CAMION: Cascade Multi-input Multi-output Network for Skeleton Extraction

Sheng Fang, Kaiyu Li, Zhe Li†
Shandong University of Science and Technology, China
fangs99@126.com, {likyoo, lizhe}@sdust.edu.cn

Abstract

Skeletonization is an important process of extracting the medial axis of the object shape while maintaining the original geometric and topological properties. Some recent studies have demonstrated that deep learning-based segmentation models can extract the main skeleton from objects more robustly. However, we find that the skeleton extracted by a vanilla segmentation process is always discontinuous and not accurate enough. In this paper, we propose a general cascade deep learning pipeline that achieves competitive performance only using a simple U-shape network. The semantic information contained in the shapes is limited, so we introduce a ConvNet with multi-source input and multi-task output, CAMION for short, on top of the basic shape-to-skeleton network. With the multi-source inputs, CAMION can converge faster than using only binary shapes; and with the introduction of multi-task learning, relevant and suitable auxiliary tasks (e.g., feature point detection and contour extraction) bring considerable gains for the extraction of skeleton. Our code used in Pixel Skel-NetOn - CVPR 2022 challenge will be released at https: //github.com/likyoo/CAMION-CVPRW2022.

1. Introduction

Skeletonization can reduce the dimensionality of an image object to a "medial axis", providing an efficient and compact "skeleton" of the image object representation. An accurate medial axis transformation can remove redundant information while maintaining critical topological and geometric properties of the object, which facilitates the extraction of object features in subsequent tasks, e.g., human action recognition [9, 13, 28], segmentation [10], etc.

Some traditional skeletonization algorithms make the object "thinner" by "peeling" layer by layer, using geometric properties. They progressively remove points from the original object, but keep the original morphology, until they get the skeleton of the target [7, 29]. Their results are smooth and continuous on the whole, but always miss some smooth corners or generate burrs lines at curved shapes,

such as the shape of a bird's head. Different from traditional skeleton extraction algorithms, deep learning-based skeleton extraction methods exhibit significant discontinuous lines and false-negative pixels in the extracted skeleton due to the impact of sample imbalance in the supervised signals. In addition, although the shape of the object directly contains the geometric information of the object, it contains limited semantic features.

In this paper, we propose our solution for the Pixel Skel-NetOn challenge in the Deep Learning for Geometric Computing - CVPR 2022 Workshop and Challenge, which includes:

• We propose CAMION, a **CA**scade **M**ulti-**I**nput multi-**O**utput **N**etwork that can obtain better performance from several auxiliary tasks such as feature point detection and contour extraction. The auxiliary tasks enrich the potential semantics and the supervision of feature points emphasizes the key features of the skeleton.

• We further show that our model, a general cascade manner, is able to generate refined predictions in a coarse-to-fine manner without changing the original formulation. The discontinuous skeleton phenomenon is alleviated without using any post-processing.

2. Related Works

In recent years, most of the outstanding solutions in SkelNetOn competition were based on deep learning approaches, more specifically on dense prediction models, such as semantic segmentation. Some well-established segmentation models, e.g., FCN [14], U-Net [18], PSPNet [30], Deeplab series [2], HRNet [22], Segformer [27], etc., can be naturally applied to the skeleton extraction task. Among them, U-Net is the most commonly used [16, 17, 21]. The semantic features of shape images are relatively simple, and U-Net can maintain low-level semantic features by skip connection between encoder and decoder. Moreover, the model complexity of U-Net is relatively low and does not easily lead to overfitting for skeleton extraction tasks with limited training data. At the same time, the limited amount of data also makes the vision transformer model does not show competitiveness in skeleton extraction [21].
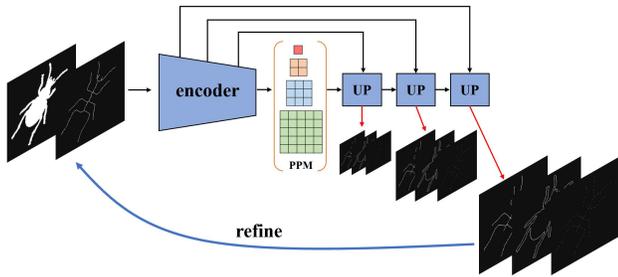
Figure 1. Illustration of the CAMION. Due to space constraints only three skip connections and up-sampling are drawn. PPM denotes pyramid pooling module.
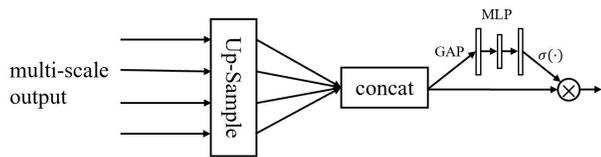
Figure 2. Ensemble attention module. "Up-Sample" module indicates up-sampling all input features to the original size.

Figure 3. The basic block of CAMION. GAP denotes global average pooling. MLP denotes multilayer perceptron.

Multi-task learning [25] allows the network to complete multiple tasks simultaneously and can also facilitate the performance of the main task through the learning of auxiliary tasks. [16] introduces auxiliary tasks in skeleton extraction but not multi-task. It embeds auxiliary heads in different layers of the decoder, but uses only the skeleton as the supervised signal, i.e., deep supervision. Different from [16], we really construct supervision from different tasks. Cascade networks can generate refined predictions in a coarse-to-fine manner and have competitive performance in fields such as semantic segmentation [3], object detection [1], and point cloud completion [26]. In the skeleton extraction task, the motivation is to generate fine and accurate skeletons, and adding the coarse outputs as input for the next level does not change the original formulation, so the same network can be used recursively to obtain finer skeletons.

3. Methods

3.1. Network Architecture

The skeleton extraction task can be viewed as a pixel-to-pixel dense prediction task, and U-shape networks can usually obtain competitive performance, as in some impressive skeleton extraction methods in recent years. Therefore, we construct a simple baseline with U-shape network. Initially, we use some mature backbone networks as encoder, e.g., ResNet [8], Efficientnet [24], Swin-transformer [11] and ConvNext [12], but unfortunately, their performance is inferior. This seems to be caused by the excessive down-sampling in their stem layers, the thin skeleton features

in deep decoder layers cannot effectively benefit from the high-resolution encoder features. Therefore, we manually design the encoder and decoder based on the available experience, as shown in Fig. 1. At a macro level, we design a basic U-shape network which performs skip-connections between the encoder and decoder at resolutions of (1, 1/2, 1/4, 1/8) of the original resolution. In our decoder, multiple auxiliary heads are designed to perform multi-scale deep supervision to speed up the convergence of the network.

At a micro level, we embed a concurrent spatial-channel attention mechanism in the basic block, referring to the concurrent spatial and channel Squeeze & Excitation (scSE) module [19], as shown in Fig. 3. To more adequately extract features at different scales and obtain more accurate global and local skeleton representations, we introduce the spatial pyramid pooling module (PPM) at the bottom of U-Net [3, 30]. Moreover, it is natural that for the multi-scale output of the feature pyramid in the decoder, we introduce an ensemble attention module that automatically accomplishes the integration of the multi-scale output and weights them in a learnable way [6].

Multi-task Learning for Skeletonization. As mentioned above, most of the deep learning-based approaches for skeleton extraction follow the shape-to-skeleton segmentation paradigm. Although the shape of the object directly contains the geometric information of a real object, it contains limited semantic information, i.e., without the texture features of the object. Therefore, we believe that introducing multi-task learning can help the network to obtain more general feature representation with limited input signal.

Specifically, for the skeleton extraction task, we introduced two auxiliary tasks: feature point detection [20] and
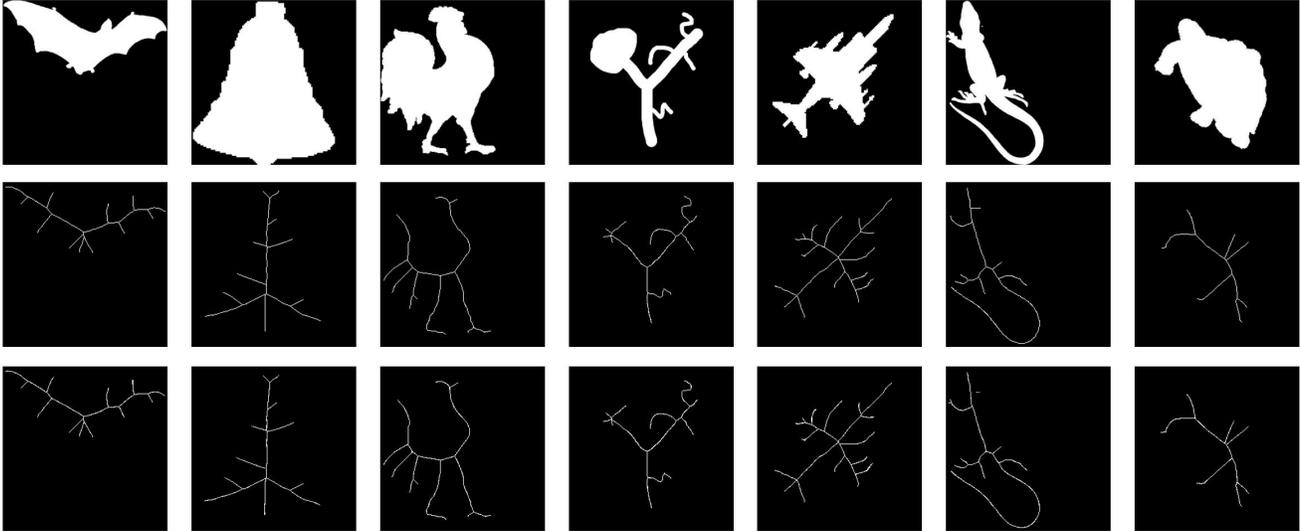
Figure 4. Visualization results. The first and second rows are the shapes and skeletons provided in Pixel SkelNetOn dataset. The third row is the results generated by CAMION.

contour extraction [23]. The motivation of using feature points as training signal is that we observe the skeleton in some non-trunk regions are frequently missed by the model, e.g., the fingertips of bats. The supervision from feature points is more like emphasis on making the model track the "good features", i.e., the corner points of the skeleton, and thus allowing the extraction of skeletal lines to benefit from these corner points. In other words, some of the skeletal lines that had been discarded are recovered. On the other hand, contour extraction can also bring improvement to the skeleton extraction, although this gain is relatively slight, but more importantly, it improves the robustness of skeleton extraction. The supervision of contours constrains the range of skeleton and can avoid false positive pixels to some extent. In CAMION, all tasks share the backbone network, and generate their own masks through several lightweight heads. The ground truth of feature point detection and contour extraction are generated by the algorithms from [20] and [23].

Cascade Network with Multi-input. Compared with traditional methods, deep learning-based skeleton extraction is more accurate and stable. However, we observed that some skeletal lines extracted using deep learning methods are discontinuous obviously, which impairs the recall of the skeletal pixels. Some post-processing algorithms can connect the discontinuous lines for better visual performance, but usually with loss of pixel accuracy. From the perspective of processing pipeline, cascade network scheme can alleviate this problem to some extent. Recently, some excellent cascade schemes have been used to optimize the network output of the previous step to get finer segmentation or detection. As shown in Fig. 1, we get the skele-

ton, feature points and contour of the shape image, through forward propagation. In the first stage, the skeleton is relatively coarse and discontinuous. Then, we concatenate the coarse skeleton with the original shape image and feed them into the network again. In this stage, we obtain a more refined skeleton than the previous one. After several refinements recursively, the final skeleton result is generated. For the skeleton input of the first stage, we use the ZHANGSUEN thinning algorithm, a traditional morphological method, which also allows our network to converge quickly in the early epochs [29].

3.2. Loss Function

Our optimization target is to minimize the overall loss, which can be formulated as:

$$\mathcal{L} = \mathcal{L}_s + \lambda_p \mathcal{L}_p + \lambda_c \mathcal{L}_c \tag{1}$$

where $\mathcal{L}_s$, $\mathcal{L}_p$ and $\mathcal{L}_c$ represent the supervised loss of skeleton extraction, feature point detection and contour extraction, respectively. λ_p and λ_c denote the weights of $\mathcal{L}_p$ and $\mathcal{L}_c$, and both are set to 0.1.

To alleviate the impact of sample imbalance, $\mathcal{L}_s$, $\mathcal{L}_p$ and $\mathcal{L}_c$ are all the hybrid loss, i.e., the combination of focal loss and dice loss:

$$\mathcal{L}_s = \mathcal{L}_{s\_focal} + \mathcal{L}_{s\_dice} \tag{2}$$

$$\mathcal{L}_p = \mathcal{L}_{p\_focal} + \mathcal{L}_{p\_dice} \tag{3}$$

$$\mathcal{L}_c = \mathcal{L}_{c\_focal} + \mathcal{L}_{c\_dice} \tag{4}$$

PPM	scSE	ensemble att.	F1-score
			0.7903
✓			0.7951
✓	✓		0.8023
✓	✓	✓	**0.8051**

Table 1. Ablation study of several modules of basic network.

skeleton	feature point	contour	cascade	F1-score
✓				0.8051
✓	✓			0.8162
✓	✓	✓		0.8188
✓	✓	✓	✓	**0.8285**

Table 2. Ablation study of multi-task learning and cascade manner.

The $\mathcal{L}_{focal}$ and $\mathcal{L}_{dice}$ are defined as:

$$\mathcal{L}_{focal} = \sum_i -\alpha(1-p_i)^\gamma log(p_i) - (1-\alpha)p_i^\gamma log(1-p_i) \tag{5}$$

$$\mathcal{L}_{dice} = 1 - \frac{2\sum_i y_i p_i + \epsilon}{\sum_i y_i + \sum_i p_i + \epsilon} \tag{6}$$

where α and γ denote the balanced variant and focusing parameter of the focal loss. y_i and p_i denote the target labels and predicted probabilities, respectively. ϵ is a small constant to avoid division by zero.

4. Empirical Evaluations on Pixel SkelNetOn

4.1. Dataset and Evaluation Metrics

The dataset provided by the CVPR 2022 "Deep Learning for Geometric Computing" workshop consists of 1725 binary images with a size of 256×256 pixels. Among them, 1218 images are used as the training set, 241 images as the validation set, and 266 images for testing. As shown in Fig. 4, in the shape images and the ground truth skeletons, the white area represents the foreground and the black area represents the background.

F1-score is used for quantitative metric evaluation and is expressed as:

$$F1 - score = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{7}$$

where the precision and recall are defined as:

$$precision = \frac{TP}{TP + FP} \tag{8}$$

$$recall = \frac{TP}{TP + FN} \tag{9}$$

TP, FP, FN respectively indicate the number of true positive, false positive and false negative.

4.2. Setting

We train CAMION for 240 epochs using SGD with a learning rate of 5e-2. There is a 30-epoch linear warm up and a cosine decaying schedule afterward. All the results

are obtained by training on single Tesla T4 GPU, with batch size set to 8.

For data augmentations, all possible flips and rotations multiple 90° are used. In addition, we randomly remove some lines from the input skeletons in order to better enable the cascade network to learn the potential task of "connecting discontinuous skeletons". In the specific implementation, we use a coarse Cutout [4] only for the skeletons, before feeding into the network.

4.3. Results

We split the original training set into a new training and testing set in the ratio of 4:1 to avoid overfitting the original validation set. Tab. 1 shows the ablation experiments of several modules of our basic network. Our U-Shape Network baseline can achieve 0.7903 F1-score. With the pyramid pooling module, our network benefits from the global shape information. The scSE attention module improves the performance to 0.8023 by adaptive adjustment of features on channel and spatial dimension, with a negligible increase in model complexity. The ensemble attention mechanism reweights multi-scale features in a learnable manner. And finally, our basic model achieves 0.8051 F1-score on the split-test data.

Tab. 2 presents the results after using the multi-task learning and cascade manner, where "cascade" indicates that only one additional cascade process was performed. The feature point detection task brings a surprising improvement for skeleton extraction, with the score increasing from 0.8051 to 0.8162. The improvement from the contour extraction task, although slight, further demonstrates that the multi-task learning is feasible in skeleton extraction. After using the cascade manner, CAMION achieves a performance of 0.8285 on the split-test data and 0.8289 on the official testing set (achieve state-of-the-art result compared to the results of previous Pixel SkelNetOn competitions), without using extra tricks, e.g., multi-model ensemble, pseudo label, exponential moving average (EMA), fine tuning of hyper-parameters, etc. The visualization results and descriptions are presented in Fig. 4 and Fig. 5.
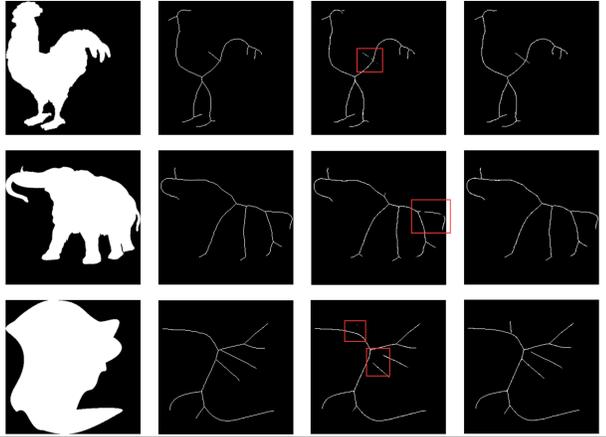
Figure 5. Visualization results of basic U-Net and CAMION. The first and second columns are the shapes and skeletons provided in Pixel SkelNetOn dataset. The third column is the results generated by our basic U-Net. The fourth column is the results generated by CAMION. Some discontinuous lines and false-negative pixels are indicated by red frames.

Mehond	Validation (242)	Test (266)
SkeletonNet [15]	0.7480	0.7711
Panichev et. al. [17]	0.7500	0.7846
Subpixel [5]	0.7708	-
CAMION	-	0.8289

Table 3. Comparison of results on Pixel SkelNetOn validation and testing data.

5. Conclusion

In this paper, we propose CAMION which demonstrates the feasibility of multi-task learning and cascade manners in skeleton extraction tasks. Furthermore, CAMION is potential and generalizable. For multi-task learning in CAMION, more relevant tasks can be introduced, such as the classification of shapes (categories are indicated in the image names). And the cascade manner in CAMION can be generalized to future models with better performance.

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 2

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1

[3] Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, and Chi-Keung Tang. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8890–8899, 2020. 2

[4] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 4

[5] Sohom Dey. Subpixel dense refinement network for skeletonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 258–259, 2020. 5

[6] Sheng Fang, Kaiyu Li, Jinyuan Shao, and Zhe Li. Snunet-cd: A densely connected siamese network for change detection of vhr images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021. 2

[7] Zicheng Guo and Richard W Hall. Parallel thinning with two-subiteration algorithms. *Communications of the ACM*, 32(3):359–373, 1989. 1

[8] 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. 2

[9] Meng Li, Howard Leung, and Hubert PH Shum. Human action recognition via skeletal and depth based feature fusion. In *Proceedings of the 9th International Conference on Motion in Games*, pages 123–132, 2016. 1

[10] Cheng Lin, Lingjie Liu, Changjian Li, Leif Kobbelt, Bin Wang, Shiqing Xin, and Wenping Wang. Seg-mat: 3d shape segmentation using medial axis transform. *IEEE Transactions on Visualization and Computer Graphics*, 2020. 1

[11] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[12] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022. 2

[13] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 143–152, 2020. 1

[14] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[15] Sabari Nathan and Priya Kansal. Skeletonnet: Shape pixel to skeleton pixel. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 5

[16] Nam Hoang Nguyen. U-net based skeletonization and bag of tricks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2105–2109, 2021. 1, 2

[17] Oleg Panichev and Alona Voloshyna. U-net based convolutional neural network for skeleton extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 5

[18] 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. 1

[19] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger. Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In *International conference on medical image computing and computer-assisted intervention*, pages 421–429. Springer, 2018. 2

[20] Jianbo Shi et al. Good features to track. In *1994 Proceedings of IEEE conference on computer vision and pattern recognition*, pages 593–600. IEEE, 1994. 2, 3

[21] Soonyong Song, Heechul Bae, and Junhee Park. Disco-u-net based autoencoder architecture with dual input streams for skeleton image drawing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2128–2135, 2021. 1

[22] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019. 1

[23] Satoshi Suzuki et al. Topological structural analysis of digitized binary images by border following. *Computer vision, graphics, and image processing*, 30(1):32–46, 1985. 3

[24] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 2

[25] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2021. 2

[26] Xiaogang Wang, Marcelo H Ang Jr, and Gim Hee Lee. Cascaded refinement network for point cloud completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 790–799, 2020. 2

[27] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[28] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1112–1121, 2020. 1

[29] Tongjie Y Zhang and Ching Y. Suen. A fast parallel algorithm for thinning digital patterns. *Communications of the ACM*, 27(3):236–239, 1984. 1, 3

[30] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 1, 2