SEFD: Learning to Distill Complex Pose and Occlusion

Anonymous ICCV submission

Paper ID 11451

Abstract

This paper addresses the problem of three-dimensional (3D) human mesh estimation in complex poses and occluded situations. Although many improvements have been made in 3D human mesh estimation using the two-dimensional (2D) pose with occlusion between humans, occlusion from complex poses and other objects remains a consistent problem. Therefore, we propose the novel Skinned Multi-Person Linear (SMPL) Edge Feature Distillation (SEFD) that demonstrates robustness to complex poses and occlusions, without increasing the number of parameters compared to the baseline model. The model generates an SMPL overlapping edge similar to the ground truth that contains target person boundary and occlusion information, performing subsequent feature distillation in a simple edge map. We also perform experiments on various benchmarks and exhibit fidelity both qualitatively and quantitatively. Extensive experiments prove that our method outperforms the state-of-the-art method by 2.8% in MPJPE and 1.9% in MPVPE on a benchmark 3DPW dataset in the presence of domain gap. Also, our method is superior in 3DPW-OCC, 3DPW-PC, RH-Dataset, OCHuman, CrowdPose, and LSP dataset in which occlusion, complex pose, and domain gap exist.

1. Introduction

Human mesh estimation, which has been used recently in various applications such as digital human and action recognition, targets to generate a 3D mesh by estimating 3D semantic human joints and human mesh vertex locations from 2D input images. However, directly estimating 3D mesh information from input 2D images poses a significant challenge due to the ambiguity in estimating body part exact locations. Especially in complex pose cases (e.g., yoga, crouching, twist) or those where occlusion between humans occur, the performance deteriorates. To address these problems, recent studies [1, 2] employ additional information predicted by 2D pose detectors. Specifically, Choi et al. [1] illustrated the problem of global average pooling and utilized 2D pose information that includes the target person in the feature map as a resolution. Khirodkar et al. [2] proposed the center map, modularized in the encoder, as well as two losses specifically, the interpenetration loss and depth ordering loss to solve the occlusion between humans. These studies have improved 3D mesh estimation accuracy by incorporating the prior 2D pose knowledge, but they still struggle with occlusions and complex poses in the wild datasets containing a large domain gap.

First, a problem with the baseline 3D mesh (3DCrowdNet [1]) in that it cannot be properly extracted as in Figs. 1 (a) and (b). The 2D pose for each image is estimated properly, but the 3D mesh estimate is inaccurate. Therefore, additional structural information is essential to estimate complex poses. Second, if an object is occluded by other people or objects, the 3D mesh estimation accuracy decreases. In Figs. 1 (c) and (d), the baseline model cannot estimate an appropriate mesh due to severe occlusion, necessitating additional guidance on occlusion.

This paper proposes a novel methodology, SMPL Edge Feature Distillation (SEFD) which can solve occlusion and complex poses as limitations of 3D pose estimation. To this end, Fig. 2 displays a novel SMPL overlapping edge serving as a ground-truth (GT) edge map. To exploit this SMPL overlapping edge under real-world conditions where GT does not exist, we utilize another edge extracted by the simple edge detector. Thereafter, feature distillation produces a simple edge map to reduce the difference between the edges extracted by the simple edge detector and the SMPL overlapping edges. Conventional knowledge distillation methods for human poses aim to lighten student models [3, 4, 5, 6, 7]. Our feature distillation aims to distill the ground-truth feature representations of the teacher model to the student model solving the real-environment condition without ground-truth. We propose a new approach robust to occlusion and complex poses while account for structural information in 3D mesh estimation. Our contribution is summarized as follows:

- We designed the SMPL overlapping edge generation, which contains occlusion information while being robust to complex poses. The superiority of the proposed

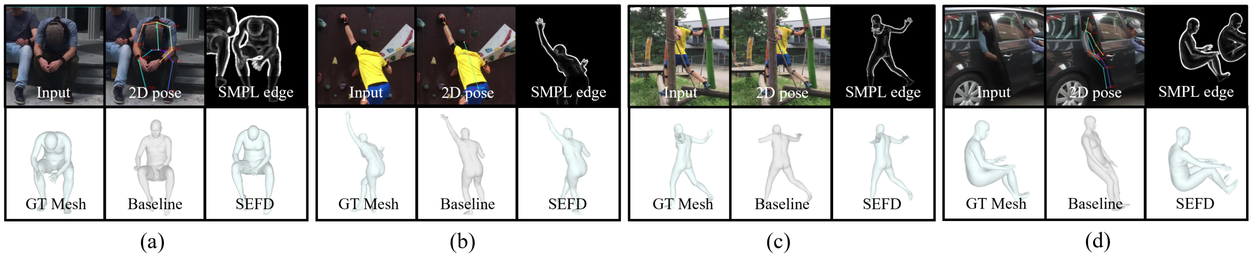

Figure 1. Qualitative comparison of the baseline [1] and the proposed feature distillation SMPL overlapping edge. (a) and (b) show complex poses and (c) and (d) shows occluded situations.

method persists through several experiments. We provide the preconfigured occlusion & complex pose annotation and codes to support SMPL overlapping edge generation.

- To efficiently utilize the SMPL overlapping edge under real-world conditions where GT does not exist, we propose a novel method, SMPL edge feature distillation (SEFD), where simple edge map is distilled to reduce the difference between the simple edge and the SMPL overlapping edge.
- Extensive studies demonstrate that the proposed method is the most effective for complex poses and occlusion on various public datasets outperforming previous state-of-the-arts methods.

2. Related Works

3D Human Mesh Recovery Estimation. 3D Human Mesh Estimation is the field of estimating mesh directly or using the pose, shape, and trans parameters, which are parameters described in [8]. In human mesh estimation, top-down approaches appear in [9, 10, 11, 12, 2, 13, 14, 15, 16, 1, 17, 18, 19, 20, 21, 22, 23, 24] and bottom-up approaches in [25, 26, 27]. The top-down approach is a method of estimating a single person's mesh in a bounding box. When used as an input for the model, the person's bounding box is characterized robustly by the similarity of human scale. Alternatively, the bottom-up approach changes to a certain size when the image is used as an input. Therefore, an originally high-resolution image changes into a low-resolution image and affects the person‘s scale. This poses a challenge for human mesh estimation of a small-scale person. Accordingly, most recent state-of-the-art algorithms are top-down approaches. However, human mesh estimation is not performed appropriately in the case of overlapping people in the top-down approach.

Occlusion-aware mesh estimation Various methods, such as [28, 29, 2, 27, 23, 1, 25, 26], have emerged to solve the occlusion concern, and each idea is logically convincing and yields favorable results. Specifically, [28] used occlusion-robust pose-maps (ORPM) to make it possible to infer full-body poses from partial occlusions. However, proximity of multiple points of the same type results in inaccuracy. Furthermore, [29] proposed an occlusion-robust method, but a disadvantage emerges when the neck is covered; identifying people becomes difficult and the model suffers in occlusions situations such as hugs between people. Conversely, [2] solved the occlusion by creating a module that can provide global and local information through the Context Normalization (CoNorm) for Human Center heatmap. However, there is a problem of relying on the Human Center map. And, the problem of relying on the 2D pose detection model can result in inaccurate human center maps. Similarly, [27, 23] solves occlusions by creating occlusion annotations, but the annotations were made from [30] and [31]. Therefore, occlusion was not solved with certainty because of their inaccuracy. Additionally, [1] offers 2d pose guidance, but does not solve occlusions with objects and people. Whereas [25, 26] estimates the mesh by estimating the human center map and mesh parameter maps. Also, [26] includes an additional relative depth to make it occlusion-aware. These methods yield promising results with occlusions, but may not be properly estimated if occlusions are present in the human heatmap. Unlike [1], SEFD is robust to occlusion because it uses the SMPL Edge Map, which explicitly informs occlusions of objects and people. Additionally, complex poses can be solved by using the edge map‘s structural features to help distinguish a person's poses.

Edge detection. Edge detection represents one of the most fundamental tasks in the field of computer vision and image processing. In early stages, hand-crafted edge detection methods such as the Canny edge detector [32] and Sobel filtering [33] have been used. With the advent of deep learning-based methods, model-based edge detection techniques have been studied. This includes [34], which used hierarchical Convolutional Neural Networks (CNNs) to utilize semantic and fine features to carry out edge detection, and [35], which developed a multi-scale CNN architecture to learn rich hierarchical representations in the edge and object boundary detection. Similarly, [36] proposed a lightweight edge detection model that utilizes traditional edge detection operators into CNNs. Edge detection has been utilized as guidance in many vision tasks [37, 38, 39, 40, 41, 42, 43].

Knowledge distillation. Knowledge distillation provides a technique to transfer general knowledge extracted from a source teacher model to a target student model [44]. Simple yet effective for model compression, response-based knowledge refers to imitating the neural response of the teacher model's final output [44, 45, 3, 46]. Taking advantage of its ability to extract essential features of deep neural networks with increasing abstraction, feature based knowledge distillation utilizes intermediate layers, i.e., feature representations, to supervise the student model [47, 48, 49, 50, 51]. Recently, many knowledge distillation methods have been applied widely in different fields of computer vision tasks [52, 53, 54, 55, 56]. We utilized feature-based knowledge distillation not to lighten but to supply guidance to the student model, to ensure it follows the teacher model's features. We provide a detailed description in the following section.

3. Proposed Method

Fig. 2 exhibits the proposed method's overall architecture, SMPL edge feature distillation (SEFD). The SMPL overlapping edge includes the structural information close to the ground-truth expressing the occlusion. It is generated by the SMPL edge map generator, concatenated with the input image for the teacher model training. We use the trained teacher model's encoder to teach the student model through the proposed feature distillation. The detailed operation includes the following elements.

3.1. SMPL Edge Map Generator

This section explains the detailed process of the SMPL edge map generator with the occlusion description from Fig. 2 (a). The SMPL edge map generator sets the world coordinates independently and projects each of them to different image planes. Thereafter, each image experiences edge detection and adaptive dilation to create an SMPL overlapping edge, $I^{overlap\_edge}$, which is the SMPL edge with occlusion description.

Mesh-to-image projection. This section refers to the mesh-to-image projection in Fig. 2 (a) onto the SMPL edge map generator. Firstly, SMPL [8] obtains a total of 6,890 3D body meshes $\overrightarrow{B}$ through the SMPL model using 23 real-valued 3D poses $\overrightarrow{\theta}$ representing human poses and 10 real-valued shapes $\overrightarrow{\beta}$, involving human shape information; $\overrightarrow{\theta} \in \mathbb{R}^{23\times3}$, $\overrightarrow{\beta} \in \mathbb{R}^{10}$, and $\overrightarrow{B} \in \mathbb{R}^{6890\times3}$. To perform image projection, camera intrinsic parameters are required. Those parameters include, the focal length, the distance from the camera to the image plane, $\{f_x, f_y\}$, and the principal point, the center coordinate of the image plane, $\{c_x, c_y\}$. Define the original image as $I_O \in \mathbb{R}^{H\times W\times3}$. If I_O includes N people of SMPL parameters, each parameter is defined as $\Theta_N = \{\overrightarrow{\beta}, \overrightarrow{\theta}, \overrightarrow{T_{SMPL}}\}$ and each focal length and principal point that corresponds to the person is called $F_N = \{f_{xN}, f_{yN}\}$ and $C_N = \{c_{xN}, c_{yN}\}$ respectively. However, the image projection may perform improperly in the following two cases.

First, when an extrinsic camera parameter $\overrightarrow{R} \in \mathbb{R}^{3\times3}$ and $\overrightarrow{T} \in \mathbb{R}^{3\times1}$ are not set to each scene's camera parameter, the image projection fails ($\overrightarrow{R}$ is the angle of the camera view and $\overrightarrow{T}$ is the camera translation parameter).

Second, if the camera intrinsic parameter is changed depending on each person, the image projection also fails. For an example displayed in Fig. 3 (b), when several people are image-projected naively using pseudo ground-truth SMPL parameters in MSCOCO [57], problems occur because F_N and C_N differ for each person in a single image. Therefore, we solved the problem such that all people presented in the image possess the same focal length and principal point (the detailed process appears in the supplementary materials).

Finally, the image projection is performed using the experimentally determined F_N and C_N. Through this, an accurate pseudo ground-truth SMPL map displays in Fig. 3 (c). Projecting single Θ_N, F_N, and C_N is called a single SMPL map I_N, and it is defined as below:

$$I_N = \mathit{Proj}(\Theta_N, F_N, C_N), \tag{1}$$

where *Proj* means projection to an image plane.

Edge detection. This section refers to edge detection in the SMPL edge map generator in Fig. 2 (a) to pass I_N to the edge detector. For robust, 3D, human mesh recovery in complex poses, the structural information including the occlusion description is necessary. Thus, we use the edge map, a widely-used structural maps, by utilizing simple edge detection such as that employed in [32], [35], and [36]. If we perform the edge detection naively, we obtain a noisy edge as it extracts unwanted texture information from the background, as well as outlines or object boundaries. However, we require an edge map that retains only human boundary information for 3D human mesh estimation (and avoids noisy edges). Hence, we distinguish these noisy edges from those where only human boundary information remains.

The results from the edge detection phase yield, an edge map from the SMPL parameter I_N^{edge} which is extracted for each people N. It can be expressed in the following equation:

$$I_N^{edge} = \mathit{Edge}(I_N), \tag{2}$$

where *Edge* represents the simple edge detector. In our case, a Canny edge detector [32] was adopted.

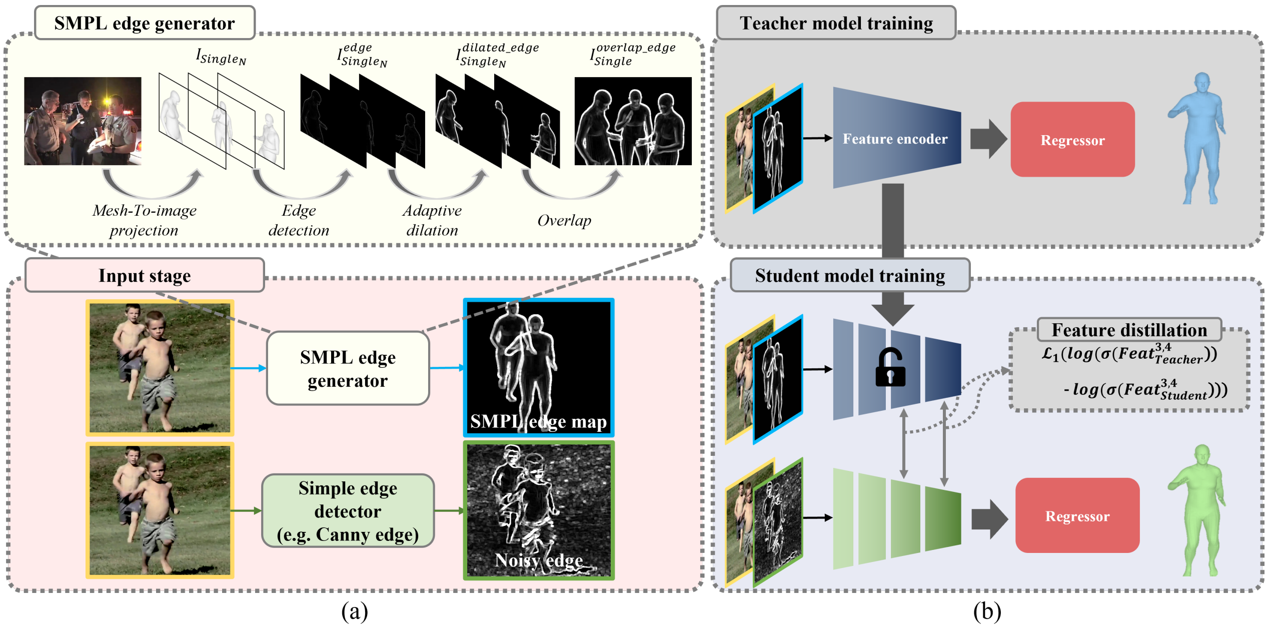

Figure 2. Visualization of the proposed SMPL edge feature distillation (SEFD). Overall processes of (a) the proposed SMPL edge map generator and (b) the proposed feature distillation.

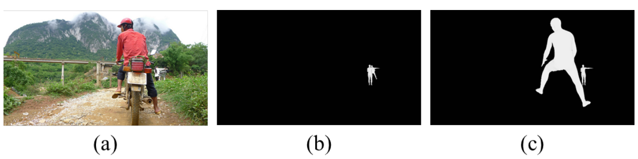

Figure 3. A failure case using MSCOCO [57] benchmark pseudo ground-truth. (a) an input image, (b) a falsely projected image, and (c) an image using the proposed method to solve the focal length and principal point problems.

Adaptive dilation. This section demonstrates the adaptive dilation phase in the SMPL edge map generator of Fig. 2 (a). This phase performs dilation to I_N^{edge}, the output from the edge detection phase. The role of the dilation involves highlighting the human boundary information over the general edges. The results from the dilation phase can be expressed in the following equation:

$$I_N^{dilated\_edge} = Edge_{n\times n}(I_N), \tag{3}$$

where $Edge_{n\times n}$ means performing the dilation using a kernel with $n \times n$ pixels after the edge detector.

We propose the effective adaptive dilation over conventional dilation methods. Adaptive dilation is a method of effectively changing the kernel size according to an object's size. Fig. 4 (c) is an edge map generated from the edge detection phase on the small object, and Fig. 4 (d) is acquired following naive $n \times n$ dilation with $n = 5$. While the person's structural information is preserved in Fig. 4 (c), the structural information is crushed in Fig. 4 (d), and

S_{area}	Dilation kernel size
$1 \times 1 \sim 16 \times 16$ pixels	1×1
$16 \times 16 \sim 64 \times 64$ pixels	3×3
$64 \times 64 \sim 128 \times 128$ pixels	5×5
$128 \times 128 \sim 256 \times 256$ pixels	7×7
$256 \times 256 \sim 512 \times 512$ pixels	9×9

Table 1. Various adaptive dilation kernel sizes according to the bounding box region of a person S_{area}.

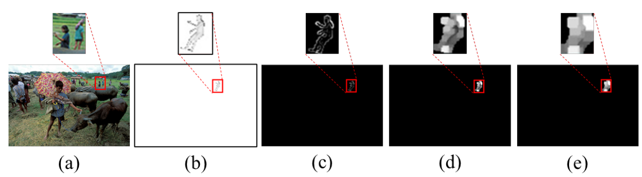

Figure 4. Results of performing Canny edge detection [32] and dilation for a small object. (a) Input image, (b) SMPL map, results extracted by Canny edge detections [32] (c) without dilation, (d) with 5×5 dilation, and (e) with 9×9 dilation.

the person's shape becomes unrecognizable in Fig. 4 (e). Correspondingly, to save a small object's boundary information, we employ a bounding box histogram to identify the histogram distribution and propose adaptive dilation that maximizes human structural information. Adaptive kernel size for dilation appears in Table 1 using a bounding box area S_{area} (the detailed explanation is described in the supplementary materials).

Using the proposed adaptive dilation as shown in Fig. 5,


(a) (b) (c) (d) (e)

Figure 5. Results for adaptive dilation depending on the size of the object. (a) Input image, (b) SMPL overlapping edge (adaptive dilation), (c) zoomed in red box from (b), (d) SMPL overlapping edge (5×5 dilation), and (e) zoomed in blue box from (d).

by applying minimal dilation to a small object and extensive dilation to a large object, it is possible to create detailed internal structural information while preserving the target person's boundary information.

Overlapping dilated SMPL edge maps. This section refers to the overlapping phase in Fig. 2 (a). All dilated SMPL edge maps $I_N^{dilated\_edge}$ are added and saturated to make the SMPL overlapping edge $I^{overlap\_edge}$. The equation is as follows.

$$I^{overlap\_edge} = \sum_{i=1}^{N} I_i^{dilated\_edge}. \quad (4)$$

Visualization of $I^{overlap\_edge}$ exhibits in the last phase of Fig. 2. This edge offers the advantage of being able to convey information clearly on the edge map's occlusion. Compared with various structural maps, the proposed SMPL overlapping edge $I^{overlap\_edge}$ achieved superior performance improvement, and therefore, we defined it as the final teacher model (verified in the experimental section).

3.2. Feature Distillation

The proposed architecture appears in Fig. 2. We adopt feature distillation to mimic the teacher model's output from the simple student model input. I_O, and $I^{overlap\_edge}$ are concatenated and imported into the teacher model as done previously. The student model is fed a new edge map, concatenation of I_O and I^{simple_edge}, where I^{simple_edge} is the naive edge detector's output. In other words, the student model starts with a simple edge from simple edge detector I^{simple_edge} to mimic the SMPL overlapping edge $I^{overlap\_edge}$, the teacher model's output. This is designed to reduce the structural gap between I^{simple} and $I^{overlap\_edge}$, given a severe difference and that direct estimation of $I^{overlap\_edge}$ proves impossible.

Utilizing the general losses used in feature distillation [51, 44, 58, 59], losses are applied such that the teacher encoder features could be distilled adequately to the student encoder. The loss configuration is as follows:

$$\mathcal{L}_1(Log(\sigma(Feat^{3,4}_{Teacher})) - Log(\sigma(Feat^{3,4}_{Student}))), \quad (5)$$

where $\mathcal{L}_1$ is the $\mathcal{L}_1$ calculation, Log the logarithmic function, and σ the Softmax function. $Feat^{3,4}$ refers to the third and fourth layers in which the feature size decreases in the encoder. The reason for this setting is that each feature map offers a different capacity that can contain the structural information difference, and various feature map combinations are included, though difficult to learn due to a large spatial information difference. The feature map connection with the best performance is used (it will be explained in the experiments section).

4. Experiments

We present the experimental results to demonstrate the proposed 2D pose and 3D mesh estimation method's effectiveness with common datasets [28, 60, 61, 25, 26, 62, 63, 64, 65] by comparing them with other state-of-the-art 3D mesh estimation methods. We analyze the structural maps results to utilize the SMPL edge fully, and prove the proposed method's superiority by thoroughly performing ablation studies on the losses, optimal feature layers for feature distillation, and various edge maps.

4.1. Datasets

Training Datasets. We used Human3.6M [66], MuCo-3DHP [28], MSCOCO [57], and MPII [67] as training datasets, according to the standard split protocols defined in [66, 28, 57, 67].

Test Datasets. The 3DPW [60] dataset, commonly used as a test dataset, was compared and classified into conditions with and without domain gaps. *Our experiment, did not use 3DPW dataset for training.* Occlusion-related datasets 3DPW-OCC [61, 60], 3DPW-PC [60, 25], RH-Dataset (RH-D) [26], OCHuman [62], and CrowdPose [63] and complex pose related datasets, LSP [64, 65] and OCHMR [2], Liu et al [27], VisDB [23] , CLIFF [24], 3DCrowd-Net [1] are compared. Also, the results of MuPoTs [28] are compared in the supplementary materials.

Evaluation Metrics. For performance evaluation, we employed mean per-joint position error (MPJPE), procrustes-aligned mean per-joint position error (PA-MPJPE), mean per-vertex point position error (MPVPE), the mean percentage of correct key points ($mPCK_h^{0.6}$), as well as AP, AP^{50}, and AP^{75}, which are standard metrics using Object Keypoint Similarity (OKS) [57].

4.2. Effectiveness of Structural Map

SMPL edge map without occlusion description. To address the validity of the proposed $I^{overlap\_edge}$ on occluded scenes, we address the SMPL edge without an occlusion description $I_{\forall}^{dilated\_edge}$. It follows a similar procedure to the SMPL edge map generator, only without overlapping each dilated edge, since all people in an image are already considered. This becomes the equation below:

	Method	MPJPE(↓)	PA-MPJPE(↓)	MPVPE(↓)
No Structure	Baseline	81.7	51.5	98.3
Simple Edge Detecter	Canny edge [32]	80.05	49.45	95.86
	Canny edge 3×3 [32]	79.87	49.86	95.22
	Canny edge 5×5 [32]	79.16	49.88	94.84
	Canny edge 7×7 [32]	80.06	49.38	95.52
	Canny edge 9×9 [32]	80.45	49.72	95.90
	HED [35]	79.53	49.91	95.09
	HED 3×3 [35]	80.03	50.12	96.26
	HED 5×5 [35]	80.84	50.35	96.60
	RCF [34]	79.78	49.94	95.67
	PiDiNet [36]	79.78	49.88	95.76
EPS	RTV [68]	79.94	50.06	95.11
Instance Segmentation	Mask2former [69]	79.43	49.80	95.13
SMPL edge (Ground Truth)	$I_{\forall}^{dilated\_edge}$	70.07	46.04	84.24
	$I^{overlap\_edge}$	**64.75**	**43.79**	**78.36**

Table 2. Evaluation on preprocessing methods using several structural information.

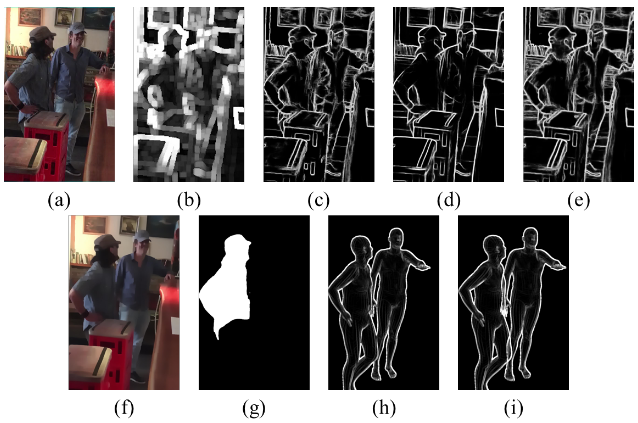

Figure 6. Results for various structural maps. (a) Input image, results by (b) Canny edge [32] 5×5 dilation, (c) HED [35], (d) PiDiNet [36], (e) RCF [34], (f) RTV method [68], (g) Mask2Former [69], (h) $I_{\forall}^{dilated\_edge}$, and (i) $I^{overlap\_edge}$.

$$I_{\forall} = Proj(\sum_{i=1}^{N}(\Theta_i), F_N, C_N), \tag{6}$$

where $I_{\forall}$ is the output of projection from all people in an image. In the same way as the $I^{overlap\_edge}$ production described above, $I_{\forall}$'s Edge detection and dilation are carried out

$$I_{\forall}^{edge} = Edge(I_{\forall}), \tag{7}$$

$$I_{\forall}^{dilated\_edge} = Edge_{5\times5}(I_{\forall}). \tag{8}$$

While $I^{overlap\_edge}$ considers both occlusions between a person and an object and those overlapping between people, $I_{\forall}^{dilated\_edge}$ only accounts for occlusions between a person and an object, as displayed in Figs. 6 (h) and (i). For convenience, we define Canny edge [32] as the output of the Canny edge detector [32], and the SMPL edge calls both the SMPL overlapping edge $I^{overlap\_edge}$ and the SMPL edge map without an occlusion description $I_{\forall}^{dilated\_edge}$.

Performance comparison for various structural map generations. Table 2 conveys the structure map results for

	MPJPE(↓)	PA-MPJPE(↓)	MPVPE(↓)
only inference	98.29	62.30	115.99
SMPL edge estimator	81.19	50.67	96.98
feature distillation	**77.37**	**49.39**	**92.6**

Table 3. Performance comparison of three approaches to efficiently use the SMPL overlapping edge $I^{overlap\_edge}$.

the simple edge detector [32, 35, 34, 36], edge preserving image smoothing [68], instance segmentation [69], and proposed SMPL edge ($I^{overlap\_edge}$ and $I_{\forall}^{dilated\_edge}$).

Except for the SMPL edge, the 3D mesh and pose estimation accuracy provided optimal performance when Canny edge [32] with 5×5 dilation was used, as shown in Fig. 6 (b). This 5×5 dilation improved the performance by highlighting the human boundary information. However, if the dilation size expanded too significantly, the accuracy decreased. Accordingly, it was confirmed that the accuracy decreases as detailed structural information disappears after applying over a certain dilation size. This occurs because edge detectors [35, 36, 34] fail to detect detailed edges when the background color matches that of a person, leading to failure in extracting human body structural information. Meanwhile, the Canny edge [32] detection improved performance by emphasizing a boundary through dilation because it detects almost all edges even with a background and person color similarity. [68] preserved the images's edge components and smoothed non-edge. Thus, it preserved the whole image's structural information, but did not emphasize the person's structural information, yielding even poorer performance than other structural maps. For instance segmentation, the state-of-the-art method [70] was adopted. As captured in Fig. 6 (g), occlusion information could not be predicted, and detailed body part information was unknown. Therefore, even considering the structural information, the performance improvement was not significant (the superiority and visualization of the Canny edge [32] also appears in the supplementary material).

Finally, the performance improved greatly when the SMPL edge map was applied. Also, the performance greatly increased by clearly providing the specific structural and target person's boundary information. $I^{overlap\_edge}$ provided the greatest performance improvement by additionally considering the occlusion information.

Performance comparison for various approaches using $I^{overlap\_edge}$. We confirmed that performance improved when the SMPL overlapping edge $I^{overlap\_edge}$ is used, though it could not be applied in the real environment as it was created using ground-truth. Therefore, we devised several novel approaches to apply the $I^{overlap\_edge}$. The first method was to input the canny edge [32] map directly into the trained model using the $I^{overlap\_edge}$. The second involved estimating the SMPL edge map directly using the image generator model [71], and the last method was to perform feature distillation. As available in Table 3, the first

(a) Input (b) 2D pose (c) I2L-MeshNet (d) SPIN (e) 3DCrowdNet **(f) SEFD (Ours)**

Figure 7. Visualization comparison with state-of-the-arts methods on benchmark datasets. (a) Input image, (b) visualization of 2D pose, (c) I2L-MeshNet [12], (d) SPIN [10], (e) 3DCrowdNet [1] (baseline), and (f) the proposed SEFD.

method exhibited extremely poor performance. This results because the boundary information for a target person and occlusion contained in the Canny edge [32] presented a remarkable difference from that of the $I^{overlap\_edge}$.

The second approach of directly estimating $I^{overlap\_edge}$ using the U-Net structure [71] yielded little improvement in performance, and was not satisfactory when it came to complex poses or occlusions. In these situations, the SMPL edge map could not be estimated properly, which led to failure in human mesh estimation (detailed analysis will be available in the supplementary materials).

Therefore, we used the feature distillation method to consider the complex poses and occlusion situation. Unnecessary Canny edge [32] boundary information was removed through $I^{overlap\_edge}$, and occlusion information was distilled through feature distillation from pseudo ground-truth SMPL edges.

4.3. Comparisons with State-of-the-Arts Methods

Fig. 7 exhibits the comparison with state-of-the-art methods [12], [11], and [1] against the proposed method. Overall, as I2L-MeshNet [12] and SPIN [10] did not consider complex poses or occlusions between people or objects, their performance was not satisfactory. Therefore, the proposed method was compared only with the baseline 3DCrowdNet [1].

Not using 3DPW training Dataset				using 3DPW training Dataset			
Method	MPJPE(↓)	PA-MPJPE(↓)	MPVPE(↓)	Method	MPJPE(↓)	PA-MPJPE(↓)	MPVPE(↓)
HMR [9]	130	76.7	-				
GraphCMR [10]	-	70.2	-	TCMR [72]	86.5	52.7	102.9
SPIN [11]	96.9	59.2	116.4	I2L-MeshNet [12]	84.5	51.1	98.2
I2L-MeshNet [12]	93.2	57.7	110.1	VIBE [73]	82.0	51.9	99.1
Liu et al [27]	93.1	-	-	MAED [74]	79.1	45.7	92.6
ROMP [25]	91.3	54.9	108.3	BEV [26]	78.5	46.9	92.3
OCHMR [2]	89.7	58.3	107.1	METRO [19]	77.1	47.9	88.2
Pose2Mesh [13]	89.5	56.3	105.3	ROMP [25]	76.7	47.3	93.4
MAED [74]	88.8	50.7	104.5	HybrIK [20]	76.2	45.1	89.1
Song et al [14]	-	55.9	-	PARE (H) [16]	74.5	46.5	88.6
Fang et al.[75]	85.1	54.8	-	Mesh Graphormer [21]	74.7	45.6	87.7
Tuch [15]	84.9	55.5	-	PyMAF-X (R) [18]	76.8	46.8	88.7
PARE [16]	82.0	50.9	97.9	PyMAF-X (H) [18]	74.2	45.3	87.0
3DCrowdNet (R) [1]	81.7	51.5	98.3	D&D [22]	73.7	42.7	88.6
TCFormer * [17]	80.6	49.3	-	VisDB [23]	72.1	44.1	83.5
PyMAF-X (R) [18]	79.7	49.0	94.4	CLIFF (R) [24]	72.0	45.7	85.3
SEFD (R) (Ours)	**77.4**	**49.4**	**92.6**	**CLIFF (H)** [24]	**69.0**	**43.0**	**81.2**

Table 4. Results of using and not using the 3DPW training data set. It can seen how robust the proposed method is to the domain gap when 3DPW training data is used. "R" stands for Resnet backbone and "H" stands for HRNet backbone. "*" stands for pseudo 3DPW training dataset.

Fig. 7's top row displays an example of occlusion. In the baseline case, the three 2D poses were estimated to a certain level, but the person at the back was not completely restored. Our method provides robustness in this occlusion case and restores the occluded person. The performance of 3D mesh estimation is influenced greatly by the 2D pose estimator's output demonstrating the limitation when an estimator fails to generate a proper 2D pose in the second row of Fig. 7. We solved this problem by considering additionally structural information through SMPL edge feature distillation. The third row reveals an improvement in complex poses accuracy. In such complex poses, the 2D pose was

Method	Occlusion													Complex
	3DPW-OCC [61, 60]			3DPW-PC [60, 25]			RH-D [26]	OCHuman [62]			CrowdPose [63]			LSP [64, 65]
	MPJPE(↓)	PA-MPJPE(↓)	MPVPE(↓)	MPJPE(↓)	PA-MPJPE(↓)	MPVPE(↓)	$PCK_h^{0.6}$(↑)	AP(↑)	AP^{50}(↑)	AP^{75}(↑)	AP(↑)	AP^{50}(↑)	AP^{75}(↑)	$PCK_h^{0.6}$(↑)
OCHMR [2]	-	-	-	117.5	77.1	149.6	-	24.8	60.7	28.6	21.4	48.3	16.5	-
Liu et al [27]	94.4	-	-	-	-	-	-	-	-	-	-	-	-	-
VisDB [23]	87.3	56.0	110.5	-	-	-	-	-	-	-	-	-	-	-
CLIFF (R) [24] Openpose [76]	-	-	-	-	-	-	0.7690	17.7	38.5	14.9	19.9	37.9	18.0	0.8247
CLIFF (R) [24] YOLOX [77]	-	-	-	-	-	-	0.7613	23.2	49.8	20.4	27.4	**51.5**	25.9	**0.8407**
3DCrowdNet (R) [1]	88.61	56.79	103.22	88.77	57.26	105.68	0.8019	39.7	68.8	41.0	25.2	44.9	24.9	0.8186
SEFD (R)	**83.48**	**55.0**	**97.12**	**85.92**	**54.94**	**100.50**	**0.8259**	**44.1**	**71.7**	**47.2**	**29.0**	48.8	**30.0**	0.8315

Table 5. This table shows comparisons with Baseline, methods for solving occlusions, and comparisons with SOTA methods in using 3DPW [60] dataaset. "R" means for using Resnet [78] backbone. The openpose [76] and YOLOX [77] next to CLIFF are detector models.

Loss	MPJPE(↓)	PA-MPJPE(↓)	MPVPE(↓)
$\mathcal{L}_1$	79.15	49.78	94.80
FSP [59]	79.10	50.11	95.12
KD [44]	78.68	49.85	94.22
GC [51]	78.51	49.71	94.50
AT [58]	78.85	**49.52**	94.46
$Log\text{-}Softmax\text{-}\mathcal{L}_1$	**78.26**	49.85	**94.22**

Table 6. Results on utilizing various losses for feature distillation.

properly extracted to restore the human mesh to a plausible degree, but the hand position was not formed accurately. However, our proposed method formed the hand position adhering to the input.

We are comparing 3DPW as well as the occlusion and complex pose datasets with other models in Tables 4 and 5. From Table 4, our method outperforms other methods when a domain gap exists. Table 5 conveys that our proposed method outperforms OCHMR [2], Liu *et al.* [27], VisDB [23], and 3DCrowdNet [1], which are specialized methods for handling occlusion. Additionally, compared using RH-D [26], OCHuman [62], CrowdPose [63], and LSP [64, 65] datasets, where domain gap, occlusion, and complex poses exist, our results exceed those of CLIFF [24].

4.4. Ablation Study

Loss configuration. First, Table 6 describes the loss configuration. In order to select the optimal loss, we conducted an experiment by adopting various losses in the $Feat^{1,2,3,4}$ of the teacher and student encoders. Various losses commonly used in feature distillation were experimented, $Log\text{-}Softmax\text{-}\mathcal{L}_1$ yielding the superior accuracy. The channel-wise Softmax was performed to consider the individual channel effects in all feature maps. However, there exists a difference in the structural information between the Canny edge [32] and the SMPL edge map, providing an input to the feature encoder and posing a challenge for distillation. To neglect this difference between edge map inputs and thereby boost feature distillation performance, we applied various combinations of feature maps.

Optimal feature branch selection. Table 7 reveals the result of feature map combination. By removing the feature maps individually, the performance improved. Specifically, removing the initial feature maps strengthened performance due to severe structural information differences. When

feature map	MPJPE(↓)	PA-MPJPE(↓)	MPVPE(↓)
1^{st}, 2^{nd}, 3^{rd}, 4^{th}	78.26	49.85	94.22
2^{nd}, 3^{rd}, 4^{th}	77.88	**49.15**	93.43
$\mathbf{3^{rd}}$, $\mathbf{4^{th}}$	**77.37**	49.39	**92.6**
4^{th}	77.92	49.5	93.77

Table 7. Results on various combinations of feature map usage.

simple edge	MPJPE(↓)	PA-MPJPE(↓)	MPVPE(↓)
Canny edge 5 × 5 [32]	**77.37(-1.8)**	49.39(-0.5)	**92.60(-2.2)**
HED [35]	77.83(-1.7)	49.36(-0.6)	93.24(-1.9)
RCF [34]	77.72(-2.1)	49.07(-0.9)	93.14(-2.5)
PiDiNet [36]	77.80(-2.0)	**49.05(-0.8)**	93.45(-2.3)
SMPL edge estimator	78.87(-2.3)	49.47(-1.2)	94.56(-2.4)

Table 8. Various edge detector results for feature distillation.

$Feat^4$ is used, the spatial area containing the structural map features is too small, so the accuracy responds improperly. Therefore, we connected $Feat^{3,4}_{student}$ and $Feat^{3,4}_{teacher}$ and used $Log\text{-}Softmax\text{-}\mathcal{L}_1$ loss when performing feature distillation. Experiments were conducted with various edge detectors as captured in Table 8 to identify the edge with the optimal feature distillation.

Optimal edge selection for feature distillation. In Table 8, all performance has improved using simple edge [35, 34, 36] or SMPL edge estimators, while Canny [32] with 5×5 dilation exhibited the greatest performance. To this end, we demonstrated that feature distillation can improve performance for all edge detectors.

5. Conclusion

We presented a novel SMPL edge feature distillation (SEFD) method to solve complex pose and occlusion problems. This method effectively solved the existing occlusion problem and reduced the structural difference between SMPL and Canny edges [32]. Thus, complex pose and occlusion problems have been solved, demonstrated both qualitatively and quantitatively. Our proposed method is simple yet effective, outperforming existing state-of-the-art methods and offering wide application in the field of human mesh and pose estimation.

References

[1] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1475–1484, 2022. 1, 2, 5, 7, 8

[2] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1715–1725, 2022. 1, 2, 5, 7, 8

[3] Feng Zhang, Xiatian Zhu, and Mao Ye. Fast human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3517–3526, 2019. 1, 3

[4] Xuecheng Nie, Yuncheng Li, Linjie Luo, Ning Zhang, and Jiashi Feng. Dynamic kernel distillation for efficient pose estimation in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6942–6950, 2019. 1

[5] Chaoyang Wang, Chen Kong, and Simon Lucey. Distill knowledge from nrsfm for weakly supervised 3d pose learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 743–752, 2019. 1

[6] Philippe Weinzaepfel, Romain Brégier, Hadrien Combaluzier, Vincent Leroy, and Grégory Rogez. Dope: Distillation of part experts for whole-body 3d pose estimation in the wild. In *European Conference on Computer Vision*, pages 380–397. Springer, 2020. 1

[7] Zheng Li, Jingwen Ye, Mingli Song, Ying Huang, and Zhigeng Pan. Online knowledge distillation for efficient pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11740–11750, 2021. 1

[8] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2, 3

[9] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 2, 7

[10] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019. 2, 7

[11] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019. 2, 7

[12] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020. 2, 7

[13] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *European Conference on Computer Vision*, pages 769–787. Springer, 2020. 2, 7

[14] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *European Conference on Computer Vision*, pages 744–760. Springer, 2020. 2, 7

[15] Lea Muller, Ahmed AA Osman, Siyu Tang, Chun-Hao P Huang, and Michael J Black. On self-contact and human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9990–9999, 2021. 2, 7

[16] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021. 2, 7

[17] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11101–11111, 2022. 2, 7

[18] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *arXiv preprint arXiv:2207.06400*, 2022. 2, 7

[19] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1954–1963, 2021. 2, 7

[20] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3383–3393, 2021. 2, 7

[21] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12939–12948, 2021. 2, 7

[22] Jiefeng Li, Siyuan Bian, Chao Xu, Gang Liu, Gang Yu, and Cewu Lu. D &d: Learning human dynamics from dynamic camera. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 479–496. Springer, 2022. 2, 7

[23] Chun-Han Yao, Jimei Yang, Duygu Ceylan, Yi Zhou, Yang Zhou, and Ming-Hsuan Yang. Learning visibility for robust dense human body estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 412–428. Springer, 2022. 2, 5, 7, 8

[24] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pages 590–606. Springer, 2022. 2, 5, 7, 8

[25] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11179–11188, 2021. 2, 5, 7, 8

[26] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13243–13252, 2022. 2, 5, 7, 8

[27] Qihao Liu, Yi Zhang, Song Bai, and Alan Yuille. Explicit occlusion reasoning for multi-person 3d human pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 497–517. Springer, 2022. 2, 5, 7, 8

[28] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018. 2, 5

[29] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. Xnect: Real-time multi-person 3d motion capture with a single rgb camera. *Acm Transactions On Graphics (TOG)*, 39(4):82–1, 2020. 2

[30] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin 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)*, pages 269–286, 2018. 2

[31] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 2

[32] John Canny. A computational approach to edge detection. *IEEE TPAMI*, (6):679–698, 1986. 2, 3, 4, 6, 7, 8

[33] Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of solid-state circuits*, 23(2):358–367, 1988. 2

[34] Yun Liu, Ming-Ming Cheng, Xiaowei Hu, Kai Wang, and Xiang Bai. Richer convolutional features for edge detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3000–3009, 2017. 2, 6, 8

[35] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015. 2, 3, 6, 8

[36] Zhuo Su, Wenzhe Liu, Zitong Yu, Dewen Hu, Qing Liao, Qi Tian, Matti Pietikäinen, and Li Liu. Pixel difference networks for efficient edge detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5117–5127, 2021. 2, 3, 6, 8

[37] Heng Liu, Zilin Fu, Jungong Han, Ling Shao, Shudong Hou, and Yuezhong Chu. Single image super-resolution using multi-scale deep encoder–decoder with phase congruency edge map guidance. *Information Sciences*, 473:44–58, 2019. 3

[38] Kamyar Nazeri, Harrish Thasarathan, and Mehran Ebrahimi. Edge-informed single image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

[39] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8779–8788, 2019. 3

[40] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7264–7273, 2019. 3

[41] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

[42] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-aware image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5840–5848, 2019. 3

[43] Sihang Zhou, Dong Nie, Ehsan Adeli, Jianping Yin, Jun Lian, and Dinggang Shen. High-resolution encoder–decoder networks for low-contrast medical image segmentation. *IEEE Transactions on Image Processing*, 29:461–475, 2019. 3

[44] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 3, 5, 8

[45] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? *Advances in neural information processing systems*, 27, 2014. 3

[46] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017. 3

[47] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 3

[48] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems*, 31, 2018. 3

[49] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3779–3787, 2019. 3

[50] Xiaobo Wang, Tianyu Fu, Shengcai Liao, Shuo Wang, Zhen Lei, and Tao Mei. Exclusivity-consistency regularized knowledge distillation for face recognition. In *European Conference on Computer Vision*, pages 325–342. Springer, 2020. 3

[51] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4643–4652, 2022. 3, 5, 8

[52] Zhimao Peng, Zechao Li, Junge Zhang, Yan Li, Guo-Jun Qi, and Jinhui Tang. Few-shot image recognition with knowledge transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 441–449, 2019. 3

[53] Tong He, Chunhua Shen, Zhi Tian, Dong Gong, Changming Sun, and Youliang Yan. Knowledge adaptation for efficient semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 578–587, 2019. 3

[54] Yuenan Hou, Zheng Ma, Chunxiao Liu, Tak-Wai Hui, and Chen Change Loy. Inter-region affinity distillation for road marking segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12486–12495, 2020. 3

[55] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2019. 3

[56] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4183–4192, 2020. 3

[57] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3, 4, 5

[58] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 5, 8

[59] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4133–4141, 2017. 5, 8

[60] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018. 5, 8

[61] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7376–7385, 2020. 5, 8

[62] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 889–898, 2019. 5, 8

[63] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019. 5, 8

[64] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR 2011*, pages 1465–1472. IEEE, 2011. 5, 8

[65] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *bmvc*, volume 2, page 5. Aberystwyth, UK, 2010. 5, 8

[66] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 5

[67] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt 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*, pages 3686–3693, 2014. 5

[68] Li Xu, Qiong Yan, Yang Xia, and Jiaya Jia. Structure extraction from texture via relative total variation. *ACM transactions on graphics (TOG)*, 31(6):1–10, 2012. 6

[69] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 6

[70] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 6

[71] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 6, 7

[72] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1964–1973, 2021. 7

[73] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 7

[74] Ziniu Wan, Zhengjia Li, Maoqing Tian, Jianbo Liu, Shuai Yi, and Hongsheng Li. Encoder-decoder with multi-level attention for 3d human shape and pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13033–13042, 2021. 7

[75] Qi Fang, Qing Shuai, Junting Dong, Hujun Bao, and Xiaowei Zhou. Reconstructing 3d human pose by watching humans in the mirror. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12814–12823, 2021. 7

[76] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 8

[77] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 8

[78] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 8