UDC 004.93
DOI https://doi.org/10.32782/IT/2022-2-8

Kostiantyn KHABARLAK
Postgraduate Student, Assistant Professor at the Department of System Analysis and Control, Dnipro University of Technology, 19 Dmytra Yavornytskoho Avenue, Dnipro, Ukraine, 49005, habarlack@gmail.com
***ORCID:** 0000-0003-4263-0871*
***Scopus-Author ID:** 57219511187*

To cite this article: Khabarlak, K. (2022). Adaptyvna pislia navchannia merezha U-Net dlia zadachi sehmentatsii zobrazhen [Post-Train adaptive U-Net for image segmentation]. *Information Technology: Computer Science, Software Engineering and Cyber Security, 2,* 73–78, doi: https://doi.org/10.32782/IT/2022-2-8

POST-TRAIN ADAPTIVE U-NET FOR IMAGE SEGMENTATION

Many fields benefit from fast and accurate image segmentation. Convolutional neural networks show the best accuracy solving the task. Applications include medical or satellite imaging, autonomous driving, etc. Typical neural network architectures used for image segmentation are expected to be fully configured before the training procedure starts. To change the network architecture additional training steps are required. This is quite limiting as the network might not only be executed on a powerful server, but also on a mobile or edge device. Adaptive neural networks offer a solution to the problem by allowing certain adaptivity after the training process is complete.

In this work for the first time, we apply Post-Train Adaptive (PTA) approach to the task of image segmentation. We introduce U-Net+PTA neural network, which can be trained once, and then adapted to different device performance categories. The two key components of the approach are PTA blocks and PTA-sampling training strategy. The PTA blocks were added into the U-Net neural network with MobileNetV2 backbone. The post-train configuration can be done at runtime on any inference device including, but not limited to mobile devices. In addition to post-train neural network configuration, the PTA approach has allowed to improve image segmentation quality (Dice score) on the CamVid dataset. The final trained model can be switched at runtime between 6 PTA configurations. These configurations differ by inference time and quality. Importantly, all of the configurations have better quality than the original U-Net (No PTA) model. The possible future research direction is to expand the inference time difference between heavy and light PTA configurations to allow a single trained PTA-based network to target even more device performance categories.

***Key words:** Adaptive convolutional neural networks, image segmentation, inference speed, mobile computing, edge computing, computer vision.*

Костянтин ХАБАРЛАК
аспірант, асистент кафедри системного аналізу та управління, Національний технічний університет «Дніпровська політехніка», просп. Дмитра Яворницького 19, Дніпро, Україна, 49005, habarlack@gmail.com
***ORCID:** 0000-0003-4263-0871*
***Scopus-Author ID:** 57219511187*

Бібліографічний опис статті: Хабарлак, К. (2022) Адаптивна після навчання мережа U-Net для задачі сегментації зображень. *Information Technology: Computer Science, Software Engineering and Cyber Security, 2,* 73–78, doi: https://doi.org/10.32782/IT/2022-2-8

АДАПТИВНА ПІСЛЯ НАВЧАННЯ МЕРЕЖА U-NET ДЛЯ ЗАДАЧІ СЕГМЕНТАЦІЇ ЗОБРАЖЕНЬ

Багато застосунків потребують швидку та точну сегментації зображень, де згорткові нейронні мережі показують найкращу точність вирішення задачі. Застосування включають медичні або супутникові зображення, автономне водіння тощо. Зазвичай необхідно, щоб архітектури нейронних мереж, які використовуються для сегментації зображень, були повністю налаштовані до початку процедури навчання. Для зміни архітектури мережі необхідні додаткові ітерації навчання. Це є обмеженням, оскільки мережа може працювати не лише на потужному сервері, а й на мобільному чи крайовому пристрої. Адаптивні нейронні мережі пропонують вирішення проблеми, дозволяючи певну адаптацію після завершення процесу навчання.

У цій роботі вперше застосовано підхід Post-Train Adaptive (PTA) до задачі сегментації зображень. Представлено нейромережу U-Net+PTA, яку можна один раз навчити, а потім адаптувати до пристроїв

із різною обчислювальною швидкістю. Двома ключовими компонентами підходу є блоки PTA та стратегія навчання із випадковою вибіркою PTA конфігурацій. Блоки PTA було додано в нейромережу U-Net з мережею кодувальником MobileNetV2. Отриману мережу можна конфігурувати після навчання на будь-якому пристрої, включаючи мобільні. Також підхід PTA дозволив покращити якість сегментації зображення в наборі даних CamVid відповідно до метрики Dice. Навчену модель можна перемикати між 6 конфігураціями PTA навіть під час виконання. Ці конфігурації відрізняються часом роботи та якістю. Важливо, що всі конфігурації мають кращу якість, ніж оригінальна модель U-Net (без PTA). Можливим напрямок подальших досліджень є збільшення різниці в часі виконання між важкою та легкою конфігураціями PTA блоків, щоб дозволити одній навченій мережі на основі PTA націлюватися на ще більшу кількість пристроїв із різною обчислювальною потужністю.

Ключові слова: *адаптивні згорткові нейронні мережі, сегментація зображень, час виконання, мобільні обчислення, крайові обчислення, комп'ютерний зір.*

Introduction. Many fields benefit from fast and accurate image segmentation. Convolutional neural networks show the best accuracy solving the task. Applications include medical imaging (Ronneberger et al., 2015), autonomous driving (Brostow et al., 2009), satellite imaging (Hnatushenko et al., 2021), etc. Typical neural network architectures used for image segmentation are expected to be fully configured before the training procedure starts. To change the network architecture additional training steps are required. This is quite limiting as the network might not only be executed on a powerful server, but also on a mobile or edge device (Khabarlak, 2022a; Khabarlak & Koriashkina, 2022). Training separate networks for each device category is quite inefficient. Ideally, the network configuration change should be performed dynamically at runtime.

Adaptive neural networks offer a solution to the problem by allowing certain adaptivity after the training process is complete. Successful approaches to building adaptive neural networks have been proposed for Recurrent Neural Networks in (Graves, 2016), Convolutional Neural Networks in (Figurnov et al., 2017; Khabarlak, 2022b, 2022c). In particular, we see the Post-Train Adaptive approach proposed in (Khabarlak, 2022b) as an easy and effective way for the neural network adaptivity. Still the approach was only applied to the image classification task.

In this work we present U-Net+PTA network for the image segmentation task. We base upon U-Net (Ronneberger et al., 2015) architecture with MobileNetV2 (Sandler et al., 2018) backbone. To enable post-train adaptivity of the network, we apply the Post-Train Adaptive approach from (Khabarlak, 2022b).

To summarize, our main contributions are as follows:

1. We introduce U-Net+PTA neural network, which can be trained once, and then adapted to devices of different performance categories.

2. We demonstrate that U-Net+PTA not only improves inference speed over the U-Net, but also shows better $\text{Dice}_{\text{score}}$ on the CamVid (Brostow et al., 2009) dataset.

Literature Overview. Many fields benefit from fast and accurate image segmentation. To solve the segmentation task with high quality, the input image should be considered at multiple scales. This can be done through feature pyramid network (Lin et al., 2017), U-Net-like architecture (Ronneberger et al., 2015) or feature exchange between multiple scales (Sun et al., 2019). Such architectures are computationally intensive. In the meantime, segmentation algorithms are often required to run on desktop as well as mobile devices, while current architectures are mostly suited for desktop applications only.

Typically, neural network architectures made to solve the segmentation task are configured before the training process starts. Different backbones can be used in the segmentation models to change their quality and inference speed, like ResNet (He et al., 2016), MobileNetV2 (Sandler et al., 2018), MobileNetV3 (Howard et al., 2019), SeNet (Hu et al., 2018) or others. Still, these backbones cannot be additionally configured after the training process is complete, which limits their applicability to devices with different computational resources.

Dynamic neural networks is a promising research direction (Figurnov et al., 2017; Graves, 2016; Khabarlak, 2022b, 2022c). The goal is to allow the neural network to change its architecture depending on expected inference time or input data complexity. However, in many cases additional adaptivity comes at increased computational cost. Thus, inference time might not be smaller on average than that of a conventional static neural network. In contrast, in (Khabarlak, 2022b) Post-Train Adaptive approach was presented, which via a simple MobileNetV2 modification has allowed to reduce actual inference time. Importantly, the approach allows to reconfigure the neural network after the training process is complete. But, to the best of our knowledge, the approach has only been

applied to the image classification task, namely face anti-spoofing. In this work we adapt the Post-Train Adaptive approach to the task of image segmentation.

Materials and Methods. We base on the U-Net architecture with MobileNetV2 backbone. To make the constructed neural network dynamic, we use the approach proposed in (Khabarlak, 2022b), and add 3 Post-Train Adaptive (PTA) blocks to the network (as is shown in Fig. 1). In this work we include PTA blocks only in the U-Net encoder (backbone), leaving the decoder part intact.

A single PTA block has 2 branches:

– Light branch. Contains a single Inverted Residual block;

– Heavy branch. Contains two Inverted Residual blocks connected sequentially.

Each block can be dynamically configured to infer either branch exclusively, or both branches at the same time averaging the resulting feature maps.

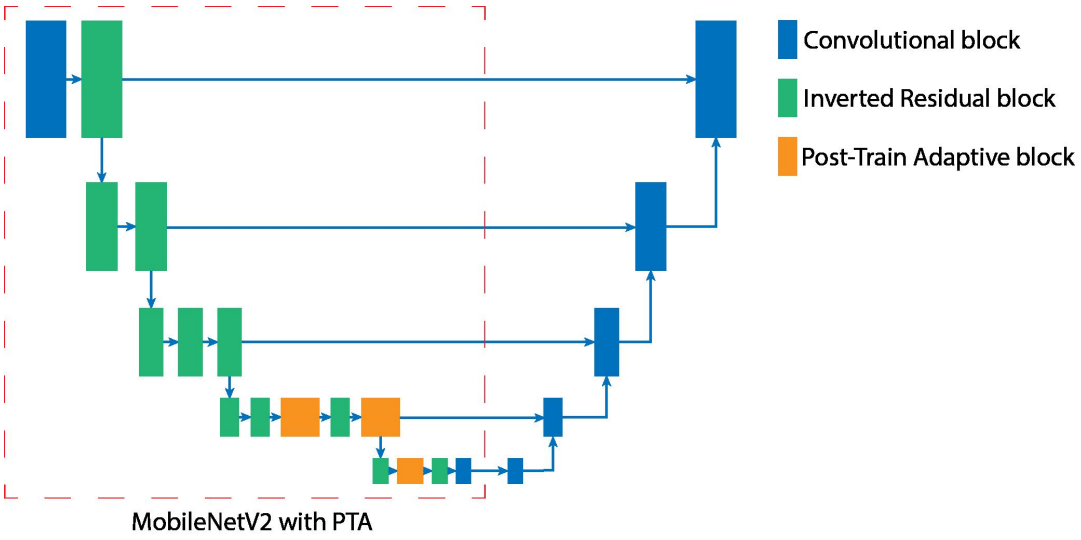

Figure 1. The proposed U-Net+PTA architecture for the image segmentation task. MobileNetV2 architecture with added PTA blocks is used as an encoder. Convolutional blocks are shown in blue, Inverted Residual in green, Post-Train Adaptive (PTA) in orange

To enable dynamic branch selection in the PTA block without retraining, a special PTA-sampling strategy is applied at training time. Specifically, several possible block configurations are selected randomly during the training procedure following the distribution shown in Table 1. All PTA block configurations that are possible, but not presented in the table are expected to be never sampled. Note, that configuration where both blocks are enabled at the same time is also never sampled.

Table 1.

U-Net+PTA train-time configuration sampling. Sampling strategy follows the original strategy from (Khabarlak, 2022b).

PTA Configuration	Sampling Probability
[Heavy, Heavy, Heavy]	0.45
[Light, Heavy, Heavy]	0.15
[Heavy, Light, Heavy]	0.15
[Heavy, Heavy, Light]	0.15
[Light, Light, Light]	0.10

To train the neural network, we use the $Dice_{loss}$ that has shown good segmentation training results, and to measure the resulting model quality, we use $Dice_{score}$(Milletari et al., 2016):

$$Dice_{loss} = 1 - \frac{2\sum_i^N p_i g_i + \epsilon}{\sum_i^N p_i^2 + \sum_i^N g_i^2 + \epsilon}, \quad (1)$$

$$Dice_{score} = \frac{2\sum_i^N p_i g_i + \epsilon}{\sum_i^N p_i^2 + \sum_i^N g_i^2 + \epsilon}, \quad (2)$$

where p_i is the predicted probability distribution, g_i is the ground true one-hot vector, N is the number of classes to distinguish between, ϵ is a small constant.

Experiments. To train and evaluate the model we use the widely known CamVid (Brostow et al., 2009) dataset. It contains images of size 480×360 pixels. The dataset is split into train (367 images), validation (101 images) and

test (233 images) subsets. All of the subsets have segmentation masks available of the same 480×360 size. The task is to learn the network to segment the images into one of the following classes: sky, building, pole, road, pavement, tree, sign symbol, fence, car, pedestrian, bicyclist, unlabeled.

For training and testing we resize the images into 256×256 size. To retain the original width to height ratio, the images are letterboxed. During training random crop and color jitter augmentations are used. Both U-Net and U-Net+PTA networks are trained for 600 epochs from scratch. No neural network pre-training is performed. Batch size is set to 8. Adam (Kingma & Ba, 2015) with the learning rate of $\alpha = 10^{-3}$ is used as an optimizer. The results are reported on the test set. NVIDIA GTX 1050Ti is used to train and test the model. In addition, we report inference time for a batch of 8 items. To ensure accurate time measurements, timings are averaged over 1000 batches. 95 % confidence interval is given for each measurement.

Results. In Table 2 we show model performance on the test set for the original U-Net model (denoted as No PTA) and the new U-Net+PTA model (denoted as PTA-*), where * is the PTA block configuration for inference. The best result is shown in red; the second best is in blue. As is clearly seen, all PTA-based configurations show better performance that the original U-Net. The best results are obtained by PTA-HLH, followed by PTA-BBB. Note, all PTA configurations have been obtained from a single model, trained only once. Thanks to the adaptive architecture, the exact configuration can be selected after the training is complete. Interestingly, PTA-HHH, which is equivalent in architecture to the No PTA model, but has been trained with PTA-sampling strategy is also better than No PTA configuration.

Table 2.

$Dice_{score}$ comparison for the U-Net+PTA model in the segmentation task on the CamVid dataset. Higher $Dice_{score}$ is better. The best configuration is highlighted in red; the second best is in blue. All Post-Train Adaptive (PTA) configurations show better quality than the original U-Net with MobileNetV2 backbone.

Configuration	$Dice_{score}$ (↑)
No PTA	0.8583
PTA-HHH	0.8666
PTA-LHH	0.8659
PTA-HLH	**0.8670**
PTA-HHL	0.8660
PTA-LLL	0.8647
PTA-BBB	**0.8667**

Table 3 shows model complexity and inference time comparison of the U-Net and the newly proposed U-Net+PTA models. The best result is shown in red; the second best is in blue. We show post-train model configuration, the number of model parameters, the number of multiplication and addition operations for inference, absolute and relative inference time. Relative time is computed with respect to the No PTA baseline. Inference time on actual device has some fluctuation due to GPU frequency change or sporadic system activity. To ensure accurate and consistent measurements, the inference time results are averaged across 1,000 measurements. Additionally, 95 % confidence interval is given. As can be seen, PTA-based models that have one or more Light branches enabled have faster than No PTA baseline inference.

Table 3.

Model complexity and inference time comparison of U-Net vs U-Net+PTA models. The following information is shown: post-train model configuration, the number of model parameters, the number of multiplication and addition operations for the inference, absolute and relative inference. Relative time is computed with respect to the No PTA model. The best result is shown in red; the second best is in blue. PTA-based models with Light branch show faster inference time.

Config	# Params (↓, M)	Multiply-Adds (↓, Mops.)	Inference Time (↓, ms)	Relative Impr. (%, ↓)
No PTA	6.63	871.80	81.37 ± 0.14	100.00
PTA-HHH	6.63	871.80	81.16 ± 0.13	99.75
PTA-LHH	6.58	868.58	80.21 ± 0.08	98.58
PTA-HLH	6.51	**864.16**	**79.78 ± 0.08**	98.05
PTA-HHL	**6.31**	866.65	79.89 ± 0.08	98.19
PTA-LLL	**6.14**	**855.49**	**78.82 ± 0.09**	96.86
PTA-BBB	7.12	888.12	83.79 ± 0.09	102.98

In Table 4 we show total training time and the best $Dice_{score}$ for the No PTA and U-Net+PTA models. The best $Dice_{score}$ is selected from all possible PTA configurations from a single training pass. Both models were trained 600 epochs. As can be seen, higher $Dice_{score}$ for the PTA-based model is achieved with slightly faster model training.

Table 4.

Training time and the best final score comparison for the U-Net and U-Net+PTA models. Higher $Dice_{score}$ for the PTA-based model is achieved with slightly faster model training.

Config	Total Training Time (↓, Min)	Best $Dice_{score}$ (↑)
U-Net	161.6	0.8583
U-Net+PTA	**158.6**	**0.8670**

Discussion. The Post-Train Adaptive method has been originally introduced for the task of face anti-spoofing in (Khabarlak, 2022b). In this work we have applied it to a different computer vision task, namely image segmentation. We based our approach on the U-Net neural network with the MobileNetV2 backbone. By adding PTA blocks to the U-Net architecture and following the PTA sampling training strategy, we have been able to successfully train the neural network. The resulting network can be trained once and reconfigured later. As can be seen from the Table 2, all of the PTA configurations show superior quality when compared to the U-Net with MobileNetV2 (No PTA). The best improvement is achieved by PTA-HLH ($Dice_{score}$ improvement of 0.0087), followed by PTA-BBB (+ 0.0084). The PTA-HHH configuration that is equivalent in architecture to the original No PTA model is also better than the No PTA configuration, which shows the benefit of the PTA-sampling training strategy. We also note that even the lightest PTA-LLL configuration is better than No PTA baseline (+ 0.0064).

The PTA-LLL configuration is the fastest configuration as is shown in Table 3. PTA-LLL model shows better $Dice_{score}$ and is 3.14 % faster. Also, the heaviest PTA-BBB model is only 2.98 % slower, while offering 0.0084 higher $Dice_{score}$. PTA-HLH has good speed and the best quality making it the best configuration in terms of speed to quality ratio.

We also note that the benefit of using 3 PTA blocks in the U-Net with MobileNetV2 backbone is smaller, than it was in the original PTA work. This can be explained by the fact that overall U-Net with MobileNetV2 backbone is a much larger model than the plain MobileNetV2 for classification as can be seen from Table 5.

Table 5.

Comparison of the number of model parameters and multiply-additions to perform classification (Class.) and segmentation (Segm.) tasks. Note, that for segmentation the baseline No PTA model is significantly larger.

Model	Task	# Params (M)	Multiply-Adds (Mops.)
MobileNetV2 No PTA	Class.	2.23	104.15
MobileNetV2 PTA-LLL	Class.	1.73	87.84
U-Net No PTA	Segm.	6.63	871.80
U-Net PTA-LLL	Segm.	6.14	855.49

The PTA training procedure is easy to integrate into existing pipelines. It offers the benefits of extra model configuration after the training is complete, higher model quality, and lower inference time. In addition to that, overall U-Net+PTA training time is no larger than that of a simple U-Net model as can be seen from Table 4.

Conclusions. In this work Post-Train Adaptive approach has been first applied to the task of image segmentation. The PTA approach has made it possible to reconfigure the architecture of the designed neural network after the training process has been complete. The two key components of the approach are PTA blocks and PTA-sampling training strategy. The PTA blocks were added into the U-Net neural network with MobileNetV2 backbone. The post-train configuration can be done at runtime on any inference device including, but not limited to mobile devices.

In addition to post-train neural network configuration, the PTA approach has allowed to improve image segmentation quality ($Dice_{score}$) on the CamVid dataset.

The final trained model can be switched at runtime between 6 PTA configurations. These configurations differ by inference time and quality. The best speed is offered by PTA-LLL configuration, that is faster and has higher quality than No PTA baseline. The best quality is achieved by PTA-HLH configuration with

better than No PTA inference speed making the best configuration in term of speed to quality ratio. Importantly, all of the configurations have better quality than the original U-Net (No PTA) model.

The possible future research direction is to expand the inference time difference between heavy and light configurations to allow a single trained PTA-based network to target even more device performance categories.

REFERENCES:

1. Brostow, G. J., Fauqueur, J., & Cipolla, R. (2009). Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, *30*(2), 88–97. https://doi.org/10.1016/j.patrec.2008.04.005

2. Figurnov, M., Collins, M. D., Zhu, Y., Zhang, L., Huang, J., Vetrov, D. P., & Salakhutdinov, R. (2017). Spatially adaptive computation time for residual networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017*, 1790–1799. https://doi.org/10.1109/CVPR.2017.194

3. Graves, A. (2016). Adaptive computation time for recurrent neural networks. *CoRR*, *abs/1603.08983*. http://arxiv.org/abs/1603.08983

4. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016*, 770–778. https://doi.org/10.1109/CVPR.2016.90

5. Hnatushenko, V. V., Zhernovyi, V., Udovyk, I., & Shevtsova, O. (2021). Intelligent system for building separation on a semantically segmented map. *Proceedings of the 2nd International Workshop on Intelligent Information Technologies & Systems of Information Security with CEUR-WS, Khmelnytskyi, Ukraine, March 24-26, 2021*, *2853*, 1–11. http://ceur-ws.org/Vol-2853/keynote1.pdf

6. Howard, A., Pang, R., Adam, H., Le, Q. V., Sandler, M., Chen, B., Wang, W., Chen, L.-C., Tan, M., Chu, G., Vasudevan, V., & Zhu, Y. (2019). Searching for MobileNetV3. *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 1314–1324. https://doi.org/10.1109/ICCV.2019.00140

7. Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 7132–7141. https://doi.org/10.1109/CVPR.2018.00745

8. Khabarlak, K. (2022a). Face detection on mobile: Five implementations and analysis. *CoRR*, *abs/2205.05572*. https://doi.org/10.48550/arXiv.2205.05572

9. Khabarlak, K. (2022b). Post-train adaptive MobileNet for fast anti-spoofing. *CEUR Workshop Proceedings*, *3156*, 44–53. http://ceur-ws.org/Vol-3156/keynote5.pdf

10. Khabarlak, K. (2022c). Faster optimization-based meta-learning adaptation phase. *Radio Electronics, Computer Science, Control*, *1*, 82–92. https://doi.org/10.15588/1607-3274-2022-1-10

11. Khabarlak, K., & Koriashkina, L. (2022). Fast facial landmark detection and applications: A survey. *Journal of Computer Science and Technology*, *22*(1), 12–41. https://doi.org/10.24215/16666038.22.e02

12. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio & Y. LeCun (Eds.), *3rd international conference on learning representations, ICLR 2015, san diego, CA, USA, may 7–9, 2015, conference track proceedings*. http://arxiv.org/abs/1412.6980

13. Lin, T.-Y., Dollár, P., Girshick, R. B., He, K., Hariharan, B., & Belongie, S. J. (2017). Feature pyramid networks for object detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017*, 936–944. https://doi.org/10.1109/CVPR.2017.106

14. Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, October 25-28, 2016*, 565–571. https://doi.org/10.1109/3DV.2016.79

15. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 – 18th International Conference Munich, Germany, October 5–9, 2015, Proceedings, Part III*, *9351*, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

16. Sandler, M., Howard, A. G., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018*, 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

17. Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019*, 5693–5703. https://doi.org/10.1109/CVPR.2019.00584