Fast CNN-based Single-Person 2D Human Pose Estimation for Autonomous Systems

Christos Papaioannidis, Ioannis Mademlis *Senior Member, IEEE,* Ioannis Pitas *Fellow, IEEE*

***Abstract*—This paper presents a novel Convolutional Neural Network (CNN) architecture for 2D human pose estimation from RGB images that balances between high 2D human pose/skeleton estimation accuracy and rapid inference. Thus, it is suitable for safety-critical embedded AI scenarios in autonomous systems, where computational resources are typically limited and fast execution is often required, but accuracy cannot be sacrificed. The architecture is composed of a shared feature extraction backbone and two parallel heads attached on top of it: one for 2D human body joint regression and one for global human body structure modelling through Image-to-Image Translation (I2I). A corresponding multitask loss function allows training of the unified network for both tasks, through combining a typical 2D body joint regression with a novel I2I term. Along with enhanced information flow between the parallel neural heads via skip synapses, this strategy is able to extract both ample semantic and rich spatial information, while using a less complex CNN; thus it permits fast execution. The proposed architecture is evaluated on public 2D human pose estimation datasets, achieving the best accuracy-speed ratio compared to the state-of-the-art. Additionally, it is evaluated on a pedestrian intention recognition task for self-driving cars, leading to increased accuracy and speed in comparison to competing approaches.**

***Index Terms*—2D human pose estimation, Convolutional Neural Networks, Generative Adversarial Networks, self-driving cars, autonomous systems.**

I. Introduction

ESTIMATING 2D human poses/skeletons from RGB images is important in many applications that involve visually captured human activities. Given the recent developments in computer vision for autonomous systems (e.g., [1]–[16]), as well as the rapidly increasing interest in self-driving cars and human-car interfaces [17]–[19], 2D human pose estimation [20], [21] is becoming more and more significant for several relevant tasks. For instance, to monitor the engagement of the driver [22], to facilitate traffic control gesture recognition [23], or to recognize pedestrian intention [24]–[26]. The latter task is of paramount importance for human safety and consists in deciding whether a visible pedestrian is about to cross the road or not. 2D human pose estimation is widely employed by state-of-the-art methods (e.g., [27]–[29]) as a preprocessing step to extract 2D pedestrian skeletons from each video frame. Subsequently, the skeletons are fed as input to an intention classifier, which yields the final prediction. Similarly, 2D human pose estimation is crucial for skeleton-based action/gesture recognition [30]–[33] in real-world applications, as it is used to extract and provide 2D skeletons to the action/gesture classifier.

Essentially, 2D skeleton/human pose estimation entails estimating the pixel coordinates of a predefined set of human body joints on a 2D image (typically RGB). It is a challenging problem, as humans can appear in very different scenes and scales, under a huge range of body postures. Moreover, occlusion of certain body joints is typical in most cases, rendering 2D human pose estimation even more challenging.

Deep Convolutional Neural Networks (CNNs) are an effective algorithmic approach for handling the above issues. Many relevant CNN architectures have been proposed, with the most successful ones [34]–[36] being those that predict high-resolution outputs. Typically, a subnetwork first processes the input to decrease the resolution and extract semantic features, while a consecutively placed subnetwork is subsequently used to raise the resolution and produce the final output, from which the 2D human body joint positions can be obtained. Overall, a single monolithic network is tasked to encode both the spatial and the semantic information required for localizing and identifying each body joint on the 2D input image (e.g., a CNN encoder-decoder architecture). However, these methods either demonstrate insufficient accuracy, or only manage to perform well at the expense of execution speed. Often, the reason is the large number of convolutional/deconvolutional layers and downsampling/upsampling calculations that are necessary to improve performance, rendering these methods incapable of achieving fast inference due to high computational load. This issue is especially pronounced in embedded AI application domains, such as autonomous systems (self-driving cars, autonomous drones, etc.), where computational capabilities are typically limited.

Motivated by the difficulties 2D human pose estimators typically face when trying to offer both high accuracy and fast execution speed, this research proposes a novel CNN architecture able to balance well between these competing demands on embedded hardware.

To this end, an approach orthogonal to previous attempts is proposed: combining a regular 2D human pose regression head with an auxiliary Generative Adversarial Network (GAN)-based [37] Image-to-Image Translation (I2I) [38] head in a unified multihead architecture, trained in a multitask fashion. The two parallel neural heads are attached on top of a common CNN backbone for shared feature extraction. In order to

Christos Papaioannidis, Ioannis Mademlis and Ioannis Pitas are with Department of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece. email: cpapaionn@csd.auth.gr, imademlis@csd.auth.gr, pitas@csd.auth.gr

This work has received funding from the European Union's Seventh Horizon 2020 research and innovation programme under grant agreement number 871479 AERIAL-CORE.

The source code and the trained neural models are available on request, by sending an e-mail to cpapaionn@csd.auth.gr, imademlis@csd.auth.gr.

accommodate a single set of image features for both tasks, the overall CNN is trained with a composite loss function, composed of a regular regression term and a novel I2I term. The latter one is specifically designed to push the auxiliary neural head towards modelling the global human body structure. Although the two tasks are essentially distinct formulations of a single problem, the different learning paradigms of the two heads (supervised and adversarial) lead to partially different patterns needed to be extracted from the common RGB image inputs.

Typically, I2I involves translating an input image to a corresponding output image which has similar general structure but is not identical with respect to local details (e.g., day-to-night images, grayscale-to-color images, etc.) [38]. In the proposed method, I2I is utilized to model the global spatial human body structure as an additional auxiliary task during training, by translating an input RGB image to the corresponding human body structure image. The related ground-truth images for training purposes are manually constructed by suitably processing the ground-truth locations of body joints on each training image. The underlying intuition is that by training the auxiliary I2I head for this objective in an adversarial manner, spatial information about the global human body structure is encoded in its intermediate features. *This is due to the fact that GAN-based I2I demands that the auxiliary neural head produces realistic outputs that resemble the ground-truth human body structure images in their overall composition, but not necessarily match them exactly in their local details*. Unsurprisingly, successful execution of such a task relies mainly on modelling the global spatial structure of the human body, instead of localized semantic and 2D positional information about each joint. Thus, *the various subtasks implicitly involved in 2D body joint estimation are explicitly partitioned among the two parallel heads*: the auxiliary/main head is assigned global spatial/local semantic+spatial modelling, respectively. Moreover, the information about global body structure internally encoded in the layers of the auxiliary I2I head is passed over to the main 2D human pose regression head, via skip synapses properly placed to facilitate interhead information flow, during both the training and the inference stage. Thus, the main neural head is able to enrich its semantic features before outputting its final predictions, in order to better identify and precisely localize each body joint on the 2D image.

The proposed unified network architecture is trained end-to-end in a multitask setting. This forces the backbone network to extract features that explicitly facilitate both global spatial human body structure modelling (by the auxiliary neural head) and localized 2D body joint regression (by the main neural head). Thus, rich baseline features for accurate 2D human pose estimation are extracted, even without the skip synapses. Thus the proposed method leads to increased 2D human pose estimation accuracy, which allows us to employ a more shallow/lightweight (therefore faster) network architecture for the backbone network and the neural heads, without any reductions in test accuracy compared to the state-of-the-art. In this manner a better accuracy-speed ratio is achieved, rendering the presented approach particularly suitable for embedded AI applications (such as autonomous systems).

Experiments on two public 2D human pose estimation datasets show that the proposed method outperforms in terms of accuracy the baseline architecture, ceteris paribus, as well as all competing methods. Moreover, its required inference runtime is equal to or smaller than that of the state-of-the-art. Thus, overall, *the proposed method offers the best accuracy-speed ratio*. As a case study for application in autonomous systems, the presented architecture is also evaluated on a pedestrian intention recognition task for self-driving cars, given this problem's importance in ensuring autonomous vehicle safety. Integrating into the proposed CNN a simple Long Short-Term Memory (LSTM) classification head leads to increased accuracy, in comparison to directly comparable competing approaches. Finally, detailed ablation studies demonstrate the contribution of each component of the proposed CNN architecture.

To summarize, the novelties of this work are twofold:

- First, a novel reformulation of the 2D human pose estimation problem as an I2I task (global body structure modelling) is proposed, along with a respective loss term for training a conditional GAN to solve it.
- Second, a novel CNN architecture is introduced that consists of a base feature extraction CNN, two parallel neural heads and a set of skip synapses conjoining them. It effectively combines regression and I2I to increase 2D human pose estimation performance compared to the baseline, by partitioning the involved computational subtasks among the two heads. Thus it offers the best accuracy-speed ratio.

The rest of the paper is organized as follows. Previous 2D human pose estimation approaches are discussed in Section II. The proposed 2D human pose estimation CNN architecture is described in Section III. The experimental setup and the extensive evaluation of the proposed CNN compared to state-of-the-art 2D human pose estimation methods, as well as detailed ablation studies can be found in Section IV. Finally, conclusions are drawn in Section V.

The source code and the trained neural models are available on request, by sending an e-mail to `cpapaionn@csd.auth.gr, imademlis@csd.auth.gr`.

II. Related Work

2D human pose estimation approaches vary: pictorial structure models [39], deformable part models [40] and deep learning [41] have all been tried. This paper focuses only on state-of-the-art CNN-based methods.

First, a set of relevant algorithms relies on directly regressing the 2D pixel coordinates of a predefined set of human body joints [41], [42]. For example, [41] aimed to achieve high-precision 2D body joint estimates by training a cascade of pose regressors, where the body joint predictions of a network stage is refined at each subsequent stage by predicting a displacement of the joint locations to the true location. In a similar fashion, the initial body joint location predictions are progressively changed in [42] to obtain the final predictions using an Iterative Error Feedback process.

More recent approaches [34]–[36], [43] predict the 2D body joint locations indirectly. These CNNs output body joint

heatmaps; then, heat maxima indicate 2D body joint locations. [44] models human body structure using a bidirectional tree structured CNN-based network, so that the feature channels at a body joint can receive information from other joints. In order to predict accurate body joint heatmaps, [45] combined CNNs with a deformable mixture of parts model in an end-to-end framework. With the same goal in mind, Stacked Hourglass [34] used a CNN architecture consisting of sequential "hourglass" modules, with features across scales combined to output high-resolution maps. Stacked hourglass networks were also adopted in [46] to generate body joint heatmaps from features at multiple resolutions, which are then utilized in Conditional Random Fields (CRFs) to refine predictions. CPN [43] decomposed the 2D human pose estimation problem into two steps. In the first step, a feature pyramid CNN is used to localize the "easy" body joints, such as hands. Then, the multi-scale feature maps are fused and given as input to a second network tasked to detect the "hard" body joints using an online hard joint mining loss function. In addition, a two-stage 2D human pose estimation framework was proposed in [20], where multiple Independent Losses Pose Nets (ILPNs) were first employed to infer body joint locations on a global level, while Convolutional Local Detectors (CLDs) were subsequently tasked to refine the body joint detections in the potential image regions indicated by the ILPNs. In contrast, [47] aimed to learn the human pose quality alongside 2D human pose regression by augmenting typical network architectures with a pose quality prediction neural block, resulting in a small increase on 2D human pose estimation accuracy.

In [35] a very simple CNN architecture based on convolutional and deconvolutional layers was used, to show that the crucial part of 2D human pose estimation is obtaining high-resolution feature maps. Similarly, a CNN architecture for 2D human pose estimation was proposed in [36], which was specifically designed to maintain high-resolution feature maps through the overall procedure. This was accomplished by connecting multiple multi-resolution subnetworks in parallel and conducting repeated multi-scale fusions, through exchanging information across these parallel subnetworks.

Certain recent approaches aimed to achieve both increased 2D human pose estimation accuracy and fast inference, similarly to the method proposed in this paper, but through different means. For instance, a Parallel Pyramid Network (PPNet) was proposed in [48], which utilized deep+wide and shallow+narrow subnetworks in a parallel configuration. The deep subnetworks were employed to obtain semantic information about body joints by processing low-resolution inputs, while the shallow ones were used to encode spatial information using high-resolution inputs. Then, the final estimated body joint heatmaps are obtained by fusing the outputs of all subnetworks. With a similar goal in mind, the process of maintaining high-resolution feature maps proposed in [36] was adjusted in [49] so as to increase inference speed. This was accomplished by introducing a conditional channel weighting module in the employed shuffle blocks [50] to efficiently facilitate information exchange between channels and features of different resolution, resulting in a lightweight network architecture that combines good 2D human pose estimation performance with low complexity. Moreover, a single-branch network architecture for real-time multi-person human pose estimation on mobile platforms was proposed in [51]. By incorporating a Fusion Deconv Head and Large Kernel Conv layers in the network architecture, the network was able to decrease latency while maintaining increased human pose estimation performance. Finally, [52] aimed to achieve efficient 2D human pose estimation by exploiting depth data instead of RGB images and by designing lightweight CNN architectures that achieved good accuracy-speed trade-off. In order to further increase the 2D human pose estimation accuracy of the designed lightweight CNNs, supplementary domain adaptation and knowledge distillation algorithms were also explored.

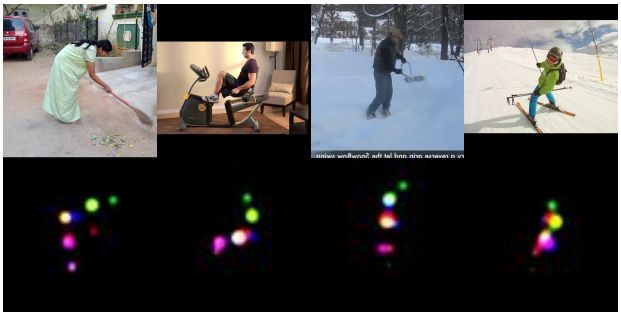


Fig. 1. Examples of input RGB images (first row), along with the corresponding constructed human body structure images $\mathbf{S}$ (second row).

Differently from these approaches, the proposed method aims to predict accurate body joint heatmaps by jointly training a main and an auxiliary neural head (for 2D body joint heatmap regression and for global human body structure modelling, respectively) under a multitask setting. This configuration efficiently extracts suitable features for 2D human pose estimation (as shown in Section IV), allowing us to use a lighter/less complex/faster CNN architecture, if we choose to do so, without sacrificing accuracy. Compared to existing methods, the main novelties are:

- an innovative multihead neural architecture, built on top of a common feature extraction backbone that feeds the two parallel heads, while a novel configuration of skip synapses conjoins them.
- an I2I-based loss function for global human body structure modelling presented for the first time.

The advantage of the proposed network architecture over previous approaches that also aim to simultaneously achieve increased accuracy and fast inference is that it explicitly partitions the subtasks implicitly involved in 2D human pose estimation among different neural heads. Thus, they can all be executed more efficiently, while the integration of their outcomes before obtaining the final predictions from the main head ensures high accuracy.

III. Proposed Architecture

In order to capture rich semantic and spatial information for 2D human pose estimation, a novel CNN architecture is

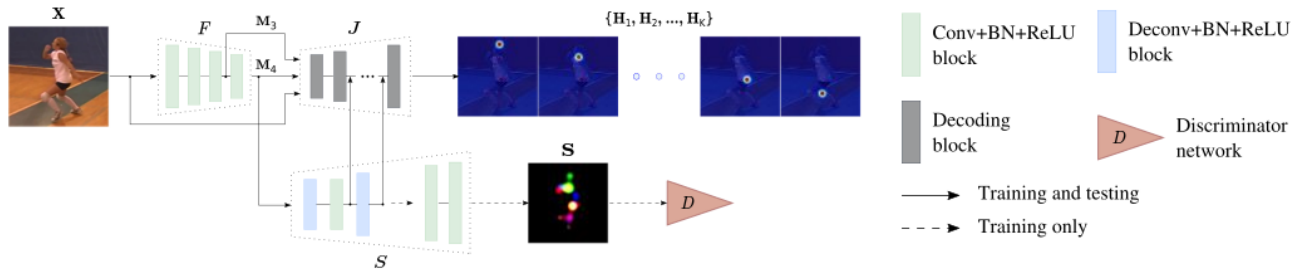


Fig. 2. The proposed unified neural architecture during training and testing. The human body structure modelling neural head S and the 2D body joint regression neural head J act on feature maps extracted by the shared CNN backbone F to predict the corresponding human body structure image and 2D body joint heatmaps, respectively. In addition, S is able to enrich the semantic features extracted by J through skip synapses.

proposed that consists of two parallel neural heads attached on a shared feature extraction backbone. The auxiliary neural head is trained under the GAN framework to encode the global human body structure and provide this information to the main 2D body joint regression neural head. The latter's purpose is to accurately locate a predefined set of body joints on the 2D image. Information flow is facilitated by placing skip synapses between neurons of the auxiliary and the main body joint regression subnetworks, conjoining the two parallel heads.

This is not an ad hoc design, but rather a targeted architecture. It explicitly partitions the subtasks that are implicitly involved in 2D body joint estimation among the two heads, while still keeping the main head aware of all relevant information during inference. The use of I2I for the auxiliary head is instrumental for ensuring that the latter's assigned subtask will mainly be to model spatial global body structure.

A. Human Body Structure modelling

The auxiliary neural head aims to encode the *coarse* spatial human body structure and provide this information to the main body joint regression neural head; the latter is the one responsible for the final, localized 2D human pose predictions. The underlying idea is that the main 2D body joint regression head is thus exempt from encoding global spatial human body structure information. Therefore, it can better focus on accurately identifying and precisely localizing each body joint on the 2D image. Human body structure modelling is essentially a reformulation of the desired task, complementary to the traditional *fine-grained* (i.e., at the body joint level) skeleton estimation objective.

Let $\mathbf{X} \in \mathbb{R}^{M \times N \times 3}$ be a cropped input RGB image (of height M and width N in pixels) of an already localized person, S be the human body structure modelling subnetwork, F be the shared CNN backbone acting as a feature extractor and $\mathbf{M}_l = F(\mathbf{X})$, $l = 1, 2, 3, 4$, be the extracted image feature maps of resolution $\frac{M}{2^l} \times \frac{N}{2^l}$, respectively derived from the l-th layer block of F. *We also define the coarse human body structure to be represented by an RGB image* $\mathbf{S} \in \mathbb{R}^{M \times N \times 3}$*, which is constructed using the ground-truth locations of body joints on the 2D image*. $\mathbf{S}$ is carefully selected to ensure that the auxiliary neural head, trained for GAN-based I2I (where the main training signal is provided indirectly by the Discriminator, i.e., a binary classifier of "reals" and "fakes"), captures information that is indeed beneficial for the main 2D human pose estimation task, in order to later provide it to the main neural head. Thus, $\mathbf{S}$ is manually constructed so as to concurrently satisfy two constraints: a) $\mathbf{S}$ should represent the global spatial human body structure and also contain semantic information identical to the target of the main 2D body joint regression neural head, ensuring that the encoded information will be beneficial for the latter one, and b) $\mathbf{S}$ should be discriminant enough to facilitate training of the GAN's Discriminator, thus providing improved training signal for S. Therefore, $\mathbf{S}$ is constructed by centering a 2D Gaussian function at the ground-truth location of each body joint, while assigning a specific RGB value to it, resulting in a coarse representation of the human body with each joint marked using a different color. The choice of having each body joint uniquely identified by a respective color was made to ensure proper modelling of the global human body structure, as well as to distinguish between ambiguous body joint representations in $\mathbf{S}$ (e.g., left and right wrists), in order to ensure that it also contains identical information to the target of the main 2D body joint regression neural head. The color assignment conventions can be arbitrary, as long as different colors are used to represent different body joints. Examples of constructed ground-truth coarse human body structure images $\mathbf{S}$, along with the corresponding input RGB images, are depicted in Fig. 1.

S is trained under the GAN-based I2I learning framework [38], which relies on the interaction between a Generator and a Discriminator. The reason behind selecting the GAN-based approach for training this subnetwork is their well-known resistance to overfitting [37]; we expect this to be imparted to the overall architecture. Since S is tasked to predict human body structure images $\mathbf{S}$ from the extracted features, it is designed as a decoding CNN [53]. This can be seen as a GAN Generator, acting on the feature maps $\mathbf{M}_4$ extracted by the last block of F, in order to produce human body structure images $S(\mathbf{M}_4)$ that fit the respective input images. Subsequently, a predicted human body structure image obtained from S and the corresponding input image are jointly fed to the Discriminator D, which processes the pair and decides whether the human body structure image is a "fake" one produced by the Generator, or a ground-truth one. As a result of this process, S learns to output realistic human

body structure images that match the human body depicted in the input image. As is typically the case with GANs, S and D are trained via the minimax game, $min_S max_D \mathcal{L}_{GAN}(S, D)$, where the objective function $\mathcal{L}_{GAN}(S, D)$ is given by [38]:

$$\mathcal{L}_{GAN}(S, D) = \mathbb{E}_{(\mathbf{X},\mathbf{S})}[\log D(\mathbf{X}, \mathbf{S})] + \mathbb{E}_{\mathbf{X}}[\log(1 - D(\mathbf{X}, S(\mathbf{M}_4(\mathbf{X}))))], \quad (1)$$

where $\mathbf{M}_4(\mathbf{X})$ denotes the final output of the last block of F, as a function of raw input $\mathbf{X}$. It is common in GAN-based I2I [38], [54] to also explicitly push the Generator towards producing outputs close to the target images, using a supervised objective complementary to the adversarial one. To this end, an additional $\mathcal{L}1$ distance-based similarity loss function is employed for training the human body structure image modelling network:

$$\mathcal{L}_{sim}(S) = \mathbb{E}_{(\mathbf{X},\mathbf{S})}[\|\mathbf{S} - S(\mathbf{M}_4(\mathbf{X}))\|_1]. \quad (2)$$

In order to further strengthen S and help it handle difficult cases where humans appear in abnormal postures or with occluded body joints, the Discriminator D is trained with an additional task. That is, given the concatenation of the input RGB image and the predicted human body structure RGB image, D is additionally tasked to predict which body joints are visible in a multi-label classification manner. As a result, the human body structure modelling subnetwork S is forced to produce more accurate human body structure images that allow body joint visibility estimation by the Discriminator. Directly training S on this task is avoided, in order to prevent it from losing focus from the human body structure modelling task. The final objective function of the human body structure modelling neural head in the proposed architecture is therefore defined as follows:

$$\mathcal{L}_S = \min_S \max_D \mathcal{L}_{GAN}(S, D) + \gamma_1 \mathcal{L}_{sim}(S) + \gamma_2 \mathcal{L}_{vis}(D), \quad (3)$$

where $\mathcal{L}_{vis}(D)$ is a cross-entropy loss function used to train the Discriminator for predicting body joint visibility and γ_1, γ_2 are scaling hyperparameters.

B. Backbone and Body Joint Regression

The CNN backbone along with the 2D body joint regression neural head constitute the main neural pathway of the proposed network architecture, from which the final 2D human pose estimations are obtained. *The 2D human pose is defined as the 2D pixel locations of a predefined set of body joints in the input image, which are regressed in the form of 2D body joint heatmaps* $\{\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_K\}$. K is the number of body joints in the set, while $\mathbf{H}_k \in \mathbb{R}^{M \times N}$ is the 2D body joint heatmap that encodes the 2D location of the k-th body joint. Each heatmap is constructed in a fashion similar to human body structure image $\mathbf{S}$, only this time each $\mathbf{H}_k$ is a grayscale image representing a single body joint.

The 2D body joint regression head J is fed the downsampled input image along with the feature maps $\mathbf{M}_3$ and $\mathbf{M}_4$, extracted by the CNN backbone. It outputs the 2D body joint heatmaps, from which the final 2D body joint locations can be obtained by choosing the location of the maximum value in the corresponding heatmap. Therefore, J is also designed as a decoding CNN. Also, since the target of J in this case has a form very similar to the one used in semantic image segmentation [55], almost any such CNN architecture can be adopted for the baseline 2D body joint regression neural pathway. It can be trained via the following loss function:

$$\mathcal{L}_J = \mathcal{L}_p + \alpha \sum_{i=2}^{3} \mathcal{L}_{a_i}, \quad (4)$$

where $\mathcal{L}_p$ is the principal loss that is used to supervise the entire main neural pathway, while $\mathcal{L}_{a_i}$, $i = 1, 2$, are similar loss terms used for intermediate supervision at stage i. Both $\mathcal{L}_p$ and $\mathcal{L}_{a_i}$ are standard Softmax loss functions. α is a hyperparameter employed to weight the contribution of the intermediate losses in the total loss.

C. Overall Architecture

The 2D body joint regression head J acts both on the downsampled input image and feature maps extracted by the shared CNN backbone F, while the human body structure modelling head S acts only on the features maps $\mathbf{M}_4$ extracted by the last block of F. In the latter case, F plays the role of the encoder and S the role of the decoder in an encoder-decoder network architecture [53], which is typically used for the Generator in GAN-based I2I. In order to augment information flow between the two neural heads, skip synapses are placed between neurons of two intermediate stages of S and J. This allows the global human body structure information that is encoded by S to flow towards J, providing complementary semantic context for 2D human pose estimation.

The overall network is jointly trained for both human body structure modelling and for 2D body joint regression, using the following multitask loss:

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_J + \lambda \mathcal{L}_S, \quad (5)$$

where λ is a hyperparameter meant to tune the contribution of the human body structure modelling loss function to the total loss. The training framework of the proposed unified CNN architecture ensures that:

- the parallel neural head S learns to compute features that capture auxiliary information about the global spatial human body structure and passes it to the main neural pathway via the added skip synapses. Thus it alleviates the latter from this subtask and allows it to focus on identifying and precisely localizing each body joint,
- the inherent resistance of GANs (training of S) to overfitting [37] is imparted to the overall architecture,
- F extracts rich features that are explicitly suitable both for global human body structure modelling (spatial subtask) and for localized 2D body joint regression (a semantic subtask and a spatial subtask),
- F's training is regularized, according to a well-known relevant side-effect of multitask training in general [56], [57].

Importantly, after training has been completed, the entire Discriminator D and the two last convolutional layers of

S can be safely discarded at the inference stage, to reduce computational costs. The overall neural architecture of the proposed method is depicted in Fig. 2.

IV. Experimental Evaluation

The proposed network architecture was evaluated both for pure 2D human pose estimation and for pedestrian intention recognition, thus showcasing its general performance and its applicability as a modular component in computer vision systems for autonomous systems. The latter task was selected due to its significance in ensuring human safety when deploying self-driving cars.

A. Implementation Details

All components of the proposed CNN's implementation utilize either lightweight or not overly complex neural architectures, so as to promote fast execution speed during inference. Thus, ResNet-50 [58] is employed as the CNN backbone F. S consists of three convolutional and two deconvolutional layers, in order to increase the feature map resolution, while maintaining a relatively low number of parameters. D is based on a standard PatchGAN [38] classifier, which was extended by an extra fully connected layer for the joint visibility classification task. J utilizes an architecture similar to the one used in [59], i.e., a state-of-the-art semantic image segmentation CNN with real-time processing capabilities, but without its downsampling network, since feature extraction for both J and S is performed by F. *In principle, the decoding part of any other fast dense image prediction CNN could be alternatively adopted in J as a building block*. Each convolutional and deconvolutional layer of S is followed by a Batch Normalization and a ReLu layer. Finally, for simplicity, all convolutional layers use 3×3 kernels with stride 1, while deconvolutional ones use 3×3 kernels with a stride equal to 2.

The subnetworks F, J, S, D were jointly trained using the proposed multitask loss function (5) for 200 epochs using the Adam optimizer [60]. The learning rate for training F, J, S subnetworks was initialized to 0.01 and is reduced in each epoch using the "poly" learning rate strategy with the power of 0.9, while for D the learning rate is kept constant to 0.00002. Batch size was set to 64. Hyperparameter λ was empirically set to $\lambda = 0.3$, favouring the 2D human pose estimation task over human body structure modelling. α was set to 1 in order to enable full intermediate supervision of the main neural pathway, while γ_1 and γ_2 were set to 1 and 0.1, respectively, in order to scale the corresponding loss terms in Eq. (3) accordingly and ensure smooth training of the auxiliary neural head S. In addition, similarly to previous methods [35], [36], the backbone was pretrained on the ImageNet classification task [61] in all cases, unless otherwise specified.

B. Evaluation Details

The proposed network architecture was evaluated on the COCO keypoints [62] and the MPII Human Pose [63] datasets. The COCO dataset consists of 250K person instances annotated with 17 body joints. We use the COCO *train2017* set consisting of 118K images, while the corresponding *val2017* and *test-dev2017* sets contain 5K and 20K images, respectively. The MPII dataset contains 40K person samples labeled with 16 body joints, where there are 3K samples for validation and 12K samples for testing. The training data augmentation policy was adopted from [36]. First, using the ground-truth human detection boxes provided with the datasets, the image is cropped and resized to a fixed size, 256×192 or 384×288 for the COCO dataset and 256×256 for the MPII dataset. Then, random rotation ($[-45°, 45°]$), random scaling ($[0.65, 1.35]$), flipping and half body data augmentation [64] is applied online on the extracted image patch. Training on four *GeForce GTX 1080 Ti* GPUs takes approximately one/three days for an input resolution of 256×192/384×288, respectively, in the COCO dataset. It requires approximately half a day for the MPII dataset.

TABLE I
Comparison between the proposed network architecture and different variants of the baselines BiSeNet [59] and Lite-HRNet [49] on COCO [62] *val2017* set and MPII [63] validation set, in terms of Average Precision (AP) and PCKh@0.5, respectively. Lite-HRNet backbones are trained from scratch.

Method	Backbone	COCO *val2017* AP		MPII val PCKh@0.5
		Input Res.		Input Res.
		256×192	384×288	256×256
BiSeNet [59]	ResNet-18	68.4	71.4	87.3
proposed	ResNet-18	**70.2**	**72.5**	**88.2**
BiSeNet [59]	ResNet-50	71.4	71.6	88.1
proposed	ResNet-50	**73.7**	**74.0**	**89.7**
BiSeNet [59]	ResNet-101	72.5	74.7	89.8
proposed	ResNet-101	**74.3**	**75.6**	**90.2**
BiSeNet [59]	ResNet-152	73.4	75.4	90.0
proposed	ResNet-152	**74.7**	**76.2**	**90.5**
Lite-HRNet [49]	Lite-HRNet-18	64.8	67.6	86.1
proposed	Lite-HRNet-18	**65.3**	**70.5**	**87.1**
Lite-HRNet [49]	Lite-HRNet-30	67.2	70.4	87.0
proposed	Lite-HRNet-30	**68.0**	**71.7**	**88.3**

Results are reported both on the *val2017* and the *test-dev2017* set of COCO, as well as on the validation set of MPII, after training on the respective training sets. For evaluating on the MPII test set, both train and validation set images of MPII were utilized for model training, similarly to [35], [36]. The common two-stage top-down evaluation paradigm [35], [36], [43] was followed for both datasets, where each person in the input image is first detected using a person detection algorithm and body joints are subsequently predicted for each detection. In COCO, the person detection algorithm provided by [35], [36] was used both for *val2017* and *test-dev2017* sets to ensure a fair comparison, while in MPII each person location is provided with the dataset.

Finally, in all cases, the final body joint heatmap is computed by averaging the heatmaps of the original and the flipped input images. For COCO, the average precision (AP) and average recall (AR)[1] is reported, similarly to [36], while

[1] https://cocodataset.org/#keypoints-eval

TABLE II

EVALUATION RESULTS ON THE COCO [62] *val2017* SET. BEST RESULT IN EACH CATEGORY IS IN BOLD. FPS-D AND FPS-M DENOTE FRAMES PER SECOND (INFERENCE SPEED) USING A GEFORCE GTX 1080 TI GPU AND A NVIDIA JETSON XAVIER COMPUTING BOARD, RESPECTIVELY. OHKM MEANS ONLINE HARD KEYPOINTS MINING [43]. "PRETRAIN" INDICATES WHETHER THE BACKBONE IS PRETRAINED ON THE IMAGENET [61] CLASSIFICATION TASK.

Method	Backbone	Pretrain	Input Res.	FPS-D	FPS-M	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
small models											
8-stage Hourglass [34]	8-stage Hourglass	N	256×192	—	—	66.9	—	—	—	—	—
CPN [43]	ResNet-50	Y	256×192	—	—	68.6	—	—	—	—	—
CPN + OHKM [43]	ResNet-50	Y	256×192	—	—	69.4	—	—	—	—	—
Lite-HRNet [49]	Lite-HRNet-18	N	256×192	28.9	15.5	64.8	86.7	73.0	62.1	70.5	71.2
Lite-HRNet [49]	Lite-HRNet-30	N	256×192	18.1	11.4	67.2	88.0	75.0	64.3	73.1	73.3
BiSeNet [59]	ResNet-50	Y	256×192	**45.4**	**22.8**	71.4	89.5	78.6	67.4	78.5	76.8
SB [35]	ResNet-50	Y	256×192	41.7	20.3	70.4	88.6	78.3	67.1	77.2	76.3
proposed	ResNet-50	Y	256×192	41.9	20.4	**73.7**	**90.2**	**80.7**	**69.8**	**80.7**	**79.0**
CPN [43]	ResNet-50	Y	384×288	—	—	70.6	—	—	—	—	—
CPN + OHKM [43]	ResNet-50	Y	384×288	—	—	71.6	—	—	—	—	—
Lite-HRNet [49]	Lite-HRNet-18	N	384×288	22.2	9.8	67.6	87.8	75.0	64.5	73.7	73.7
Lite-HRNet [49]	Lite-HRNet-30	N	384×288	13.6	7.6	70.4	88.7	77.7	67.5	76.3	76.2
BiSeNet [59]	ResNet-50	Y	384×288	**42.3**	**17.1**	71.6	89.5	79.4	67.9	78.4	77.3
SB [35]	ResNet-50	Y	384×288	31.3	11.9	72.2	89.3	78.9	68.1	79.7	77.6
proposed	ResNet-50	Y	384×288	37.7	13.4	**74.0**	**90.0**	**81.3**	**70.0**	**80.9**	**79.2**
large models											
SB [35]	ResNet-101	Y	256×192	28.2	15.0	71.4	89.3	79.3	68.1	78.1	77.1
BiSeNet [59]	ResNet-101	Y	256×192	**29.3**	**16.3**	72.5	89.9	80.7	68.8	79.5	78.2
proposed	ResNet-101	Y	256×192	27.1	11.9	74.3	90.3	81.4	70.4	**81.4**	79.6
SB [35]	ResNet-152	Y	256×192	21.3	10.1	72.0	89.3	79.8	68.7	78.9	77.8
BiSeNet [59]	ResNet-152	Y	256×192	21.2	10.1	73.4	90.4	81.5	69.7	80.4	79.0
HRNet-W32 [36]	HRNet-W32	Y	256×192	11.4	5.5	74.4	**90.5**	81.9	70.8	81.0	79.8
proposed	ResNet-152	Y	256×192	20.3	9.9	**74.7**	90.4	**82.1**	**71.0**	**81.4**	**79.9**
SB [35]	ResNet-101	Y	384×288	23.5	10.3	73.6	89.6	80.3	69.9	81.1	79.1
BiSeNet [59]	ResNet-101	Y	384×288	**27.6**	**13.8**	74.7	90.1	81.8	70.8	81.8	79.8
proposed	ResNet-101	Y	384×288	25.1	10.4	75.6	90.4	82.2	71.6	82.7	80.6
SB [35]	ResNet-152	Y	384×288	18.9	8.7	74.3	89.6	81.1	70.5	79.7	79.7
BiSeNet [59]	ResNet-152	Y	384×288	20.5	9.4	75.4	90.6	82.4	71.6	82.3	80.4
HRNet-W32 [36]	HRNet-W32	Y	384×288	8.8	5.3	75.8	90.6	82.7	71.9	82.8	81.0
proposed	ResNet-152	Y	384×288	19.0	8.7	**76.2**	**90.8**	**82.9**	**72.2**	**83.3**	**81.1**

for MPII, the head-normalized probability of correct keypoint (PCKh@0.5) metric is used as a measure of 2D human pose estimation performance. Inference speed is measured in Frames Per Second (FPS), including the flipping and heatmap averaging calculations that are necessary to obtain the final estimations. Note that in the context of this work and in contrast to previous methods [35], [36], inference speed in FPS, measured in the same machine, is reported as a fair model complexity measurement, instead of the number of trainable model parameters and/or the number of flops. This is due to the fact that the reported numbers of model parameters and flops typically involve only convolutional and linear layers of the model, ignoring any extra layers and/or calculations required by other operations (e.g., resizing, addition, multiplication, flipping, etc.).

First, a comparison between the proposed network architecture and the baseline *BiSeNet* [59] for both COCO *val2017* and MPII validation sets, as well as for backbones of different complexity (ResNet-18, ResNet-50, ResNet-101, ResNet-152) is conducted and presented in Table I, in order to evaluate the effectiveness of the proposed architecture in terms of 2D human pose estimation accuracy. This is because the differences between the two architectures are only the auxiliary neural head and the additional skip synapses of the proposed CNN. The comparison results show that the latter one outperforms *BiSeNet* for all backbone variants in both datasets. The proposed architecture increased AP score on the COCO *val2017* set and PCKh@0.5 score on the MPII validation set by a margin up to 2.4 and 1.6 (in the ResNet-50 backbone case), respectively, proving that the information encoded by the auxiliary neural head and passed to the main neural head through the skip synapses, which conjoin the two heads, enabled the main 2D body joint regression neural pathway to predict more accurate 2D human poses. Besides *BiSeNet*, the proposed method is also compared against the *Lite-HRNet* [49] baseline. That is, the stem network of Lite-HRNet is employed as the backbone F of the proposed network architecture, while the three remaining stages of Lite-HRNet were utilized in J. S utilizes the same network architecture as in the previous case. As in [49], the unified

TABLE III
EVALUATION RESULTS ON THE COCO [62] *test-dev2017* SET. THE BEST METHOD IS IN BOLD. † DENOTES THAT THE MODEL IS TRAINED FROM SCRATCH.

Method	Backbone	Input Res.	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
small models								
OpenPose [21]	—	—	61.8	84.9	67.5	57.1	68.2	66.5
Associative Embedding [65]	—	—	65.5	86.8	72.3	60.6	72.6	70.2
PersonLab [66]	—	—	68.7	89.0	75.4	64.1	75.5	75.4
MultiPoseNet [67]	—	—	69.6	86.3	76.6	65.0	76.3	73.5
Mask-RCNN [68]	ResNet-50-FPN	—	63.1	87.3	68.7	57.8	71.4	—
Lite-HRNet [49]†	Lite-HRNet-18	384×288	66.9	89.4	74.4	64.0	72.2	72.6
Lite-HRNet [49]†	Lite-HRNet-30	384×288	69.7	90.7	77.5	66.9	75.0	75.4
SB [35]	ResNet-50	384×288	71.5	91.1	78.7	67.8	78.0	76.9
RMPE [69]	PyraNet [70]	320×256	72.3	89.2	79.1	68.0	78.6	—
proposed	ResNet-50	384×288	**73.3**	**92.1**	**81.3**	**70.0**	**79.0**	**78.6**
large models								
G-RMI [71]	ResNet-101	353×257	64.9	85.5	71.3	62.3	70.0	69.7
Integral Pose Regression [72]	ResNet-101	256×256	67.8	88.2	74.8	63.9	74.0	—
G-RMI + extra data [71]	ResNet-101	353×257	68.5	87.1	75.5	65.8	73.3	73.3
CPN [43]	ResNet-Inception	384×288	72.1	91.4	80.0	68.7	77.2	78.5
proposed	ResNet-101	384×288	74.9	92.3	82.5	71.2	81.1	80.1
SB [35]	ResNet-152	384×288	73.7	91.9	81.1	70.3	80.0	79.0
BiSeNet [59]	ResNet-152	384×288	74.5	92.4	82.6	71.1	80.6	79.8
HRNet-W32 [36]	HRNet-W32	384×288	74.9	92.5	82.8	71.3	80.9	80.1
OKS-Net [47]	HRNet-W32	384×288	75.2	**92.7**	83.0	71.7	81.2	80.4
proposed	ResNet-152	384×288	**75.6**	92.5	**83.3**	**71.9**	**81.7**	**80.6**

network architecture is trained from scratch in this case. The comparisons between the Lite-HRNet-based implementation of the proposed method and the baseline *Lite-HRNet* model for both Lite-HRNet-18, Lite-HRNet-30 variants and for both COCO *val2017*, MPII validation sets are presented in Table I. They show that the proposed method increased 2D human pose estimation accuracy in all cases, proving the efficiency of the proposed method for baselines/backbones of different network architectures as well. Overall, the results reported in Table I indicate that the proposed method is able to increase the 2D human pose estimation accuracy of different baseline network architectures, with different backbones of varying complexity. Moreover, it can also be seen that small/low-complexity models enjoy the most benefits in 2D human pose estimation accuracy. This is the property that renders the proposed method most suitable for embedded AI scenarios.

Comparisons between the proposed and competing methods [34]–[36], [43], [49], [59] in terms of 2D human pose estimation accuracy and inference speed on the COCO *val2017* set are depicted in Table II, for input image resolutions of 256×192 and 384×288 pixels. Inference speed in FPS is measured for all available competing methods using both a high-end desktop PC equipped with a *GeForce GTX 1080 Ti* GPU and an *Nvidia Jetson Xavier* embedded AI computing board. When relying on a ResNet-50 as the feature extraction CNN backbone, the proposed architecture is able to outperform all competing methods that use feature extraction backbones of similar complexity for both input resolutions, while maintaining increased inference speed. It achieves higher accuracy than the best performing competing methods *BiSeNet* and *SB* for low (256×192) and high (384×288) input resolution, respectively, increasing AP score by a margin up to 2.3 in the first case and 1.8 in the latter case. When compared to *SB*, the proposed architecture is faster by 6.4 FPS and 2.5 FPS in the high input resolution case when using a desktop GPU and an embedded computing board, respectively, while it runs at the same speed in the low input resolution case. The AP score increase of the proposed architecture in these two cases is 1.8 and 3.3, respectively. Similarly, the comparison between the proposed architecture and *BiSeNet* shows that the proposed architecture consistently increases AP score, while being slower only by 4.6 FPS in the worst case scenario (high input resolution), where it increased AP score by 2.4.

Moreover, the ResNet-50 variant of the proposed architecture is able to achieve highly competitive 2D human pose estimation performance when compared to methods that use larger and more complex/slower feature extraction CNN backbones (ResNet-101, ResNet-152, HRNet-W32). Notably, it demonstrates increased AP score when compared to both deep variants of *SB* in the low input resolution case, while running faster up to 20.6 FPS and 10.3 FPS on a desktop GPU and an embedded computing board, respectively. In comparison to *HRNet-W32*, our small variant is less accurate but up to 4x and 3x faster at the inference stage for high-end GPU and embedded execution, respectively. However, as it can also be seen in Table II, by employing the ResNet-152 CNN for F in the proposed network architecture, the latter manages to outperform all competing methods for both input resolutions, while also being 1.5x-2x faster than the second-best method (*HRNet-W32*). Overall, Table II shows that the proposed network architecture offers the best accuracy-speed ratio among all competing methods.

TABLE IV
EVALUATION RESULTS ON THE MPII [63] VALIDATION SET. THE BEST METHOD IS IN BOLD. † DENOTES THAT THE CORRESPONDING MODEL IS TRAINED FROM SCRATCH. INPUT RESOLUTION IS 256×256 IN ALL CASES.

Method	Backbone	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
small models									
Lite-HRNet [49]†	Lite-HRNet-18	—	—	—	—	—	—	—	86.1
Lite-HRNet [49]†	Lite-HRNet-30	—	—	—	—	—	—	—	87.0
Integral Pose Regression [72]	ResNet-50	—	—	—	—	—	—	—	87.3
BiSeNet [59]	ResNet-50	96.5	94.8	88.2	83.2	88.1	83.1	79.2	88.1
SB [35]	ResNet-50	96.4	95.3	89.0	83.2	88.4	84.0	79.6	88.5
proposed	ResNet-50	**96.8**	**95.5**	**89.6**	**85.6**	**88.9**	**85.9**	**82.1**	**89.7**
large models									
SB [35]	ResNet-101	96.9	95.9	89.5	84.4	88.4	84.5	80.7	89.1
BiSeNet [59]	ResNet-101	97.0	95.8	90.2	85.2	89.1	85.9	81.9	89.8
proposed	ResNet-101	96.8	95.9	90.3	85.9	89.6	86.4	83.1	90.2
SB [35]	ResNet-152	97.0	95.9	90.0	85.0	89.2	85.3	81.3	89.6
BiSeNet [59]	ResNet-152	97.0	95.8	90.5	85.7	88.8	86.2	82.3	90.0
HRNet-W32 [36]	HRNet-W32	**97.1**	95.9	90.3	**86.4**	89.1	**87.1**	**83.3**	90.3
proposed	ResNet-152	**97.1**	**96.4**	**90.8**	**86.4**	**90.0**	86.8	82.9	**90.5**

TABLE V
EVALUATION RESULTS ON THE MPII [63] TEST SET. THE BEST METHOD IS IN BOLD. INPUT RESOLUTION IS 256×256 IN ALL CASES.

Method	Backbone	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
small models									
CPMs [73]	—	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
SB [35]	ResNet-50	98.2	**96.4**	91.0	86.0	90.4	86.3	82.3	90.5
8-stage Hourglass [34]	—	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Integral Pose Regression [72]	ResNet-50	—	—	—	—	—	—	—	91.0
GLN [74]	8-stage hourglass	98.1	96.2	91.2	87.2	89.8	87.4	84.1	91.0
SA-GCN [75]	ResNet-50	97.1	96.1	91.5	86.8	**90.7**	87.6	84.3	91.1
proposed	ResNet-50	**98.3**	96.3	**91.7**	**87.6**	90.3	**88.0**	**84.5**	**91.3**
large models									
DeeperCut [76]	ResNet-152	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Part Heatmap Regression [77]	ResNet-152	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
DU-Net [78]	U-Net	97.4	96.4	92.1	87.7	90.2	87.7	84.3	91.2
BiSeNet [59]	ResNet-152	98.2	96.4	91.7	87.5	90.5	87.5	83.5	91.2
SB [35]	ResNet-152	98.5	96.6	91.9	87.6	91.1	88.1	84.1	91.5
proposed	ResNet-101	98.4	96.6	92.1	88.1	90.7	88.5	84.4	91.6
proposed	ResNet-152	98.4	96.7	92.4	88.7	91.4	**89.1**	85.2	92.1
HRNet-W32 [36]	HRNet-W32	**98.6**	**96.9**	**92.8**	**89.0**	**91.5**	89.0	**85.7**	**92.3**

The 2D human pose estimation accuracy of the proposed architecture is also evaluated on the COCO *test-dev2017* set. Comparisons are presented in Table III and indicate similar behaviour. The ResNet-50 variant of the proposed architecture outperforms all competing methods which use CNN backbones of similar complexity, while its accuracy in 2D human pose estimation remains highly competitive against methods with much more complex/slow backbones. Finally, the ResNet-152 variant of the proposed method again yields the best 2D human pose estimation accuracy among all competing methods.

Comparisons against competitors [34]–[36], [49], [59], [72]–[78] in the MPII validation and test sets are reported in Tables IV and V, respectively. An input resolution of 256×256 pixels is used in all cases for a fair comparison. As it can be seen, the ResNet-50 variant of the proposed architecture yielded increased 2D human pose estimation accuracy compared to competitors of similar complexity, while it lags only by 1.0 PCKh@0.5 score against the best-performing method (*HRNet-W32* on MPII test set) that uses the more complex and significantly slower (as shown in Table II) HRNet-W32 backbone. Most remarkably, it manages to outperform (*DeeperCut*, *Part Heatmap Regression*, *DU-Net*) or achieve highly-competitive accuracy (*SB*) to approaches utilizing very complex feature extraction backbones (U-Net, ResNet-152), while it only employs the much lighter and faster ResNet-50 architecture. Finally, our ResNet-152 variant outperforms all competing methods on the MPII validation set and is only behind *HRNet-W32* by 0.2 PCKh@0.5 score on the MPII test set. However, since it is 1.5x-2x faster than *HRNet-W32* as

Fig. 3. Proposed method predictions for random test images. First row: test images from COCO and MPII datasets. Second row: unseen UAV-captured images. The first/last two columns show the output of our model trained on COCO/MPII train sets, respectively.

presented in Table II, the proposed architecture offers the best accuracy-speed ratio in this case as well.

Quantitative comparisons in Tables I - V show the efficiency of the proposed approach, which demonstrates the best accuracy-speed ratio. A complementary qualitative evaluation can be seen in Fig. 3, which depicts random test images from the MPII test set and the COCO *test-dev2017* set, as well as random, previously unseen UAV-captured images for a real-world human-UAV visual interaction scenario. The proposed method is able to yield accurate predictions, even when humans appear in abnormal poses and different scenes. Its performance under challenging conditions (e.g., cluttered scenes, occlusion) is qualitatively evaluated by inspecting the final outputs of both the auxiliary and the main neural heads of the proposed method (global human body structure images $\mathbf{S}$ and 2D skeletons, respectively) when using inputs that depict the person of interest in such conditions. The results presented in Fig. 4 show that the proposed method manages to estimate accurate 2D human poses (rightmost image of each triplet), despite the fact that the person of interest appears in complicated scenes (top-left, top-right) or under occlusion (top-right, bottom-left). In addition, the bottom-right triplet once again demonstrates the ability of the proposed method to handle weird postures. Finally, as it can be seen in the middle image of each triplet, the auxiliary neural head of the proposed method successfully predicts human body structure images in all cases.

Finally, the proposed architecture is utilized as the 2D pedestrian skeleton extraction stage of a common two-step pedestrian intention recognition approach, in order to prove its effectiveness and generalization ability. To this end, the ResNet-50 variant of the proposed architecture was first pre-trained on COCO (256×192 input resolution) and subsequently used to extract the 2D skeletons of all pedestrians in the JAAD dataset [79], which is commonly used for the cross/no-cross prediction problem. Following [27], a simple LSTM classifier with a hidden dimension equal to 64 was adopted, followed by a fully connected layer. The pedestrian intention recognition results in Table VI demonstrate that when

TABLE VI
EVALUATION OF THE PROPOSED NETWORK ARCHITECTURE AS A 2D PEDESTRIAN POSE ESTIMATION COMPONENT OF A TWO-STEP PEDESTRIAN INTENTION RECOGNITION APPROACH USING THE JAAD DATASET [79] AND A SINGLE NVIDIA GTX 1080 TI GPU. † DENOTES THAT AN NVIDIA TITAN X GPU WAS USED INSTEAD.

Method	Runtime (ms)	Accuracy (%)
RMPE + LSTM [27]	50+1.6	78.0
Res-EnDec [80]	116.4	81.0
ST-DenseNet [81]†	50.0	84.8
autoencoder + Prediction [82]†	102.5	86.7
proposed + LSTM	**22.2+0.9**	**87.0**

using the proposed method for pedestrian skeleton extraction, a simple LSTM classifier outperforms all directly comparable competing methods. Apart from pedestrian intention recognition accuracy, Table VI also shows speed comparisons between all competing methods by considering a real-world pedestrian intention recognition scenario: if d denotes the length of the sequence that is required to predict a pedestrian intention label (cross/no-cross), a prediction is made for each new video frame that becomes available, while all previous $d-1$ video frames are assumed to have already been processed and stored in a buffer. The reported runtime for the proposed method and [27] includes both the estimation of the 2D pedestrian skeleton from a new video frame and the prediction of the pedestrian intention label from the full 2D skeleton sequence. For methods [80]–[82] only pedestrian intention label prediction is considered, since they act directly on a video frame sequence of length d. Also, all reported runtimes have been measured using high-end desktop GPUs in order to ensure a fair comparison, since code for [81], [82] is not publicly available and thus their runtime is directly cited from the corresponding sources. For similar reasons, d was set to 16. The comparison shows that the proposed method is at least 2x faster than all competing methods, thus experimentally verifying that it offers the best accuracy-speed ratio in the pedestrian intention recognition task as well. The average time required by the proposed network architecture for extracting a 2D pedestrian skeleton was measured at 22.2 ms, while the processing of the 2D skeleton sequence by the employed LSTM classifier requires only 0.9 ms. Given these results, the proposed method seems especially suitable for embedded applications such as self-driving cars.

C. Ablation Study

In order to show the importance of each component of the proposed method, detailed ablation studies were performed. In all cases, the ResNet-50-based variant of the proposed network architecture was evaluated on the COCO *val2017* set for input resolution of 256×192.

First, an ablation study on the different building blocks of the proposed CNN architecture is presented in Table VII, in order to demonstrate their effect on overall 2D human pose estimation accuracy. The first column of Table VII lists the components of the presented network architecture that are utilized in each case, while the second column indicates the

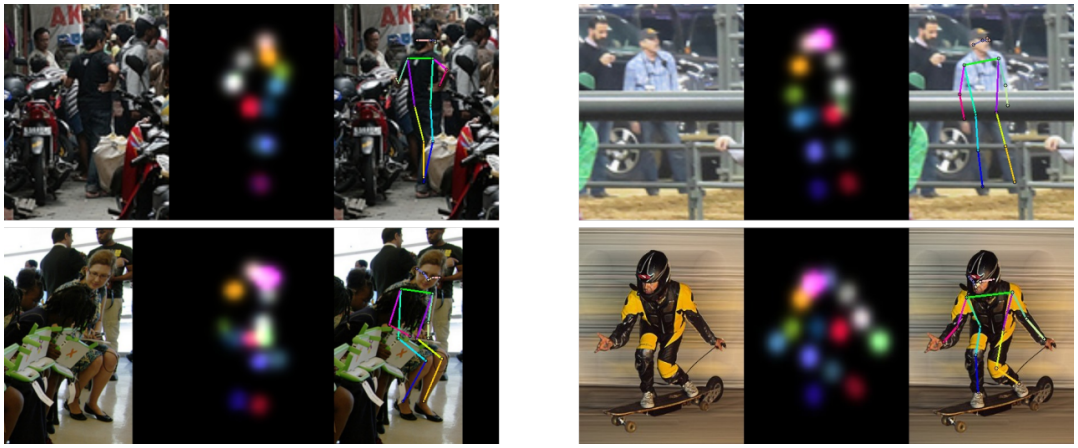


Fig. 4. Qualitative evaluation of the ResNet-50 variant of the proposed method on challenging images from the COCO [62] *val2017* set, where the person of interest appears in complicated scenes (top-left, top-right), under occlusion (top-right, bottom-left), or under weird postures (bottom-right). In each triplet, the left image shows the input image, while the middle and right images show the output of the auxiliary neural head and the final 2D pose estimation, respectively. In all cases the proposed method manages to estimate accurate 2D poses, while the auxiliary neural head is also able to successfully predict global human body structure images according to its objective.

TABLE VII
ABLATION STUDY ON THE PROPOSED NETWORK ARCHITECTURE USING THE COCO [62] *val2017* SET. THE FIRST COLUMN SHOWS THE COMPONENTS OF THE PRESENTED NETWORK ARCHITECTURE THAT ARE UTILIZED IN EACH CASE, WHILE THE SECOND COLUMN INDICATES THE MODIFICATION APPLIED TO THE PROPOSED METHOD IN THE CORRESPONDING EXPERIMENT.

	Modification	AP
CNN (BiSeNet [59])	baseline	71.4
CNN + S + D	without skip synapses	72.1
CNN + S + syn	without Discriminator	73.1
CNN + S + syn	person segmentation as auxiliary task	73.3
CNN + S + syn + D	input image as S target	73.0
CNN + S + syn + D	2D body joint heatmaps as S target	73.1
CNN + S + syn + D (*proposed*)	—	**73.7**

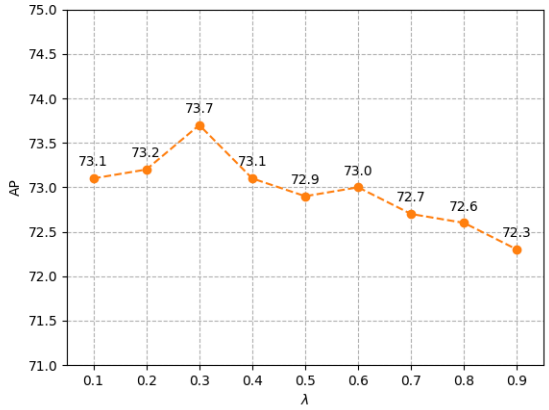


Fig. 5. Ablation study on hyperparameter λ, used in the overall multitask loss function (Eq. (5)) to balance the two tasks during training.

modification applied to the proposed method. By completely removing the skip synapses that conjoin the two parallel neural heads J and S (CNN + S + D) AP score dropped to 72.1, since information exchange between the auxiliary human body structure modelling and main 2D body joint regression heads S and J is not possible during training and inference. Subsequently, an experiment where the Discriminator network D of the proposed method is absent during training was conducted to demonstrate the effectiveness of the GAN training framework in the human body structure modelling task. Omitting the Discriminator during training (i.e., transforming the GAN training objective to a typical unsupervised one) led to decreased performance compared to the proposed method (AP score dropped by 0.6), since the utilization of GANs enables S to model the global human body structure more efficiently and introduces additional regularization to the overall model, due to their overfitting-resistant nature.

The importance of selecting the RGB image representation of the human body structure $\mathbf{S}$ as the target of the auxiliary neural head S was also evaluated by utilizing two alternative targets for S in its place. In the first case, the input image $\mathbf{X}$ was used as target, essentially tasking S with an image reconstruction objective, while in the second case, S is tasked to predict a set of 2D body joint heatmaps $\{\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_K\}$ (similar to J). As it can be seen in the fifth and sixth rows of Table VII, AP score dropped in both cases, proving that the objective selected for S by the proposed method is more effective in increasing 2D human pose estimation performance compared to the two alternatives. This can be possibly explained by the fact that $\mathbf{S}$ was manually designed to be discriminant enough for the GAN training framework, while simultaneously containing semantic information identical to the target of J (2D body joint heatmaps). Thus, it enables the Discriminator network of the GAN framework to introduce a strong supervision signal during training, helping S provide rich information to the main 2D body joint regression neural head. For completeness, an experiment where S was tasked to perform person instance segmentation was also conducted (fourth row of Table VII). Despite the fact that the 2D human

pose estimation accuracy in this case is slightly increased compared to the previous two alternative approaches, it is again outperformed by the proposed method. Moreover, in contrast to the global human body structure modelling task and the two other approaches presented in the fifth and sixth row of Table VII, training S for person instance segmentation requires additional, pixel-level annotated person instance segmentation maps, which are very hard to obtain (costly, time-consuming).

Finally, experiments were also conducted to evaluate the importance of the hyperparameter λ in (5), which is used to balance focus between the 2D body joints regression and the human body structure modelling tasks during training. Results are presented in Fig. 5, using a step of 0.1 for λ. As it can be seen, the best AP score was achieved for $\lambda = 0.3$ (73.7), with $\lambda = 0.2$ (73.2) to follow. $\lambda = 0.1$ and $\lambda = 0.4$ yielded a slightly decreased AP score (73.1), while for $\lambda \geq 0.5$ AP score dropped below 73.0. Getting increased performance for $\lambda < 0.5$ is expected, since the primary goal of the training stage is to ensure that the model performs as well as possible on the main 2D body joints regression task, while the human body structure modelling task acts as an auxiliary one, assisting the main task.

V. Conclusions

This paper proposed a novel, multihead CNN architecture for fast 2D human pose estimation, suitable for embedded execution in autonomous systems. It consists of two neural heads, i.e., an auxiliary I2I-based global human body structure modelling head S and a main 2D body joint regression head J, on top of a common feature extraction backbone F. The proposed architecture is trained in a unified multitask manner and allows information to flow from S to J through skip synapses, in order to enrich features extracted by J with information about the global body structure encoded by S. The end-result is that the different subtasks implicitly involved in 2D body pose estimation are explicitly partitioned among the different heads, with their outcomes properly integrated before obtaining the final predictions. The increased accuracy allows us to use comparatively lightweight CNN components, resulting in fast execution without sacrificing precision. Evaluation on common datasets showed that the proposed method achieves the best accuracy-speed ratio when compared to the state-of-the-art, rendering it a good candidate for autonomous systems employing embedded AI computational hardware. This result was validated by further evaluation on a pedestrian intention recognition dataset for self-driving cars.

References

[1] C. Ebert and M. Weyrich, "Validation of autonomous systems," *IEEE Software*, vol. 36, no. 5, pp. 15–23, 2019.

[2] M. A. Goddard, Z. G. Davies, S. Guenat, M. J. Ferguson, J. C. Fisher, A. Akanni, T. Ahjokoski, P. M. Anderson, F. Angeoletto, C. Antoniou, *et al.*, "A global horizon scan of the future impacts of robotics and autonomous systems on urban ecosystems," *Nature Ecology & Evolution*, vol. 5, no. 2, pp. 219–230, 2021.

[3] J. Athavale, A. Baldovin, R. Graefe, M. Paulitsch, and R. Rosales, "AI and reliability trends in safety-critical autonomous systems on ground and air," in *Proceedings of the IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, 2020.

[4] I. Karakostas, I. Mademlis, N. Nikolaidis, and I. Pitas, "UAV cinematography constraints imposed by visual target tracking," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2018.

[5] ——, "Shot type feasibility in autonomous UAV cinematography," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[6] ——, "Shot type constraints in UAV cinematography for autonomous target tracking," *Information Sciences*, vol. 506, pp. 273–294, 2020.

[7] 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*, vol. 65, no. 3, pp. 627–635, 2019.

[8] E. Kakaletsis, E. Symeonidis, M. Tzelepi, I. Mademlis, T. A., N. Nikolaidis, and I. Pitas, "Computer vision for autonomous UAV flight safety: An overview and a vision-based safe landing pipeline example," *ACM Computing Surveys*, 2021, accepted.

[9] I. Mademlis, N. Nikolaidis, A. Tefas, I. Pitas, T. Wagner, and A. Messina, "Autonomous unmanned aerial vehicles filming in dynamic unstructured outdoor environments," *IEEE Signal Processing Magazine*, vol. 36, pp. 147–153, 2018.

[10] ——, "Autonomous UAV cinematography: a tutorial and a formalized shot-type taxonomy," *ACM Computing Surveys (CSUR)*, vol. 52, no. 5, pp. 1–33, 2019.

[11] 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.

[12] 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.

[13] 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.

[14] P. Nousi, I. Mademlis, I. Karakostas, A. Tefas, and I. Pitas, "Embedded UAV real-time visual object detection and tracking," in *Proceedings of the IEEE International Conference on Real-time Computing and Robotics (RCAR)*, 2019.

[15] F. Patrona, P. Nousi, I. Mademlis, A. Tefas, and I. Pitas, "Visual object detection for autonomous UAV cinematography," in *Proceedings of the Northern Lights Deep Learning Workshop*, 2020.

[16] C. Symeonidis, E. Kakaletsis, I. Mademlis, N. Nikolaidis, A. Tefas, and I. Pitas, "Vision-based UAV safe landing exploiting lightweight deep neural networks," in *Proceedings of the International Conference on Image and Graphics Processing (ICIGP)*, 2021.

[17] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. M. Paixao, and F. Mutz, "Self-driving cars: A survey," *Expert Systems with Applications*, vol. 165, p. 113816, 2021.

[18] A. Chowdhury, G. Karmakar, J. Kamruzzaman, A. Jolfaei, and R. Das, "Attacks on self-driving cars and their countermeasures: A survey," *IEEE Access*, vol. 8, pp. 207 308–207 342, 2020.

[19] Q. Rao and J. Frtunikj, "Deep learning for self-driving cars: Chances and challenges," in *Proceedings of the International Workshop on Software Engineering for AI in Autonomous Systems*, 2018.

[20] L. Dong, X. Chen, R. Wang, Q. Zhang, and E. Izquierdo, "ADORE: An adaptive holons representation framework for human pose estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2803–2813, 2017.

[21] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[22] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara, "Poseidon: Face-from-depth for driver pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[23] J. Wiederer, A. Bouazizi, U. Kressel, and V. Belagiannis, "Traffic control gesture recognition for autonomous vehicles," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[24] M. Jeong, B. C. Ko, and J.-Y. Nam, "Early detection of sudden pedestrian crossing for safe driving during summer nights," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 6, pp. 1368–1380, 2016.

[25] Z. Fang and A. M. López, "Intention recognition of pedestrians and cyclists by 2D pose estimation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 11, pp. 4773–4783, 2019.

[26] K. Chen, X. Song, and X. Ren, "Pedestrian trajectory prediction in heterogeneous traffic using pose keypoints-based convolutional encoder-decoder network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1764–1775, 2020.

[27] A. Marginean, R. Brehar, and M. Negru, "Understanding pedestrian behaviour with pose estimation and recurrent networks," in *Proceedings of the IEEE International Symposium on Electrical and Electronics Engineering (ISEEE)*, 2019.

[28] Z. Fang and A. M. López, "Is the pedestrian going to cross? Answering by 2D pose estimation," in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, 2018.

[29] P. R. G. Cadena, M. Yang, Y. Qian, and C. Wang, "Pedestrian graph: Pedestrian crossing prediction based on 2D pose estimation and Graph Convolutional Networks," in *Proceedings of the IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019.

[30] F. Yang, Y. Wu, S. Sakti, and S. Nakamura, "Make skeleton-based action recognition model smaller, faster and better," in *Proceedings of the ACM Multimedia Asia*, 2019.

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

[32] D. Makrygiannis, C. Papaioannidis, I. Mademlis, and I. Pitas, "Optimal video handling in on-line hand gesture recognition using Deep Neural Networks," in *Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI)*, 2021.

[33] F. Patrona, I. Mademlis, and I. Pitas, "Self-supervised Convolutional Neural Networks for fast gesture recognition in Human-Robot Interaction," in *Proceedings of the IEEE International Conference on Information and Automation for Sustainability (ICIAfS)*, 2021.

[34] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[35] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[36] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[37] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[38] P. Isola, J. Zhu, T. Zhou, and A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[39] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[40] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[41] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[42] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[43] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[44] X. Chu, W. Ouyang, H. Li, and X. Wang, "Structured feature learning for pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[45] W. Yang, W. Ouyang, H. Li, and X. Wang, "End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[46] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[47] L. Zhao, J. Xu, C. Gong, J. Yang, W. Zuo, and X. Gao, "Learning to acquire the quality of human pose estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 4, pp. 1555–1568, 2020.

[48] L. Zhao, N. Wang, C. Gong, J. Yang, and X. Gao, "Estimating human pose efficiently by parallel pyramid networks," *IEEE Transactions on Image Processing*, vol. 30, pp. 6785–6800, 2021.

[49] C. Yu, B. Xiao, C. Gao, L. Yuan, L. Zhang, N. Sang, and J. Wang, "Lite-hrnet: A lightweight high-resolution network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[50] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[51] Y. Wang, M. Li, H. Cai, W.-M. Chen, and S. Han, "Lite pose: Efficient architecture design for 2d human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[52] A. Martínez-González, M. Villamizar, O. Canévet, and J.-M. Odobez, "Efficient convolutional neural networks for depth-based multi-person pose estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 4207–4221, 2019.

[53] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[54] C. Papaioannidis, V. Mygdalis, and I. Pitas, "Domain-translated 3D object pose estimation," *IEEE Transactions on Image Processing*, vol. 29, pp. 9279–9291, 2020.

[55] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[56] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[57] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2004.

[58] 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.

[59] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[61] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[62] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[63] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[64] Z. Wang, W. Li, B. Yin, Q. Peng, T. Xiao, Y. Du, Z. Li, X. Zhang, G. Yu, and J. Sun, "Mscoco keypoints challenge 2018," in *Joint Recognition Challenge Workshop at ECCV*, vol. 5, 2018.

[65] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," *arXiv preprint arXiv:1611.05424*, 2016.

[66] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, and K. Murphy, "Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[67] M. Kocabas, S. Karagoz, and E. Akbas, "Multiposenet: Fast multi-person pose estimation using pose residual network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[68] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[69] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

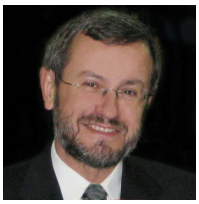
[70] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[71] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[72] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
[73] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[74] K. Sun, C. Lan, J. Xing, W. Zeng, D. Liu, and J. Wang, "Human pose estimation using global and local normalization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.
[75] Y. Bin, Z.-M. Chen, X.-S. Wei, X. Chen, C. Gao, and N. Sang, "Structure-aware human pose estimation with graph convolutional networks," *Pattern Recognition*, vol. 106, p. 107410, 2020.
[76] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.
[77] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.
[78] Z. Tang, X. Peng, S. Geng, L. Wu, S. Zhang, and D. Metaxas, "Quantized densely connected u-nets for efficient landmark localization," in *European Conference on Computer Vision (ECCV)*, 2018.
[79] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (CVPRW)*, 2017.
[80] P. Gujjar and R. Vaughan, "Classifying pedestrian actions in advance using predicted video of urban driving scenes," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2019.
[81] K. Saleh, M. Hossny, and S. Nahavandi, "Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal DenseNet," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2019.
[82] M. Chaabane, A. Trabelsi, N. Blanchard, and R. Beveridge, "Looking ahead: Anticipating pedestrians crossing with future frames prediction," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020.

Christos Papaioannidis received his Diploma in Electrical & Computer Engineering (2015) from the Aristotle University of Thessaloniki (AUTH). He is currently pursuing a Ph.D. in deep learning and computer vision at AUTH.

Dr. Ioannis Mademlis (IEEE Senior Member) received a Ph.D. in machine learning and computer vision (2018), from the Aristotle University of Thessaloniki, Greece (AUTH). Presently, he is employed as a postdoctoral research associate at AUTH. He has co-authored more than 50 publications in academic journals and international conferences. His current research interests include machine learning, computer vision, natural computing, autonomous robotics and intelligent cinematography.



Prof. Ioannis Pitas (IEEE Fellow, IEEE Distinguished Lecturer, EURASIP Fellow) received a Ph.D. in Electrical Engineering from the Aristotle University of Thessaloniki, Greece. His research interests cover image/video processing, machine learning, computer vision, intelligent digital media and biomedical imaging. He has published over 920 papers, contributed in 45 books in his areas of interest and edited or (co-)authored another 11 books. He has participated in 71 R&D projects, primarily funded by the European Union and is/was principal investigator/researcher in 43 such projects. He has 34400+ citations (Google Scholar) to his work and h-index 87+ (Google Scholar).