review," *Neural Process. Lett.*, vol. 51, no. 2, pp. 2007–2028, Dec. 2019.

[13] W. Xiong, J. Yu, Z. Lin, J. Yang, X. Lu, C. Barnes, and J. Luo, "Foreground-aware image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5840–5848.

[14] B. Shen, W. Hu, Y. Zhang, and Y.-J. Zhang, "Image inpainting via sparse representation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 697–700.

[15] Z. Li, H. He, H. M. Tai, Z. Yin, and F. Chen, "Color-direction patch-sparsity-based image inpainting using multidirection features," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 1138–1152, Mar. 2015.

[16] C. Guillemot and O. Le Meur, "Image inpainting : Overview and recent advances," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 127–144, Jan. 2014.

[17] M. Ghorai, S. Samanta, S. Mandal, and B. Chanda, "Multiple pyramids based image inpainting using local patch statistics and steering kernel feature," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5495–5509, Nov. 2019.

[18] A. Newson, A. Almansa, Y. Gousseau, and P. Pérez, "Non-local patch-based image inpainting," *Image Process. On Line*, vol. 7, pp. 373–385, Dec. 2017, doi: 10.5201/ipol.2017.189.

[19] H. Wang, L. Jiang, R. Liang, and X.-X. Li, "Exemplar-based image inpainting using structure consistent patch matching," *Neurocomputing*, vol. 269, pp. 90–96, Dec. 2017.

[20] D. Zhang, Z. Liang, G. Yang, Q. Li, L. Li, and X. Sun, "A robust forgery detection algorithm for object removal by exemplar-based image inpainting," *Multimedia Tools Appl.*, vol. 77, no. 10, pp. 11823–11842, 2018.

[21] N. Zhang, H. Ji, L. Liu, and G. Wang, "Exemplar-based image inpainting using angle-aware patch matching," *EURASIP J. Image Video Process.*, vol. 2019, no. 1, pp. 1–13, Jul. 2019.

[22] T. Ruzic and A. Pizurica, "Context-aware patch-based image inpainting using Markov random field modeling," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 444–456, Jan. 2015.

[23] W. Li and M. Wozniak, "A hole filling and optimization algorithm of remote sensing image based on bilateral filtering," *Mobile Netw. Appl.*, vol. 27, no. 2, pp. 743–751, Apr. 2022.

[24] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6721–6729.

[25] H. Liu, B. Jiang, Y. Xiao, and C. Yang, "Coherent semantic attention for image inpainting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019 pp. 4170–4179.

[26] X. Zhu, Y. Qian, X. Zhao, B. Sun, and Y. Sun, "A deep learning approach to patch-based image inpainting forensics," *Signal Process., Image Commun.*, vol. 67, pp. 90–99, Sep. 2018.

[27] U. Demir and G. Unal, "Patch-based image inpainting with generative adversarial networks," 2018, *arXiv:1803.07422*.

[28] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4471–4480.

[29] Z. Qin, Q. Zeng, Y. Zong, and F. Xu, "Image inpainting based on deep learning: A review," *Displays*, vol. 69, Sep. 2021, Art. no. 102028.

[30] H. Xiang, Q. Zou, M. A. Nawaz, X. Huang, F. Zhang, and H. Yu, "Deep learning for image inpainting: A survey," *Pattern Recognit.*, vol. 134, Feb. 2023, Art. no. 109046.

[31] X. Zhang, D. Zhai, T. Li, Y. Zhou, and Y. Lin, "Image inpainting based on deep learning: A review," *Inf. Fusion*, vol. 90, pp. 74–94, Feb. 2023.

[32] H. Zheng, Z. Zhang, H. Zhang, Y. Yang, S. Yan, and M. Wang, "Deep multi-resolution mutual learning for image inpainting," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 6359–6367.

[33] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 85–100.

[34] W. Wang, J. Zhang, L. Niu, H. Ling, X. Yang, and L. Zhang, "Parallel multi-resolution fusion network for image inpainting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14559–14568.

[35] H. Xu, X. Li, K. Zhang, Y. He, H. Fan, S. Liu, C. Hao, and B. Jiang, "SR-inpaint: A general deep learning framework for high resolution image inpainting," *Algorithms*, vol. 14, no. 8, p. 236, Aug. 2021.

[36] X. Wang, S. Niu, and H. Wang, "Image inpainting detection based on multi-task deep learning network," *IETE Tech. Rev.*, vol. 38, no. 1, pp. 149–157, Jun. 2020.

[37] T. Bapaume, E. Come, J. Roos, M. Ameli, and L. Oukhellou, "Image inpainting and deep learning to forecast short-term train loads," *IEEE Access*, vol. 9, pp. 98506–98522, 2021.

[38] S. Lee, S. W. Oh, D. Won, and S. J. Kim, "Copy-and-paste networks for deep video inpainting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4413–4421.

[39] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, "Recurrent feature reasoning for image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7760–7768.

[40] Y. Zhou, C. Barnes, E. Shechtman, and S. Amirghodsi, "TransFill: Reference-guided image inpainting by merging multiple color and spatial transformations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2266–2276.

[41] M. Suin, K. Purohit, and A. N. Rajagopalan, "Distillation-guided image inpainting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2481–2490.

[42] Y. Ren, X. Yu, R. Zhang, T. H. Li, S. Liu, and G. Li, "StructureFlow: Image inpainting via structure-aware appearance flow," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 181–190.

[43] X. Guo, H. Yang, and D. Huang, "Image inpainting via conditional texture and structure dual generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14134–14143.

[44] H. Liu, B. Jiang, Y. Song, W. Huang, and C. Yang, "Rethinking image inpainting via a mutual encoder–decoder with feature equalizations," in *Computer Vision—ECCV 2020*. Springer, 2020, pp. 725–741.

[45] T. Wang, H. Ouyang, and Q. Chen, "Image inpainting with external-internal learning and monochromic bottleneck," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5120–5129.

[46] N. Wang, Y. Zhang, and L. Zhang, "Dynamic selection network for image inpainting," *IEEE Trans. Image Process.*, vol. 30, pp. 1784–1798, 2021.

[47] M. Zhu, D. He, X. Li, C. Li, F. Li, X. Liu, E. Ding, and Z. Zhang, "Image inpainting by end-to-end cascaded refinement with mask awareness," *IEEE Trans. Image Process.*, vol. 30, pp. 4855–4866, 2021.

[48] J. Jam, C. Kendrick, K. Walker, V. Drouard, J. G.-S. Hsu, and M. H. Yap, "A comprehensive review of past and present image inpainting methods," *Comput. Vis. Image Understand.*, vol. 203, Feb. 2021, Art. no. 103147.

[49] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[50] Y. Zeng, J. Fu, H. Chao, and B. Guo, "Learning pyramid-context encoder network for high-quality image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1486–1494.

[51] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi, "EdgeConnect: Generative image inpainting with adversarial edge learning," 2019, *arXiv:1901.00212*.

[52] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "EdgeConnect: Structure guided image inpainting using edge prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–10.

[53] W. Cai and Z. Wei, "PiiGAN: Generative adversarial networks for pluralistic image inpainting," *IEEE Access*, vol. 8, pp. 48451–48463, 2020.

[54] H. Zhang, Z. Hu, C. Luo, W. Zuo, and M. Wang, "Semantic image inpainting with progressive generative networks," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 1939–1947.

[55] Y. Chen, H. Zhang, L. Liu, X. Chen, Q. Zhang, K. Yang, R. Xia, and J. Xie, "Research on image inpainting algorithm of improved GAN based on two-discriminations networks," *Appl. Intell.*, vol. 51, no. 6, pp. 3460–3474, Nov. 2020.

[56] H. Liu, Z. Wan, W. Huang, Y. Song, X. Han, and J. Liao, "PD-GAN: Probabilistic diverse GAN for image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 9371–9381.

[57] H. Wu, J. Zhou, and Y. Li, "Deep generative model for image inpainting with local binary pattern learning and spatial attention," *IEEE Trans. Multimedia*, vol. 24, pp. 4016–4027, 2022.

[58] L. Zhao, Q. Mo, S. Lin, Z. Wang, Z. Zuo, H. Chen, W. Xing, and D. Lu, "UCTGAN: Diverse image inpainting based on unsupervised cross-space translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5741–5750.

[59] J. Yang, Z. Qi, and Y. Shi, "Learning to incorporate structure knowledge for image inpainting," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, Apr. 2020, pp. 12605–12612.

[60] Y. Zeng, Z. Lin, H. Lu, and V. M. Patel, "CR-fill: Generative image inpainting with auxiliary contextual reconstruction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14164–14173.

[61] A. Xia, Y. Gui, L. Yao, L. Ma, and X. Lin, "Exemplar-based object removal in video using GMM," in *Proc. Int. Conf. Multimedia Signal Process.*, vol. 1, May 2011, pp. 366–370.

[62] S. Thomas and J. Mathew, "Analysis of image inpainting and object removal methodologies," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 1085, no. 1, Feb. 2021, Art. no. 012007.

[63] B. H. Patil and P. M. Patil, "Image inpainting based on image mapping and object removal using semi-automatic method," in *Proc. Int. Conf. Adv. Commun. Comput. Technol. (ICACCT)*, Feb. 2018, pp. 368–371.

[64] T. T. Le, A. Almansa, Y. Gousseau, and S. Masnou, "Object removal from complex videos using a few annotations," *Comput. Vis. Media*, vol. 5, no. 3, pp. 267–291, Aug. 2019.

[65] D. J. Tuptewar and A. Pinjarkar, "Robust exemplar based image and video inpainting for object removal and region filling," in *Proc. Int. Conf. Intell. Comput. Control (I2C2)*, Jun. 2017, pp. 1–4.

[66] C. Guillemot, M. Turkan, O. Le Meur, and M. Ebdelli, "Object removal and loss concealment using neighbor embedding methods," *Signal Process., Image Commun.*, vol. 28, no. 10, pp. 1405–1419, Nov. 2013.

[67] A. J. Malhotra, A. Chopra, R. Dahiya, P. Yadav, and A. Singhal, "Background object removal and image inpainting to fill irregular holes," in *Innovations in Cyber Physical Systems*. Singapore: Springer, 2021, pp. 287–295.

[68] M. Mahajan and P. Bhanodia, "Image inpainting techniques for removal of object," in *Proc. Int. Conf. Inf. Commun. Embedded Syst. (ICICES)*, Feb. 2014, pp. 1–4.

[69] S. S. Mirkamali and P. Nagabhushan, "Object removal by depth-wise image inpainting," *Signal, Image Video Process.*, vol. 9, no. 8, pp. 1785–1794, Jul. 2014.

[70] M. Ebdelli, O. Le Meur, and C. Guillemot, "Video inpainting with short-term windows: Application to object removal and error concealment," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3034–3047, Oct. 2015.

[71] B. Besic and A. Valada, "Dynamic object removal and spatio-temporal RGB-D inpainting via geometry-aware adversarial learning," *IEEE Trans. Intell. Vehicles*, vol. 7, no. 2, pp. 170–185, Jun. 2022.

[72] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.

[73] J. Pyo, Y. G. Rocha, A. Ghosh, K. Lee, G. In, and T. Kuc, "Object removal and inpainting from image using combined GANs," in *Proc. 20th Int. Conf. Control, Autom. Syst. (ICCAS)*, Oct. 2020, pp. 1116–1119.

[74] X. Cai and B. Song, "Semantic object removal with convolutional neural network feature-based inpainting approach," *Multimedia Syst.*, vol. 24, no. 5, pp. 597–609, Feb. 2018.

[75] J. Zhang, T. Fukuda, and N. Yabuki, "Automatic object removal with obstructed Façades completion using semantic segmentation and generative adversarial inpainting," *IEEE Access*, vol. 9, pp. 117486–117495, 2021.

[76] J. Kim, J. Hyeon, and N. Doh, "Generative multiview inpainting for object removal in large indoor spaces," *Int. J. Adv. Robot. Syst.*, vol. 18, no. 2, Mar. 2021, Art. no. 172988142199654.

[77] R. Zhang, W. Li, P. Wang, C. Guan, J. Fang, Y. Song, J. Yu, B. Chen, W. Xu, and R. Yang, "AutoRemover: Automatic object removal for autonomous driving videos," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, Apr. 2020, pp. 12853–12861.

[78] Z. Li, C.-Z. Lu, J. Qin, C.-L. Guo, and M.-M. Cheng, "Towards an end-to-end framework for flow-guided video inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17562–17571.

[79] J. Wang, K. Lu, D. Pan, N. He, and B.-K. Bao, "Robust object removal with an exemplar-based image inpainting approach," *Neurocomputing*, vol. 123, pp. 150–155, Jan. 2014.

[80] N. Neelima and M. Arulvani, "Object removal by region based filling inpainting," in *Proc. Int. Conf. Emerg. Trends VLSI, Embedded Syst., Nano Electron. Telecommun. Syst. (ICEVENT)*, Jan. 2013, pp. 1–5.

[81] V. Krishnamoorthy and S. Mathi, "An enhanced method for object removal using exemplar-based image inpainting," in *Proc. Int. Conf. Comput. Commun. Informat. (ICCCI)*, Jan. 2017, pp. 1–5.

[82] M. Lu and S. Niu, "A detection approach using LSTM-CNN for object removal caused by exemplar-based image inpainting," *Electronics*, vol. 9, no. 5, p. 858, May 2020.

[83] A. V. Pinjarkar and D. J. Tuptewar, "Robust exemplar-based image and video inpainting for object removal and region filling," in *Computing, Communication and Signal Processing* (Advances in Intelligent Systems and Computing). Singapore: Springer, Sep. 2018, pp. 817–825.

[84] A. Volkov, V. Efimova, V. Shalamov, and A. Filchenkov, "Keypoint-based static object removal from photographs," *Proc. SPIE*, vol. 11605, Jan. 2021, Art. no. 116050O.

[85] G. Tudavekar, S. R. Patil, and S. S. Saraf, "Dual-tree complex wavelet transform and super-resolution based video inpainting application to object removal and error concealment," *CAAI Trans. Intell. Technol.*, vol. 5, pp. 314–319, Dec. 2020.

[86] M. Banday, "Efficient object removal and region filling image refurbishing approach," in *Renewable Power for Sustainable Growth* (Lecture Notes in Electrical Engineering). Singapore: Springer, 2021, pp. 87–98.

[87] S. V. Bonde and R. P. Borole, "Object removal and image restoration within subspaces by prioritized patch optimization," in *Proc. 6th Int. Conf. Signal Process. Commun. (ICSC)*, Mar. 2020, pp. 128–137.

[88] P. M. Patil and B. H. Deokate, "Image mapping and object removal in image inpainting using wavelet transform," in *Proc. Int. Conf. Inf. Process. (ICIP)*, Dec. 2015, pp. 114–118.

[89] L. Zhang, "A novel image inpainting method for object removal based on structure sparsity," *Int. J. Performability Eng.*, vol. 14, no. 11, p. 2777, 2018.

[90] L. Zhang and M. Chang, "An image inpainting method for object removal based on difference degree constraint," *Multimedia Tools Appl.*, vol. 80, no. 3, pp. 4607–4626, Sep. 2020.

[91] L. Zhang and M. Chang, "Image inpainting for object removal based on adaptive two-round search strategy," *IEEE Access*, vol. 8, pp. 94357–94372, 2020.

[92] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 724–732.

[93] S. Caelles, A. Montes, K.-K. Maninis, Y. Chen, L. Van Gool, F. Perazzi, and J. Pont-Tuset, "The 2018 Davis challenge on video object segmentation," 2018, *arXiv:1803.00557*.

[94] S. Caelles, J. Pont-Tuset, F. Perazzi, A. Montes, K.-K. Maninis, and L. Van Gool, "The 2019 Davis challenge on VOS: Unsupervised multi-object segmentation," 2019, *arXiv:1905.00737*.

[95] L. Yang, Y. Fan, and N. Xu, "Video instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5188–5197.

[96] N. Xu, L. Yang, Y. Fan, J. Yang, D. Yue, Y. Liang, B. Price, S. Cohen, and T. Huang, "YouTube-VOS: Sequence-to-sequence video object segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 585–601.

[97] N. Xu, L. Yang, Y. Fan, D. Yue, Y. Liang, J. Yang, and T. Huang, "YouTube-VOS: A large-scale video object segmentation benchmark," 2018, *arXiv:1809.03327*.

[98] A. Kulikajevas, R. Maskeliunas, and R. Damaševičius, "Detection of sitting posture using hierarchical image composition and deep learning," *PeerJ Comput. Sci.*, vol. 7, p. e442, Mar. 2021.

[99] W. Wan and J. Liu, "Nonlocal patches based Gaussian mixture model for image inpainting," *Appl. Math. Model.*, vol. 87, pp. 317–331, Nov. 2020.

[100] F. Orujov, R. Maskeliūnas, R. Damaševičius, and W. Wei, "Fuzzy based image edge detection algorithm for blood vessel detection in retinal images," *Appl. Soft Comput.*, vol. 94, Sep. 2020, Art. no. 106452.

[101] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device real-time body pose tracking," 2020, *arXiv:2006.10204*.

[102] I. Grishchenko, V. Bazarevsky, A. Zanfir, E. G. Bazavan, M. Zanfir, R. Yee, K. Raveendran, M. Zhdanovich, M. Grundmann, and C. Sminchisescu, "BlazePose GHUM holistic: Real-time 3D human landmarks and pose estimation," 2022, *arXiv:2206.11678*.

[103] X. Zhang, Y. Yan, J.-H. Xue, Y. Hua, and H. Wang, "Semantic-aware occlusion-robust network for occluded person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2764–2778, Jul. 2021.

[104] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with Fourier convolutions," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 2149–2159.

[105] Z. Zhang, X. Zhou, S. Zhao, and X. Zhang, "Semantic prior guided face inpainting," in *Proc. ACM Multimedia Asia*, Dec. 2019, pp. 1–6.

[106] H. Chu, H. Ma, and X. Li, "Pedestrian instance segmentation with prior structure of semantic parts," *Pattern Recognit. Lett.*, vol. 149, pp. 9–16, Sep. 2021.

[107] J. Snell, K. Ridgeway, R. Liao, B. D. Roads, M. C. Mozer, and R. S. Zemel, "Learning to generate images with perceptual similarity metrics," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 4277–4281.

[108] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[109] E. J. Nunn, P. Khadivi, and S. Samavi, "Compound Frechet inception distance for quality assessment of GAN created images," 2021, *arXiv:2106.08575*.

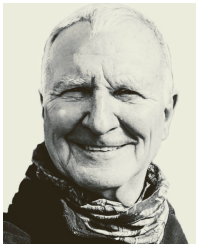
[110] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates, 2017, pp. 6629–6640.

[111] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[112] Y. Ma, X. Liu, S. Bai, L. Wang, A. Liu, D. Tao, and E. R. Hancock, "Regionwise generative adversarial image inpainting for large missing areas," *IEEE Trans. Cybern.*, early access, Aug. 17, 2022, doi: 10.1109/TCYB.2022.3194149.

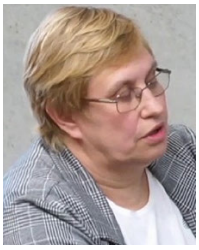
[113] Y. Zeng, J. Fu, H. Chao, and B. Guo, "Aggregated contextual transformations for high-resolution image inpainting," 2021, *arXiv:2104.01431*.

[114] H. Wu and J. Zhou, "IID-Net: Image inpainting detection network via neural architecture search and attention," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1172–1185, Mar. 2022.



RYTIS MASKELIUNAS (Member, IEEE) received the Ph.D. degree in computer science, in 2009. He is currently a Professor with the Department of Multimedia Engineering, Kaunas University of Technology, Kaunas, Lithuania, and an Invited Professor with the Faculty of Applied Mathematics, Silesian University of Technology, Poland. He is an Honorable Verbalizer. He is the author or coauthor of more than 150 refereed scientific articles and serves as a reviewer/committee member for various refereed journals. His research interests include multimodal signal processing, as well as modeling, development, and analysis of associative, multimodal interfaces, also targeted at the elderly and people with major disabilities. He has won various awards/honors, including the Best Young Scientist Award of 2012, the National Science Academy Award for Young Scholars of Lithuania, in 2010, and others.

ROBERTAS DAMAŠEVIČIUS (Member, IEEE) received the Ph.D. degree in informatics engineering from the Kaunas University of Technology, Lithuania, in 2005. He is currently a Professor with the Department of Applied Informatics, Vytautas Magnus University, Lithuania, and an Adjunct Professor with the Faculty of Applied Mathematics, Silesian University of Technology, Poland. He also lectures software maintenance, human–computer interface, and robot programming courses. He is the author of more than 500 articles and a monograph published by Springer. His research interests include sustainable software engineering, human–computer interfaces, assisted living, and explainability. He is also the Editor-in-Chief of the *Information Technology and Control* journal. He has been the Guest Editor of several invited issues of international journals, such as *BioMed Research International*, *Computational Intelligence and Neuroscience*, the *Journal of Healthcare Engineering*, IEEE Access, *Sensors*, and *Electronics*.



DAIVA VITKUTE-ADZGAUSKIENE received the degree and the Engineering degree in mathematics from the Faculty of Computing, Kaunas Polytechnic Institute (now Kaunas University of Technology), the master's degree in business administration from the Baltic Institute of Management, and the joint Ph.D. degree in informatics from VMU and the Institute of Mathematics and Informatics. Since the beginning of the restoration of Vytautas Magnus University, in 1989, she was with the Faculty of Informatics as a Researcher, and a docent, from 1995 to 2009. She was in managerial positions with Omnitel Company, developing and presenting internet and mobile communication services to the Lithuanian market. Currently, she is the Head of the Applied Informatics Department, Faculty of Informatics, Vytautas Magnus University. She is the coauthor of scientific publications and educational books. She also participates in the expert evaluation of university study programs.

SANJAY MISRA received the Ph.D. degree in information and knowledge engineering (software engineering) from the University of Alcalá, Spain, and the M.Tech. degree in software engineering from the MLN National Institute of Technology, India. Before coming to the prestigious HIOF, Norway, he has been the Research President and a top Professor (since October 2010) with Covenant University and chaired the Honorary Center of ICT/ICE Research and Lead–Soft Engineering Modeling and Intelligent Systems Research Group. He was also a Professor and the Head of the Federal University of Technology Minna and Atilim University, Turkey. He is a passionate research leader and is included among the top precious 2% of Scientists in the world for 2020 and 2021 published by Stanford University, USA. As per SciVal (Scopus–Elsevier) analysis, he has successfully guided eight doctorates and 14 master's theses as a sole supervisor. He has delivered more than 100 prestigious keynotes/invited talks/public lectures at highly reputed conferences and institutes. He is an Editor-in-Chief of *IT Personnel and Project Management* and *International Journal of Human Capital and Information Technology Professionals* (IGI Global), and an Editor and an Associate Editor in various top-level journals, including *Scientific Reports* (Nature) (IF:4.996) and *AEJ* (Elsevier) (I.F.:6.626). He has published 200 major publications and has edited (with colleagues) 58 LNCSs, four LNEEs, two LNNSs, three CCISs, ten IEEE proceedings, and six books.