Gesture Recognition by Self-Supervised Moving Interest Point Completion for CNN-LSTMs

1st Fotini Patrona
Department of Informatics
Aristotle University of Thessaloniki
fotinip@aiia.csd.auth.gr

2nd Ioannis Mademlis
Department of Informatics
Aristotle University of Thessaloniki
imademlis@csd.auth.gr

3rd Ioannis Pitas
Department of Informatics
Aristotle University of Thessaloniki
pitas@csd.auth.gr

***Abstract*—Gesture recognition, i.e., classification of videos depicting humans who perform hand gestures, is essential for Human-Computer Interaction. To this end, coupled Convolutional Neural Networks-Long Short-Term Memory architectures (CNN-LSTMs) are employed for fast semantic video analysis, but the typical transfer learning approach of initializing the CNN backbone using pretraining for whole-image classification is not necessarily ideal for spatiotemporal video understanding tasks. This paper investigates self-supervised CNN pretraining for a novel pretext task, relying on spatiotemporal video frame corruption via a set of low-level image/video processing building blocks that jointly force the CNN to learn to complete missing content. This is likely to coincide with visible moving object boundaries, including human body silhouettes. Such a CNN parameter set initialization is then able to augment gesture recognition performance, after retraining for this video classification downstream task, without inducing any runtime overhead during the inference stage. Evaluation on a gesture recognition dataset for autonomous Unmanned Aerial Vehicle (UAV) handling demonstrates the effectiveness of the proposed method, against both traditional ImageNet initialization and a competing self-supervised pretext task-based initialization.**

***Index Terms*—Self-Supervised Learning, Gesture Recognition, Convolutional Neural Networks, Long Short-Term Memory**

I. INTRODUCTION

Deep learning is a prominent area of AI research that is applicable to a wide variety of domains (e.g., medicine, financing, media and robotics [16]–[19], [23], [24], [29]). However, Deep Neural Networks (DNNs) require massive amounts of labeled training datasets, to avoid overfitting and to retain accuracy at the inference stage. *Transfer learning* offers a partial solution. A DNN is initially trained on a well-known, public, large-scale benchmark dataset for solving a basic task (e.g., object recognition from an entire image). The very large size of this dataset minimizes the chance of overfitting, even for very complex neural models. Subsequently, the pretrained DNN is retrained on the actual task we need, using a small, domain-specific labeled dataset. Initializing DNN parameters not randomly, but to the parameter set obtained by pretraining on the first, generic task, allows immediate transfer of learnt feature extraction patterns to the actually desired task and, thus, minimizes the risk of overfitting even with a small domain-specific dataset.

When using Convolutional Neural Networks (CNNs), it is common to initialize the feature extraction backbone CNN by training it on the ImageNet large-scale dataset for whole-image classification [3]. Then, small domain-specific labeled datasets can be used for training on the desired task. ImageNet pretraining allows the CNN to extract semantically rich and meaningful features from an image, that are subsequently further tuned to the desired task. However, generic ImageNet pretraining may not lead to optimal CNN initialization. This is particularly evident in video analysis tasks relying on spatiotemporal data relations. Gesture recognition constitutes exactly such a problem that is highly significant for several application domains, most notably Human-Computer Interaction. Given a sequence of RGB video frames, gesture recognition methods predict a gesture class belonging to a predefined set of supported gestures, thus classifying an entire video sequence. Various neural architectures have been proposed for handling this, such as combining CNNs with Long Short-Term Memory networks (CNN-LSTM), 3D CNNs, CNN-LSTMs that process precomputed 2D human body skeletons [25] instead of the raw RGB video frames [26], etc. 3D CNNs are the most accurate, but they are highly complex models and rather slow during the inference stage. Therefore, CNN-LSTM architectures are still commonly employed; typically by initializing the backbone CNN using ImageNet pretraining, despite the dataset's static image nature.

Self-supervised learning (SSL) can be utilized as a possible improvement over naive ImageNet pretraining. SSL focuses on extracting high-level, semantic visual representations from the input data by leveraging automatically created pseudo-labels. This is performed in a DNN pretraining stage using a so-called *pretext task*, i.e., learning to map variants of the training input data to pseudo-labels that are being automatically generated from the data themselves. Pretraining the DNN in a regular supervised manner on a suitable pretext task enforces the network to learn improved context-invariant features, easily transferable to another desired *downstream task*, such as gesture recognition, thus augmenting its performance on the latter one by reducing overfitting. In essence, SSL by pretext pretraining provides us with a better DNN parameter initialization to be used when training for the downstream task, therefore giving rise to increased accuracy at the inference stage, without *any* architectural modification and/or runtime

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No871479 (AERIAL-CORE).

overhead.

The most prominent types of pretext tasks involve content generation (e.g., GANs [11], colorization [41]), context structure (e.g., jigsaw puzzles [1] and geometric transformations [6]) or context similarity [38]. Pretext pretraining on images is usually exploited for downstream tasks like object detection [22], image classification or segmentation [6], while pretext pretraining on videos is mostly used for activity/gesture recognition from videos [35].

In this paper, a novel pretext task for SSL CNN pretraining is proposed. It leverages temporal video information by embedding it in each spatial 2D video frame representation, so as to increase accuracy in a gesture recognition downstream task. This is accomplished by suitably processing the input training video frames of a large-scale benchmark human activity recognition dataset, so that camera motion is first eliminated and, then, pixels surrounding spatial interest points located at the edges of visible moving objects are randomly corrupted. Subsequently, the proposed pretext task consists in learning to map each distorted training input video frame to its original version, thus forcing the feature extraction CNN to complete missing content by paying attention to video frame regions likely to depict moving human body silhouettes. A lightweight CNN can be pretrained in this manner, resulting in a better-than-ImageNet parameter initialization, and then employed (as the feature extraction component of a CNN-LSTM architecture) for regular downstream gesture recognition training.

II. Related Work

SSL methods can be divided into three categories, based on the data type used for pretext and downstream training: *image-based*, *video-based pretext* and *video-based*. The first two types focus on learning image representations in pretext pretraining and exploiting them on image-related downstream tasks. Their difference is that purely image-based/video-based pretext methods use image/video data in pretext pretraining, respectively. On the other hand, in video-based SSL approaches both the pretext and the downstream task concern videos.

Regarding image-based SSL, [2] adopts one of the most widely studied pretext tasks, jigsaw puzzle solving. The original image is spatially decomposed along a 3×3 grid and transformed into a jigsaw puzzle by shuffling the patches. During pretext pretraining, the DNN concurrently learns to identify the permutation indices and classify the original images. This method is enhanced in [22].Furthermore, image colorization is performed in [41], prediction of image rotation and exemplar are exploited in [40], while rotation prediction is handled as a classification problem in [6]. One-shot view grid prediction from a single 2D view is the pretext task posed in [11], while in [38] representations are learnt that are capable of discriminating among individual object instances. In [31] synthetic RGB images are leveraged for estimating instance contour, depth and surface normal. A different multitask approach is [4], advocating the combination of multiple pretext tasks (relative position, colorization, exemplar, motion segmentation).

Numerous pretext tasks for obtaining good image representations that leverage spatiotemporal information present in videos have been proposed over time, since videos can capture scene dynamics unavailable in static images. In [34] image-, shot- and video-level context is exploited to instill information from the different video granularities to their representations, while in [21] motion cues are combined with images in order to predict possible changes over time. Sequence sorting and foreground-background segmentation are adopted in [13] and [27], respectively. Finally, in [37], a siamese-triplet learns to predict similar representations for two tracked image patches of the same video and different representations for randomly sampled patches.

In video-based SSL, where the downstream task also concerns videos, an important milestone was [7], i.e., a pretext task for learning spatiotemporal video embeddings by predicting the future content. In contrast, video pace prediction is employed in [36], based on the assumption that video content understanding is a prerequisite for distinguishing between edited video variants with different pace. In [39], video clip order prediction is employed by pairwise concatenating video clip features extracted using 3D CNNs, while in [12] a *Space-Time Cubic Puzzles* pretext task is presented, i.e., the equivalent of image jigsaw puzzle for videos, only fusing clip features at the final fully-connected network layers and considering it as a permutation problem. This is also adopted by [1], which proposes a novel permutation strategy that preserves spatial coherence. In an entirely different approach, [35] aspires to learn spatiotemporal representations by employing prediction of video sequence motion and appearance statistics.

III. Moving Interest Point Completion

The method proposed in this paper is essentially a video-based SSL approach, since both the pretext and the downstream task act on video data, but the goal is to obtain good per video frame representations, as is typically the case with image-based pretext SSL approaches. Thus, during pretext pretraining, the CNN learns to embed temporal information from video content into the individual video frame descriptions it outputs. Subsequently, this pretrained CNN can be used as a feature extraction backbone in a typical CNN-LSTM setting and trained regularly in an end-to-end manner for a video classification downstream task, such as gesture recognition from RGB camera feed. The proposed *Moving Interest Point completion* pretext task, or MIP completion, is described below.

Basic image processing/computer vision methods are first utilized in concert, so as to automatically identify visible video frame areas most likely depicting moving object edges. The employed algorithms include optical flow estimation [5], Fast Fourier Transform [9] and SIFT keypoint detection [14]. These are exploited as building blocks in the context of the proposed algorithm, which purposefully corrupts in a very focused and

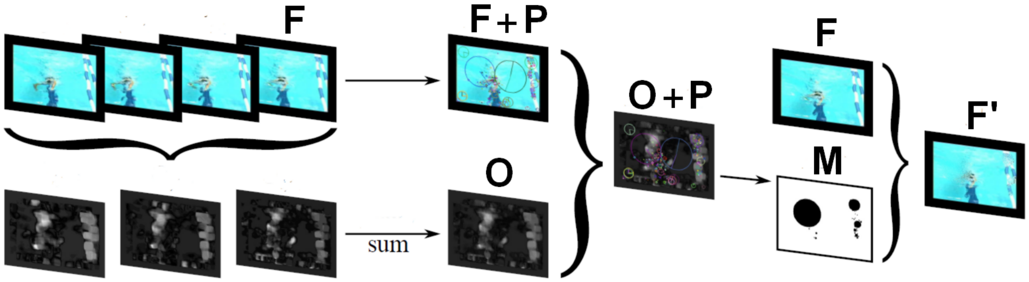


Fig. 1. Simplified training dataset preparation pipeline for the MIP completion pretext task. Given a video frame $\mathbf{F}$ and its three temporally preceding video frames, a carefully corrupted version $\mathbf{F}'$ is computed. $\mathbf{O}$ is the combined optical flow magnitude map, $\mathbf{P}$ is the interest point map and $\mathbf{M}$ is the moving interest point map.

specific manner all training video frames of a large-scale pretext dataset for human activity recognition.

Assuming that the currently processed video is composed of T consecutive RGB video frames $\mathcal{F}_t \in \mathbb{R}^{W \times H \times 3}$, $1 \leq\leq T$, a set of three dense optical flow magnitude maps $\mathbf{O}_{t,t-1}, \mathbf{O}_{t,t-2}, \mathbf{O}_{t,t-3}$ are computed for each video frame $\mathcal{F}_t$, relative to its 3 preceding video frames $\mathcal{F}_{t-1}, \mathcal{F}_{t-2}, \mathcal{F}_{t-3}$[1]. Subsequently, considering possible camera motion as undesired dominant noise, we zero-out the DC term of each optical flow map's 2D FFT [30] in order to eliminate the influence of camera motion. Afterwards, each optical flow map is transformed back to image space and, thus, the filtered optical flow maps $\tilde{\mathbf{O}}_{t,t-1}, \tilde{\mathbf{O}}_{t,t-2}, \tilde{\mathbf{O}}_{t,t-3}$ are obtained. Finally, a single, combined optical flow map $\mathbf{O}_t$ is acquired by weighted summation: $\mathbf{O}_t = (\tilde{\mathbf{O}}_{t,t-1} + 0.5 * \tilde{\mathbf{O}}_{t,t-2} + 0.5 * \tilde{\mathbf{O}}_{t,t-3})/2$. This initial process provides us with a scalar visible object motion estimate for each pixel of each video frame.

Simultaneously, an interest point map $\mathbf{P}_t \in \mathbb{R}^{W \times H}$ of the original video frame $\mathcal{F}_t$ is computed using the SIFT keypoint detector [14], although alternative interest point detectors would most likely be equally acceptable. SIFT keypoints can nowadays be computed extremely fast on images, with each one denoting a small, internally consistent image area whose local texture-wise or illumination-wise variability on multiple resolution scales makes it a distinctive spatial locus. SIFT keypoints are typically located at the peaks of blobs and along the ridges of lines [15], [20], [33], therefore along visible human silhouettes as well. Each such interest point is characterized by a center (in 2D pixel coordinates), a range/area (in pixels), and a strength (a numerical "degree of interest"). Thus, $\mathbf{P}_t$ is a grayscale image of resolution equal to that of $\mathcal{F}_t$, where all detected SIFT keypoints have been appropriately marked according to their center, range and strength.

At the next algorithm step, $\mathbf{P}_t$ and $\mathbf{O}_t$ are merged into a single moving interest point map $\mathbf{M}_t \in \mathbb{R}^{W \times H}$, that only contains SIFT keypoints spatially coinciding with video frame areas where local combined optical flow magnitude is larger that a dataset-specific intensity threshold o_c. Finally, $\mathbf{M}_t$ is applied as a mask over the original corresponding video frame $\mathcal{F}_t$, so that the RGB values of all pixels of $\mathcal{F}_t$ which fall within non-zero regions of $\mathbf{M}_t$ are substituted by the values of randomly selected neighboring pixels, while the rest of the video frame content remains unaltered. The end result is $\mathcal{F}_t^{'}$, i.e., a carefully corrupted version of original video frame $\mathcal{F}_t$, where image regions most likely depicting parts of edges of moving objects (thus, including human body silhouettes) have been randomly distorted in an automated manner, without relying on any human body location ground-truth information.

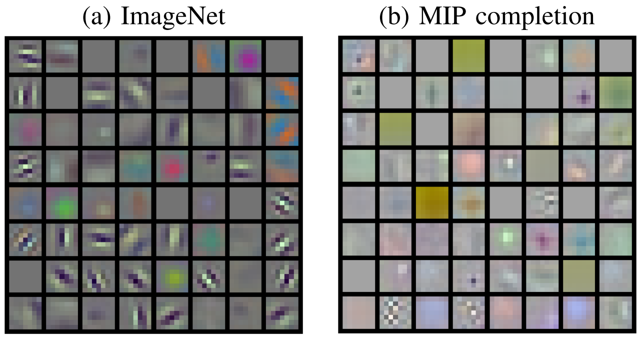
Figure 1 depicts a simplified schematic of the overall, above-described process. This is applied across the entire pretext training set, with the resulting corrupted video frames being subsequently employed as input video frames to the supervised pretext task. The corresponding original/undistorted video frames function as the respective pseudo-labels. A CNN trained on this task attempts to reconstruct each original video frame, given as input a corrupted version of itself. Thus, it learns to focus on missing video content which must be completed in the output and, due to the targeted nature of the input corruption, on visible regions that depict moving objects. Note that the approach may not work correctly if the pretraining video dataset, or the actual downstream (gesture recognition) video dataset, contains many rapid scale changes across consecutive video frames; however this is almost never the case with regular footage.

IV. Quantitative Evaluation

As in the vast majority of SSL approaches for videos, we chose "split1" of the UCF101 human activity recognition video dataset [32] for pretext pretraining. UCF101 is a benchmark dataset comprised of 13320 videos of 101 different activity categories with spatial resolution 320 × 240 pixels. Transferability of the features produced through pretext pretraining is evaluated by regular supervised training on a gesture recognition task using a randomly selected subset of the recently introduced AUTH UAV Gesture Dataset [28], designed for autonomous drone/Unmanned Aerial Vehicles (UAV) handling. The employed subset used consists of 275 videos in total, captured both indoors and outdoors, with

[1] 3 was determined empirically.

static and moving cameras, divided into 6 classes. Its data were amassed from several preexisting datasets, thus spatial resolutions vary.



(a) ImageNet

(b) MIP completion

Fig. 2. Conv1 filter visualization.

TABLE I
EVALUATION RESULTS ON THE EMPLOYED SUBSET OF THE AUTH UAV GESTURE DATASET, USING THE CORRECT CLASSIFICATION RATE (CCR) METRIC.

Initialization	CCR
Random	56.90 %
ImageNet	60.19 %
RotNet [6]	60.00 %
MIP (proposed)	**64.08 %**

Implementation Details. We employed a ResNet-18-based [8] CNN autoencoder for pretext pretraining. Starting with a regular ResNet-18 architecture in the role of a CNN encoder, which is the desired feature extraction backbone network, a mirror CNN decoder (composed of consecutive deconvolutional layers) was inserted just before the final average pooling layer. Thus, a CNN autoencoder was obtained and trained using binary cross-entropy (BCE) loss function for minimizing the original video frame reconstruction loss. Encoder parameters were initialized with ResNet-18 weights pretrained for whole-image classification on the ImageNet dataset.

Training input RGB video frames from "split1" of the UCF101 dataset were normalized using ImageNet dataset mean and std values, resized to 256×256 and afterwards randomly cropped to 224×224. Then, the proposed process for pretext task data preparation (described in Section III) was followed. The batch size used for pretext pretraining was 64 and a SGD optimizer with momentum 0.98 and weight decay of 0.001 was employed. The initial learning rate was 0.1, decaying every 30 epochs, and training was stopped after 200 epochs.

After pretext pretraining, the ResNet-18 encoder was attached as backbone feature extraction network to an LSTM [10] network with input dimension equal to 4096, i.e., the length of the feature vector produced by the backbone, 2 layers of 128 neurons each, and unrolling for 15 time steps. The entire CNN-LSTM architecture was trained end-to-end for the desired gesture recognition task on the AUTH UAV Gesture Dataset, using truncated backpropagation through time (BPTT) for 100 epochs. Learning rate was set to 0.001, decaying every 30 epochs. The batch size used was 20 and an SGD optimizer with momentum 0.98 and weight decay of 0.001 was employed.

Before downstream training, all gesture input videos were first set to a temporal length equal to 15 video frames, so that all samples are equally handled by the LSTM network. This was done by random video frame subsampling, for videos of length larger than 15 video frames, and random video frame duplication for videos of less than 15 frames.

Evaluation Results. The proposed MIP completion pretext task was examined with regard both to the nature of the representations it produces and their transferability to a gesture recognition downstream task. A widespread approach to qualitatively demonstrate the image representations produced by CNNs pretrained on pretext tasks is by visualizing the filters of their first convolutional layer ("Conv1"). Thus, Figure 2 compares the Conv1 filters obtained by: a) pretraining the ResNet-18 encoder alone on ImageNet for whole-image classification, and b) attaching to the ResNet-18 encoder a corresponding decoder and pretraining the overall model for the MIP completion pretext task. As expected, ImageNet classification pretraining provides good gradient/edge detection convolutional filters, while UCF101 MIP completion pretraining provides filters sensitive to spatial texture frequency and appearance. Thus, initializing the CNN for the downstream task using the proposed method renders the model able to better distinguish between visible human body regions (which are characterized by specific texture/appearance patterns) from the background, or from very differently looking objects.

Quantitative evaluation results on the selected subset of AUTH UAV Gesture Dataset are shown in Table I. Evidently, initializing the ResNet-18 part of the CNN-LSTM architecture for the desired gesture recognition downstream task using the parameter set obtained by MIP completion pretraining, gives rise to higher Correct Classification Rate (CCR) than both ImageNet initialization and than using competing pretext pretraining [6]. The reported MIP completion pretraining results were obtained by using a combined optical flow per-pixel intensity threshold of $o_c = 100$ (falling within the pixel intensity integer interval $[0, 255]$) when computing each $\mathbf{M}_t$, as described in Section III.

V. CONCLUSIONS

This paper presented a novel pretext task for pretraining a CNN under a video-based self-supervised learning (SSL) setting, where the intention is to later employ this model as a per video frame feature extraction backbone for a gesture recognition downstream task, in the context of a CNN-LSTM architecture. The proposed Moving Interest Point (MIP) completion pretext pretraining algorithm, relying on basic image processing building blocks such as optical flow estimation, image frequency transform and interest point detection, was shown to provide better CNN parameter set initialization than typical pretraining for whole-image classification on ImageNet, as well as than a competing SSL approach. Gesture classification accuracy was significantly augmented in a relevant gesture recognition dataset for Unmanned Aerial Vehicle (UAV) handling, without any modifications to the lightweight neural architecture or the input data. Thus, overall, the proposed method showcases the importance of proper, video-oriented parameter set initialization for CNN-LSTM architectures trained for analyzing spatiotemporal video content, when relying on transfer learning for combating CNN overfitting.

References

[1] U. Ahsan, R. Madhok, and I. Essa. Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019.

[2] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[4] C. Doersch and A. Zisserman. Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[5] G. Farneback. Very high accuracy velocity estimation using orientation tensors, parametric motion, and simultaneous segmentation of the motion field. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2001.

[6] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[7] T. Han, W. Xie, and A. Zisserman. Video representation learning by dense predictive coding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCV)*, 2019.

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

[9] M. Heideman, D. Johnson, and C. Burrus. Gauss and the history of the Fast Fourier Transform. *IEEE Signal Processing Magazine*, 1:14–21, 1984.

[10] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

[11] D. Jayaraman, R. Gao, and K. Grauman. ShapeCodes: self-supervised feature learning by lifting views to viewgrids. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[12] D. Kim, D. Cho, and I. S. Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[13] H.-Y. Lee, J.-B. Huang, M. Singh, and M.-H. Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision (ICIP)*, 2017.

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[15] I. Mademlis, A. Iosifidis, A. Tefas, N. Nikolaidis, and I. Pitas. Exploiting stereoscopic disparity for augmenting human activity recognition performance. *Multimedia Tools and Applications*, 75(19):11641–11660, 2016.

[16] I. Mademlis, V. Mygdalis, N. Nikolaidis, M. Montagnuolo, F. Negro, A. Messina, and I. Pitas. High-level multiple-UAV cinematography tools for covering outdoor events. *IEEE Transactions on Broadcasting*, 65(3):627–635, 2019.

[17] I. Mademlis, V. Mygdalis, N. Nikolaidis, and I. Pitas. Challenges in autonomous UAV cinematography: An overview. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2018.

[18] I. Mademlis, N. Nikolaidis, A. Tefas, I. Pitas, T. Wagner, and A. Messina. Autonomous unmanned aerial vehicles filming in dynamic unstructured outdoor environments [applications corner]. *IEEE Signal Processing Magazine*, 36(1):147–153, 2018.

[19] I. Mademlis, N. Nikolaidis, A. Tefas, I. Pitas, T. Wagner, and A. Messina. Autonomous UAV cinematography: a tutorial and a formalized shot-type taxonomy. *ACM Computing Surveys (CSUR)*, 52(5):1–33, 2019.

[20] I. Mademlis, A. Tefas, N. Nikolaidis, and I. Pitas. Compact video description and representation for automated summarization of human activities. In *Proceedings of the INNS Conference on Big Data*. Springer, 2016.

[21] A. Mahendran, J. Thewlis, and A. Vedaldi. Cross pixel optical-flow similarity for self-supervised learning. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Springer, 2018.

[22] I. Misra and L. Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[23] S. Papadopoulos, I. Mademlis, and I. Pitas. Neural vision-based semantic 3D world modeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021.

[24] S. Papadopoulos, I. Mademlis, and I. Pitas. Semantic image segmentation guided by scene geometry. In *Proceedings of the IEEE International Conference on Autonomous Systems (ICAS)*, 2021.

[25] C. Papaioannidis, I. Mademlis, and I. Pitas. Fast CNN-based single-person 2D human pose estimation for autonomous systems. *IEEE Transactions on Circuits and Systems for Video Technology (under review)*, 2022.

[26] C. Papaioannidis, D. Makrygiannis, I. Mademlis, and I. Pitas. Learning fast and robust gesture recognition. In *Proceedings of the EURASIP European Signal Processing Conference (EUSIPCO)*. IEEE, 2021.

[27] D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan. Learning features by watching objects move. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[28] F. Patrona, I. Mademlis, and I. Pitas. An overview of hand gesture languages for autonomous UAV handling. In *Proceedings of the Aerial Robotic Systems Physically Interacting with the Environment Workshop (AIRPHARO)*, 2021.

[29] F. Patrona, I. Mademlis, A. Tefas, and I. Pitas. Computational UAV cinematography for intelligent shooting based on semantic visual analysis. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2019.

[30] I. Pitas. *Digital image processing algorithms and applications*. John Wiley & Sons, 2000.

[31] Z. Ren and Y. Jae Lee. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[32] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[33] C. Symeonidis, I. Mademlis, N. Nikolaidis, and I. Pitas. Improving neural Non-Maximum Suppression for object detection by exploiting interest-point detectors. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2019.

[34] M. Tschannen, J. Djolonga, M. Ritter, A. Mahendran, N. Houlsby, S. Gelly, and M. Lucic. Self-supervised learning of video-induced visual invariances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[35] J. Wang, J. Jiao, L. Bao, S. He, Y. Liu, and W. Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[36] J. Wang, J. Jiao, and Y.-H. Liu. Self-supervised video representation learning by pace prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020.

[37] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[38] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[39] D. Xu, J. Xiao, Z. Zhao, J. Shao, D. Xie, and Y. Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[40] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[41] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016.