Medical & Biological Engineering & Computing (2024) 62:73–82
https://doi.org/10.1007/s11517-023-02881-4

ORIGINAL ARTICLE

Knowledge distillation for efficient standard scanplane detection of fetal ultrasound

Jacopo Dapueto[1] · Luca Zini[2] · Francesca Odone[1]

Received: 16 November 2022 / Accepted: 16 June 2023 / Published online: 1 September 2023

Abstract

In clinical practice, ultrasound standard planes (SPs) selection is experience-dependent and it suffers from inter-observer and intra-observer variability. Automatic recognition of SPs can help improve the quality of examinations and make the evaluations more objective. In this paper, we propose a method for the automatic identification of SPs, to be installed onboard a portable ultrasound system with limited computational power. The deep Learning methodology we design is based on the concept of Knowledge Distillation, transferring knowledge from a large and well-performing *teacher* to a smaller *student* architecture. To this purpose, we evaluate a set of different potential teachers and students, as well as alternative knowledge distillation techniques, to balance a trade-off between performances and architectural complexity. We report a thorough analysis of fetal ultrasound data, focusing on a benchmark dataset, to the best of our knowledge the only one available to date.

Keywords Standard scanplane detection · Fetal ultrasound · Knowledge distillation · Machine Learning

1 Introduction

Abnormalities are one of the leading reasons for perinatal mortality in both industrialized and developing countries [1], thus mid-trimester fetal ultrasound (US) scans are carried out to provide accurate diagnostic information for the delivery of optimized prenatal care with the best possible outcomes for mother and fetus. During the obstetric US examinations, standard planes (SPs, sectional images containing key anatomical structures) are selected with care to compute biometric measurements to evaluate fetal growth and congenital malformations. They are also used to estimate the pre-birth weight and the gestational age of a fetus [1]. SPs are identified following international guidelines, promoted by scientific committees, so that images are obtained following the same protocols, in theory with the purpose of guaranteeing repeatability and reliability.

In clinical practice, the SPs selection based on the abovementioned protocols is experience-dependent, cumbersome, and suffers from inter-observer and intra-observer variability. The situation is even more critical in some countries, especially in the developing world, where according to the World Health Organization (WHO) individuals with no formal training carry out ultrasound scans [1]. Hence, the *automatic identification of SPs* will facilitate more objective evaluations and overall workflow improvement.

The task is challenging for several reasons, we highlight two important ones: *first*, the properties of US images [3]: low contrast, low signal-to-noise ratio, non-uniform acoustic densities and the presence of scattered noise. In addition, due to the high intra-class and low inter-class variations of US images, often non-SPs images are very similar to SPs [4, 5], so they are difficult to distinguish.

Second, in the identification of SPs of a given anatomical district, one should consider their difference with non-standard planes (non-SPs) of the same district may be minute — see Fig. 1. Indeed, SPs detection is different from the related task of US classification, since different images of

Luca Zini and Francesca Odone contributed equally to this work.

✉ Francesca Odone
francesca.odone@unige.it

Jacopo Dapueto
jacopo.dapueto@edu.unige.it

Luca Zini
luca.zini@esaote.com

[1] MaLGa-DIBRIS, Università degli Studi di Genova, Genova, Italy

[2] Esaote S.p.A, Genova, Italy

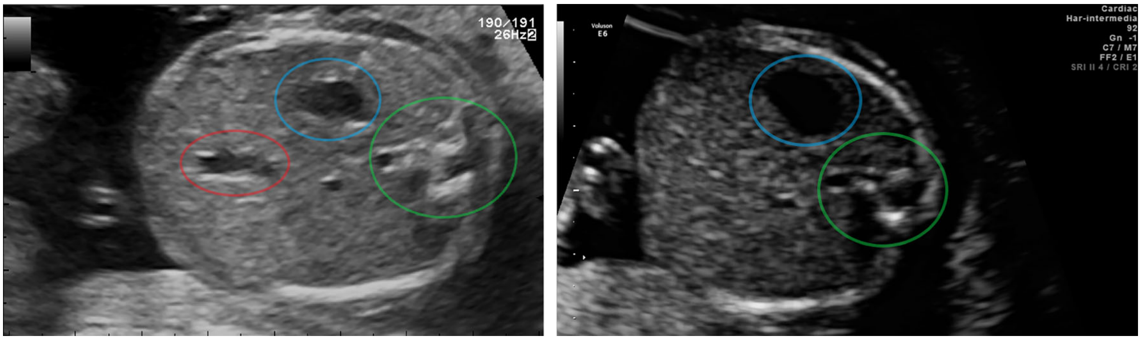


Fig. 1 Examples of fetal *abdomens*, from [2]. Left: a standard plane compliant to fetal anomaly screening programme [1]: it includes the spine (*green*), the umbilical vein (*red*), and the stomach (*blue*). Right: a non-standard one, where the umbilical vein is missing

the same anatomical district may be classified differently (standard or non-standard), while the goal of US classification is to associate an anatomical label to each example.

In this paper, we address the above-mentioned challenges by adopting a data-driven approach. Since the field is also suffering from a lack of data, data-driven solutions reduce the challenges related to data quality and variability, but raise a *third* issue: because of the legal rights to protect the privacy of patients, the example collection and annotation workload make data collection hard or sometimes even impossible. For this reason, medical images are valuable and often kept private: the literature is often reporting results based on private datasets [5–9], with a negative impact on methods benchmarking and reproducibility. In our work, we rely on the FETAL_PLANE_DB[1], to the best of our knowledge, the first publicly available dataset of fetal US scanplanes [2].

Specifically, we focus on the problem of designing a computationally efficient network for SPs detection. Indeed, as a main functional requirement of our research, we require the network to be able to run onboard portable ultra-sound scanners, whose computational power is limited and fully dedicated to signal acquisition and processing. Portable units are cheaper and more flexible than traditional ultrasound machines and more appropriate for the market of developing countries, where automatic analysis functionality is particularly needed.

Following previous works in the literature, we formulate SPs detection as a multi-class classification task and investigate the usage of Knowledge Distillation (KD) [10] to keep the computational cost under control, transferring the generalization ability of a high-performing network (called *teacher*) into a second one (called *student*), with a two-fold benefit (1) reducing the dimension and the complexity of the model in terms of the number of layers and parameters; (2) mitigating the over-fitting effect due to the limited amount of training data. We propose a simple strategy distilling knowledge from Logits and from three hidden layers (see Fig. 2), resorting to a hand-crafted association.

To summarize, the *paper contributions* are as follows:

- A simple yet effective knowledge distillation strategy, based on a loss that combines a response-based and a feature-based distillation.
- A thorough evaluation of candidate choices for teacher and student models, taking into account a budget on data availability due to the specific application and inference time performance limits guided by the target portable US system.
- An efficient pipeline for standard SPs detection we assess on a fetal image classification benchmark to provide measurable evidence of the potential of the approach in the clinical domain.

2 Methods

2.1 Related works

SPs detection. With the development of deep learning, general-purpose CNNs have been applied in many medical applications, including US classification [2, 11, 12]. In the case of US classification, a recent study assesses the performances of several architectures [2], showing the potential of large CNNs in this domain. It does not discuss CNNs limitations due to the stringent computational requirements of the device, and the limited availability of data, and this lack of information motivates our analysis, while the architectures considered by the authors are the candidates for our teacher models.

Other works learn the relationship between 2D plane images and 3D volumes [13], exploit the temporal relation between the frames in video clips [11, 14], other completely based their approach on the identification of the ASs [7]. The limited availability of data in the field, pushed the researchers to improve the generalization ability of the

[1] https://doi.org/10.5281/zenodo.3904280 Version: 1.0.

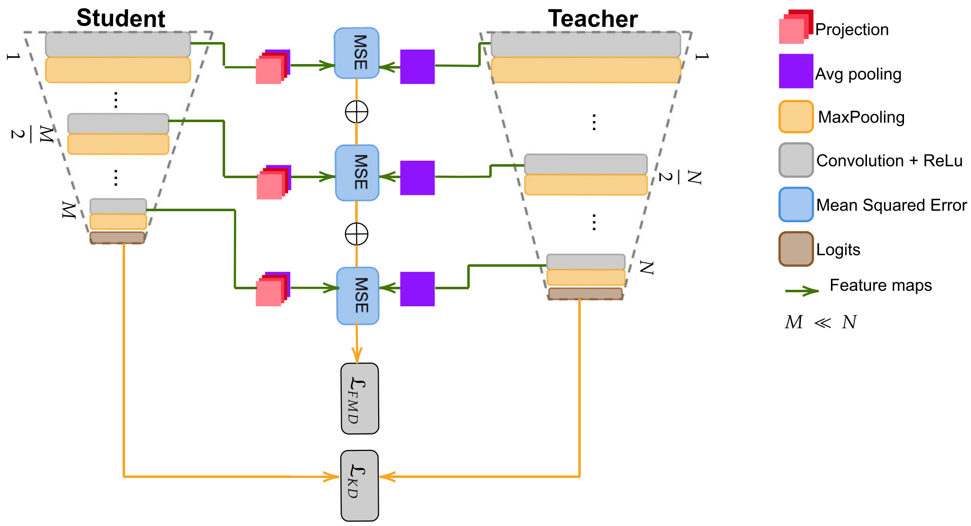


Fig. 2 A sketch of the proposed knowledge distillation scheme, involving three convolutional layers of the student and teacher networks, and the final logits. Average pooling is employed to align the shapes of teacher and student feature maps, and a projection of student feature maps is learnt to reduce the gap between the capability of the two networks

models with different strategies, for instance designing ad hoc features[15], adopting multi-task learning classifying and localizing ASs [8, 16], and including clinical prior knowledge [8, 16] increasing the burden of data acquisition and annotation, or integrating an *attention mechanism* so that the predictions are performed thanks to local features at different scales [9].

Fewer, *ad hoc* architectures are also available in the literature: SonoNet is a family of CNNs specifically designed to solve the detection of SPs in real-time during the US scanning [5].

Knowledge distillation. KD is a class of methods from the deep learning literature, whose goal is to control the complexity of a network by distilling knowledge from a large teacher model to a smaller student model mainly comparing either the logits [17] or the outputs of the hidden layers of both models [18–21], after the distillation only the student will be used for inference. In the literature, hidden layer associations are either done by hand before training [18–20] or learnt during the optimization of the student [21]. As it will be clarified in the following, we adopt the former approach.

Knowledge distillation has been previously applied to SPs detection in ultrasound imaging to classify echo cine series into 12 standard views [22], and to detect the abdomen, femur and head from fetal images [6]; both methods proposed memory-efficient student models, with no specific focus on the inferring time. Both approaches take into account a richer input, with respect to our work: [22] analyze US clips, while [6] complement US images with relevant information acquired by an eye tracker.

2.2 Proposed methodology

We formalize SPs detection as a multi-class classification task, where each image is associated with one of *n* predefined classes and the set of classes may include SPs and non-SPs of the same anatomical district, and for this reason, some classes may be very similar to one another.

The task is complex and fine-grained, for this reason, small and specific architectures may not be sufficient. At the same time, large architectures may not be applicable, to the stringent computational requirements. We adopt a KD strategy, where we distil the knowledge from a more complex general purpose *teacher model* to a simpler more specific *student model*.

In the method we propose, we combine two alternative KD approaches: a first approach, BasicKD, using the output of the last layer before the softmax, called *Logits*, so the student is trained to produce Logits equal to those generated by the teacher [17].

A second intermediate KD considering the outputs of the hidden layers of the CNN, comparing the feature maps generated by the student with those of the teacher, involving compatible layers, otherwise the optimization will lead to suboptimal solutions. Here, we follow a hand-crafted association, motivated by the fact, we are interested in distilling knowledge between architectures with the same hierarchical structure, including layers with similar semantic meanings.

Figure 2 provides a sketch of the methodology we propose.

The distillation loss we propose combines a response-based and a feature-based part, the hyperparameter β controls the importance of features:

$$\mathcal{L} = \mathcal{L}_{KD} + \beta\mathcal{L}_{FMD}. \qquad (1)$$

$\mathcal{L}_{\mathbf{KD}}$: it is a loss computed on the Logits, as the linear combination of two components, whose relative weight is controlled by a hyperparameter $\lambda \in [0, 1]$.

$$\mathcal{L}_{KD} = \lambda\mathcal{L}_{SL} + (1 - \lambda)\mathcal{L}_{logits} \qquad (2)$$

The mismatch between the output of the student model and the ground-truth label is computed with the function H representing the cross-entropy loss:

$$\mathcal{L}_{SL} = H(softmax(a_s), y_r) \tag{3}$$

a_s is the student network logit and y_r the ground-truth label.

The difference between the probability distributions generated by the *softmax* is computed with the *Kullback–Leibler* (*KL*) divergence loss.

$$\mathcal{L}_{logits} = \tau^2 KL(softmax(\frac{a_t}{\tau}), softmax(\frac{a_s}{\tau})) \tag{4}$$

A hyperparameter τ, called *temperature*, controls the softening of the signal arising from the output of the teacher network; a_t and a_s are the logits of the teacher and student networks, respectively. More details on this loss can be found at [17] (Sec. Distillation) where the Cross-Entropy loss is adopted, but it has been proved it is equivalent to the KL divergence ([23] Sec. 2).

$\mathcal{L}_{\mathbf{FMD}}$: it is computed on the intermediate feature maps, as a *hand-crafted* version of the adaptive method proposed in [21]. As shown in Fig. 2 the knowledge transfer is performed taking into account only three convolutional layers with *RELU* activation function for the student and for the teacher: the *first*, the *intermediate*, and the *last* layers of the teacher as associated with the corresponding layers of the student, through an appropriate feature map loss, combining Mean Square Error (MSE) losses:

$$\mathcal{L}_{FMD} = MSE_1 + MSE_2 + MSE_3. \tag{5}$$

Since the two networks should be of different size scales and based on different architectures, in order to align the shapes of the features between the teacher and the student, an average pooling [24] is performed to both student and teacher's features map. A projection including a stack of convolutional layers of the size 1 × 1, 3 × 3, and finally 1 × 1 is further applied to the student's features to reduce the gap between capabilities of the networks.

2.3 Implementation details

To estimate the above-mentioned hyperparameters, we perform *grid search* on the following ranges: $\lambda \in \{i(0.1) \mid i \in \mathbb{Z}, 1 \leq i \leq 9\}$, $\tau \in \{i \mid i \in \mathbb{Z}, 1 \leq i \leq 10\}$, $\beta \in \{1, 5, 10, 100, 200, 400, 800\}$, weight decay regularization $wt \in \{5e^{-5}, 5e^{-6}, 1e^{-5}, 1e^{-6}, 1e^{-7}, 1e^{-8}\}$, learning rate $lr \in \{0.01, 0.001, 0.0001, 0.00005\}$.

We train the models with 10% of the training set as a validation, and the models achieving maximum validation accuracy, in a given epoch, are selected.

Because of the limited size of the datasets available, we adopt the following data augmentation:

1. Randomly *flipped horizontally*, with a probability of 50%;
2. Randomly *flipped vertically*, with a probability of 50%;
3. *Gamma Correction* with randomly selected gamma between 0.3 and 1.7;
4. Normalized so that the values of the pixels range from 0 to 1;
5. *Zoom* with a random factor between 80% (zoom out) and 120 (zoom in);
6. Random rotation in a range between −10° and 10°.

Gamma correction is performed to mimic different gain settings used during the examination to overcome US attenuation and different tissue echogenicity, while the other transformations encode the possible settings the sonographer could choose to visualize US images on the US scanner and the possible fetal position in the placenta.

The tables report the *validation accuracy* (VA) used for the model selection, and the *balanced* validation/test accuracy (VAB, TAB). The *balanced* accuracy is computed as the mean of the diagonal of the *confusion matrix* and it is not considered during the optimization phase.

As teachers, we considered state-of-the-art architectures of different kinds, like DenseNet and ResNet, characterized by the potential to reach high accuracy but a huge amount of parameters and layers. For the student, we choose the SonoNet family, which is characterized by the absence of fully-connected layers, replaced with convolutions also for the final prediction before the softmax.

3 Results

In this section, we report and discuss the results we obtain on the FETAL_PLANE_DB, benchmark, available online [2].

The dataset is composed of over 12400 images from 1792 patients, acquired on 6 different machines by several different operators with similar experience. They are divided into 6 classes (Table 1): four of the most widely used fetal anatomical planes (**abdomen**, **brain**, **femur** and **thorax**), the mother's **cervix** (used for prematurity screening) and a general category to include any **other** less common image planes (not necessarily fetal or maternal), further dataset statistics can be found in [2] (Table 1). The authors already divided into the *training* and *test subsets* to simplify the comparison with the state-of-the-art, containing 7129 and 5271 images, respectively. The *validation* is 10% of the training set, as in [2].

Table 1 FETAL_PLANE_DB: cardinality of the entire dataset

Anatomical plane	N. patients	N. images
Fetal abdomen	595	711
Fetal brain	1082	3092
Fetal femur	754	1040
Fetal thorax	755	1718
Maternal cervix	917	1626
Other	734	4213

3.1 Choice of a teacher model

We first aim to find the best performing existing model on the data in hand, to be the *teacher* model.

We start by reproducing the results presented in [2]: in Table 2, we report our performances, higher than the ones in [2] (average improvement of 2.13%) as we change *augmentation* and make the learning rate lr and the weight decay wd parameters *decay exponentially*, starting from an initial value $5e-5$ as weight decay and $1e-4$ as learning rate, updating every 100 steps and a decay rate of 0.9. The number of layers and parameters clearly state these architectures are not appropriate for the task at hand; 0.25-MobileNet is suitable in terms of the number of parameters but its performances are not sufficient for our purpose. DenseNet169 remains the one achieving the best performances, as reported in [2], and for this reason, we consider it a valid teacher model.

We include in our analysis an architecture from the SonoNet family, SonoNet64, which has a comparable size and performance to DenseNet169. It will also be considered in the following, as a teacher candidate.

3.2 Choice of a student model

The smaller architectures from Table 2 do not appear to be suitable student models because the inferring time is still quite high. We evaluate a selection from the family of SonoNet architectures instead, training them with a fixed learning rate of $1e-3$, a batch size of 16 and ADAM optimizer. The results are in Table 3. While the dataset is rich enough for bigger networks to get better performances, here, we are interested in smaller networks in terms of layers (to decrease the time of the prediction) and in terms of filters (i.e., number of parameters). We select SmallNet32 as our student model, and baseline for the forthcoming experiments because it represents a trade-off between size (w.r.t. parameters and layers), accuracy and computational performances.

3.3 Knowledge distillation

Table 4 reports the results comparing our approach to KD with BasicKD [17] and SemCKD [21]. The best performances are achieved with DenseNet169 as teacher and our hand-crafted KD strategy. In the case of BasicKD the validation accuracy reaches 97.15%, also surpassing the performance of the teacher, but the same behavior is not appreciated in the balanced validation accuracy. SemCKD is instead ineffective, with a degradation of the results w.r.t. the baseline student model. It appears that the automatic association of the layers is possibly too challenging (the attention mechanism needs to learn additional parameters) for the limited amount of training data available. Moreover, the teacher and student architecture structures are very similar, therefore the data-driven association of SemCKD appears to be unnecessary. Instead, our hand-crafted association reports very promising performances, both in terms of validation and

Table 2 FETAL_PLANE_DB: state-of-the-art CNNs assessed as *teacher models* (number of Layers and Total parameters from [25])

Network	Layers	Total parameters	Sec/Frame (ms)	VA	VAB	TAB
VGG16 [26]	21	14,717,766	14.9	95.17%	93.74%	93.35%
MobileNet [27]	88	3,235,014	35.4	94.60%	94.88%	93.55%
0.75-MobileNet	88	1,837,590	36.0	94.31%	93.66%	93.59%
0.25-MobileNet	88	220,086	35.6	87.50%	81.77%	83.94%
Inception-v3 [28]	313	21,815,078	104.0	95.88%	95.75%	95.02%
ResNet50 [29]	177	23,600,006	71.9	95.17%	95.52%	94.75%
ResNet101 [29]	347	42,670,470	132.4	96.30%	95.68%	94.69%
ResNet152 [29]	517	58,383,238	196.5	95.31%	94.83%	94.45%
DenseNet121 [30]	429	7,043,654	145.4	96.02%	95.89%	94.66%
DenseNet169 [30]	597	12,652,870	204.7	96.44%	96.27%	95.13%
SonoNet64	53	14,864,350	25.8	96.59%	96.17%	95.10%

Seconds required for the forward stage through the network for 1000 samples and a batch size of 1 with Quadro RTX 5000

Table 3 FETAL_PLANE_DB: candidate *student* models assessment

Network	Layers	Total parameters	Sec/Frame (ms)	VA	VAB	TAB
SmallNet16	18	71,334	9.4	95.88%	93.60%	90.11%
SmallNet32	18	282,950	9.4	95.59%	93.43%	91.29%
SonoNet8	53	234,966	25.3	96.16%	95.14%	93.48%
SonoNet16	53	933,646	25.4	96.44%	93.82%	93.23%
SonoNet32	53	3,722,238	25.1	95.59%	95.15%	91.77%

testing accuracy, with an increase of the balanced accuracy of SmallNet32 of about 1.73% with respect to the reference student model.

We also report the performance of SonoNet32 obtained with BasicKD, leading to higher test accuracy. However, this model cannot be a suitable candidate because its dimensionality is incompatible with our needs (in particular the number of parameters).

To support the choice of a three-tier association of our hand-crafted approach, we perform the grid-search with 2 and 4 correspondences, where the Teacher is DenseNet169 and while the Student is SmallNet32. In this case

- 2-layers associates the *first* and the *last* layers;
- 4-layers associates the first, the *median*, the last layer, plus a layer in the middle of the second half of the model.

We report the results of the selected models in Table 5 where we notice the three-tier association is the one achieving the best performances for all the metrics. Considering the size of the Student, it is not appropriate to associate a higher number of layers.

For an interpretation of the KD benefits, we compare the GradCam [31] heatmaps of the SmallNet32 model from Table 3 with the heatmaps of the best model in Table 4 (thus, we interpret the results before and after knowledge distillation). Notice that, even though SonoNet has been designed with a weakly supervised localisation embedded into the architecture, here we, use a generic GradCam for an effective comparison between the two outputs.

Figure 3 compares the same abdomens generated from the baseline model and the distilled one. The abdomen plane is denoted by the presence of the spine, the umbilical vein and the stomach (Fig. 1). On the first row (prior KD), the model focuses on one structure only (either the vein or the stomach). On the second row (after KD) the model captures both the umbilical vein and stomach bubble. Both models never take the spine, possibly because the spine is always present in abdominal images. The distilled model seems to rely on wider regions and it allows to generate features maps with a higher discrimination power.

Finally, Fig. 4 shows the confusion matrices obtained by SmallNet32, again before and after knowledge distillation. Each row represents the instances in a ground-truth class, while each column represents the instances in a predicted class. Most misclassifications involve the class *Other*, which includes non-SPs of several anatomical districts. The errors are mitigated by knowledge distillation.

4 Discussion

The experimental analysis we carried out confirms that, with an appropriate choice of a distillation strategy, we may transfer knowledge from larger well-performing networks to smaller ones. The latter can be installed on a portable device, allowing us to achieve high efficiency (since they rely on few parameters and require a low inference time) while maintaining a good detection performance.

We achieve an *inference time improvement* with our best distilled student (SmallNet32) requiring $\frac{1}{21}$ of the time needed by the reference teacher network, DenseNet169 (while the model proposed in [22] is only $\frac{1}{6}$). As for the

Table 4 FETAL_PLANE_DB, dataset: knowledge distillation. *Improvement*: difference of balanced test accuracy with respect to the corresponding baseline student model. The distillation may improve (↑) or worsen (↓) the performances

Teacher	Student	KD Method	VA	VAB	TAB	Improvement
DenseNet169	SonoNet32	BasicKD	97.30%	95.76%	93.89%	↑(2.12%)
DenseNet169	SmallNet32	BasicKD	97.15%	94.13%	92.05%	↑(0.76%)
SonoNet64	SmallNet32	BasicKD	97.15%	96.44%	92.77%	↑(1.48%)
DenseNet169	SmallNet32	SemCKD	94.46%	93.01%	89.27%	↓(2.02%)
SonoNet64	SmallNet32	SemCKD	96.02%	95.14%	91.58%	↑(0.29%)
DenseNet169	SmallNet32	Ours	96.87%	96.39%	93.02%	↑(1.73%)
SonoNet64	SmallNet32	Ours	96.73%	97.03%	92.80%	↑(1.51%)

Table 5 FETAL_PLANE_DB, dataset: knowledge distillation with hand-crafted association. The column *Associations* reports the number of layers of the teacher and student used to perform the distillation

Teacher	Student	Associations	VA	VAB	TAB
DenseNet169	SmallNet32	2	95.59%	95.66%	92.88%
DenseNet169	SmallNet32	3	96.87%	96.39%	93.02%
DenseNet169	SmallNet32	4	94.74%	94.09%	92.13%

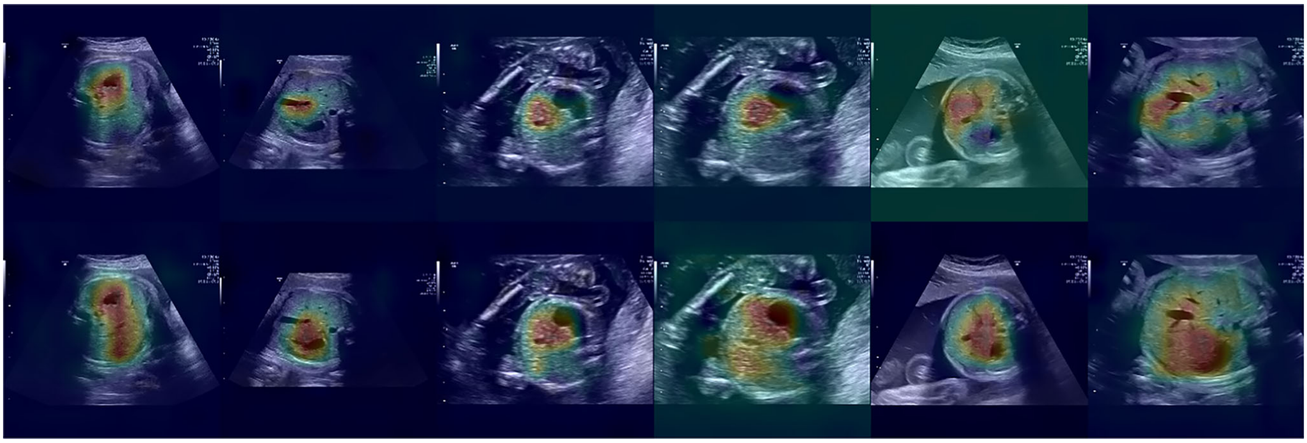
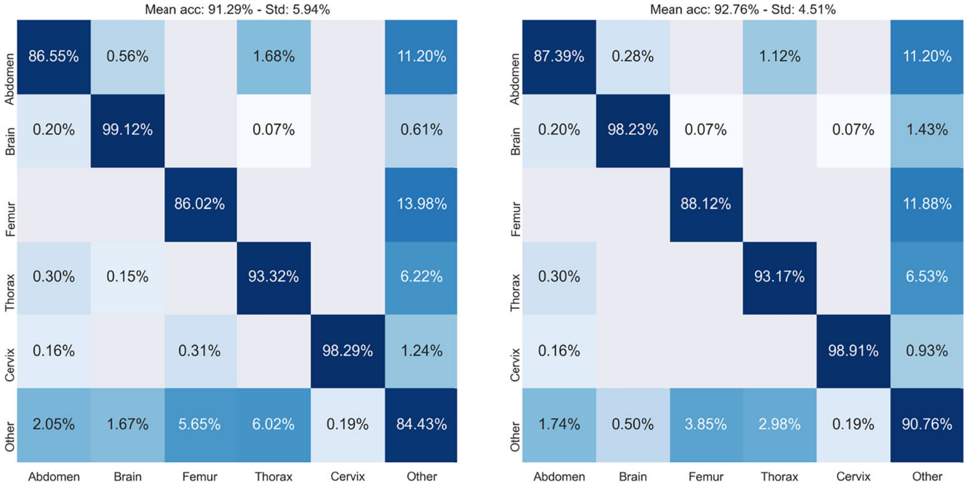


Fig. 3 Heatmaps obtained via gradcam of 6 random abdomen samples from FETAL_PLANE_DB, test set [2], before (top) and after (bottom) KD (see text)

Fig. 4 Confusion matrix of distilled SmallNet32 on the FETAL_PLANE_DB, test set [2]. Left: before the knowledge distillation. Right: after distillation. The colours scale is logarithmic to better highlight the misclassifications



space occupancy of the student network, it contains about $\frac{1}{22}$ of the teacher's parameters and $\frac{1}{33}$ layers.

Moreover, with our hand-crafted distillation strategy, we obtain a *test accuracy improvement* with respect to the student model prior to distillation The improvement mainly consists of a reduced misclassification on the class "Other", the most challenging one because of its internal variability and its appearance overlap with several classes. This is due to the way the dataset was built [2], since only images complying with minimum quality requirements were selected by the clinicians, while low quality and inappropriate anatomical planes (cropped or badly taken) were labelled as *Other*, together with images containing calipers.

Apart from "Other", the most challenging class that received a benefit from the distillation procedure is "Abdomen", where most of the errors are mis-classifications with "Thorax" (see Fig. 4). Besides this, we could not appreciate any interpretable pattern in our errors.

It should be noted that SPs detection would be better addressed if we dispose of a dataset with several non-standard classes, one per anatomical district of interest. Unfortunately, to date, a dataset with these characteristics is not available.

If we compare the results of the student after distillation with the original teacher model, we notice a small performance degradation of 2.11% with SmallNet32, while SonoNet32 (Table 4) architecture, whose inference time is $\frac{1}{8}$ of DenseNet169, leads to a higher test accuracy just with BasicKD, but a much higher dimensionality with respect to SmallNet32. Notice that classical CNNs like VGG16 have a similar inference time, a much larger spatial occupancy (they have a very high number of parameters, higher than DenseNet169), and lower performances.

5 Conclusions

In this paper, we addressed the detection of SPs, with the objective of implementing the method onboard of a portable US scanner. First, we explored the performances of general state-of-the-art CNNs but the design of such networks is incompatible with the system's computational requirements. The family SonoNet, explicitly designed for the detection of SPs, was tested looking for a small and fast but effective combination. The obtained results are encouraging, but the performances of the smallest models (e.g., SmallNets) are not competitive with the state of the art. To reduce this gap, we identified knowledge distillation strategies: KD allows to reduce the size of the network while maintaining satisfactory performances. We transferred the capabilities from a teacher to a student thanks to the logits and features maps of intermediate convolutional layers; we found out a manual association leads to better results, possibly because the available data are not enough to learn an automatic and optimal association. The best solution we obtained is a distilled SmallNet32, leading led to the best trade-off between speed and performance on the benchmark dataset.

An interesting direction for future works would be to further push the performances of the models with additional expert knowledge, similarly to [6]. The outcome of this research will be the development of software tools to help inexperienced operators with US acquisition tasks, highlighting SPs of interest as the examination takes place and recording them automatically during the exam. From a practical point of view, it would be highly beneficial to obtain this automatic classification on board the ultrasound device, as a guideline to the operator during scanning. The outputs would also be available for an offline analysis carried out by an expert ultrasonographer.

Author Contributions All authors contributed to the study's conception and design. Methodology design: Luca Zini, Jacopo Dapueto; implementation: Jacopo Dapueto; experimental design: Jacopo Dapueto; experimental analysis: all authors; original draft preparation: Jacopo Dapueto, Francesca Odone; review and editing: all authors.

Funding Open access funding provided by Universitá degli Studi di Genova within the CRUI-CARE Agreement.

Data Availability FETAL PLANE DB is a publicly available dataset. For the ethics declaration, we refer to the original datasets publication.

Declarations

Conflict of interest The authors declare no competing intersts.

References

1. Salomon, LJ, Alfirevic, Z, Berghella, V, Bilardo, C, Hernandezandrade, E, Johnsen, SL, Kalache, K, Leung, K-y, Malinger, G, Munoz, H, Prefumo, F, Toi, A (2010) Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. Ultrasound Obstet Gynecol. https://doi.org/10.1002/uog.8831
2. Burgos-Artizzu X, Coronado-Gutiérrez D, Valenzuela B, Bonet-Carne E, Eixarch E, Crispi F, Gratacós E (2020) Evaluation of deep convolutional neural networks for automatic classification of

common maternal fetal ultrasound planes. Sci Rep 10. https://doi.org/10.1038/s41598-020-67076-5

3. Noble JA, Boukerroui D (2006) Ultrasound image segmentation: a survey. IEEE Trans Medical Imaging 25(8):987–1010. https://doi.org/10.1109/TMI.2006.877092
4. Yaqub, M, Kelly, B, Papageorghiou, AT, Noble, JA (2015) Guided random forests for identification of key fetal anatomy and image categorization in ultrasound scans. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 687-694
5. Baumgartner CF, Kamnitsas K, Matthew J, Fletcher TP, Smith S, Koch LM, Kainz B, Rueckert D (2017) Sononet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE Trans. Medical Imaging 36(11):2204–2215. https://doi.org/10.1109/TMI.2017.2712367
6. Patra A, Cai Y, Chatelain P, Sharma H, Drukker L, Papageorghiou AT, Noble JA (2019) Efficient ultrasound image analysis models with sonographer gaze assisted distillation. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, Yap P-T, Khan A (eds) Medical Image Computing and Computer Assisted Intervention - MICCAI 2019. Lecture Notes in Computer Science. Springer, Cham, pp 394–402. https://doi.org/10.1007/978-3-030-32251-9_43
7. He S, Lin Z, Yang X, Chen C, Wang J, Shuang X, Deng Z, Liu Q, Cao Y, Lu X, Huang R, Ravikumar N, Frangi A, Zhang Y, Xiong Y, Ni D (2021) Statistical dependency guided contrastive learning for multiple labeling in prenatal ultrasound. In: Lian C, Cao X, Rekik I, Xu X, Yan P (eds) Machine Learning in Medical Imaging. Springer, Cham, pp 190–198
8. Lin Z, Li S, Ni D, Liao Y, Wen H, Du J, Chen S, Wang T, Lei B (2019) Multi-task learning for quality assessment of fetal head ultrasound images. Medical Image Anal. 58. https://doi.org/10.1016/j.media.2019.101548
9. Schlemper, J, Oktay, O, Chen, L, Matthew, J, Knight, CL, Kainz, B, Glocker, B, Rueckert, D (2018) Attention-gated networks for improving ultrasound scan plane detection. CoRR arXiv:1804.05338
10. Gou J, Yu B, Maybank S, Tao D (2021) Knowledge distillation: a survey. Int J Comp Vis 129. https://doi.org/10.1007/s11263-021-01453-z
11. Chen H, Dou Q, Ni D, Cheng J-Z, Qin J, Li S, Heng P-A (2015) Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks. In: Navab N, Hornegger J, Wells WM, Frangi A (eds) Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. Springer, Cham, pp 507–514
12. Wu J, Zeng P, Liu P, Lv G (2022) Automatic classification method of liver ultrasound standard plane images using pre-trained convolutional neural network. Connection Science 34(1):975–989. https://doi.org/10.1080/09540091.2021.2015748
13. Li, Y, Khanal, B, Hou, B, Alansary, A, Cerrolaza, JJ, Sinclair, M, Matthew, J, Gupta, C, Knight, C, Kainz, B, Rueckert, D (2018) Standard plane detection in 3d fetal ultrasound using an iterative transformation network. In: Frangi, AF, Schnabel, JA, Davatzikos, C, Alberola-López, C, Fichtinger, G (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2018. Springer, Cham , pp. 392-400
14. Pu B, Li K, Li S, Zhu N (2021) Automatic fetal ultrasound standard plane recognition based on deep learning and iiot. IEEE Trans. Ind. Informatics 17(11):7771–7780. https://doi.org/10.1109/TII.2021.3069470
15. Wang, X, Liu, Z, Du, Y, Diao, Y, Liu, P, Lv, G, Zhang, H (2021) Recognition of fetal facial ultrasound standard plane based on texture feature fusion. Comput Math Methods Med
16. Zhao L, Li K, Pu B, Chen J, Li S, Liao X (2022) An ultrasound standard plane detection model of fetal head based on multi-task learning and hybrid knowledge graph. Future Generation Computer Systems 135:234–243. https://doi.org/10.1016/j.future.2022.04.011
17. Hinton, GE, Vinyals, O, Dean, J (2015) Distilling the knowledge in a neural network. CoRR arXiv:1503.02531
18. Romero, A, Ballas, N, Kahou, SE, Chassang, A, Gatta, C, Bengio, Y (2015) Fitnets: hints for thin deep nets. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. arXiv:1412.6550
19. Jiao, X, Yin, Y, Shang, L, Jiang, X, Chen, X, Li, L, Wang, F, Liu, Q (2020) TinyBERT: distilling BERT for natural language understanding. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 4163-4174. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.findings-emnlp.372. https://aclanthology.org/2020.findings-emnlp.372
20. Sun, S, Cheng, Y, Gan, Z, Liu, J (2019) Patient knowledge distillation for bert model compression. In: EMNLP
21. Chen D, Mei J-P, Zhang Y, Wang C, Wang Z, Feng Y, Chen C (2021) Cross-layer distillation with semantic calibration. Proc AAAI Conf Artif Intell 35(8):7028–7036. https://doi.org/10.1609/aaai.v35i8.16865
22. Vaseli, H, Liao, Z, Abdi, AH, Girgis, H, Behnami, D, Luong, C, Dezaki, FT, Dhungel, N, Rohling, R, Gin, K, et al. (2019) Designing lightweight deep learning models for echocardiography view classification. In: Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling, vol. 10951, pp. 93-99. SPIE
23. Van der Maaten, L, Hinton, G (2008) Visualizing data using t-sne. J Mach Learning Res 9(11)
24. Bieder, F., Sandkühler, R, Cattin, PC (2021) Comparison of methods generalizing max- and average-pooling. CoRR arXiv:2103.01746
25. Abadi, M, Agarwal, A, Barham, P, Brevdo, E, Chen, Z, Citro, C, Corrado, GS, Davis, A, Dean, J, Devin, M, Ghemawat, S, Goodfellow, I, Harp, A, Irving, G, Isard, M, Jia, Y, Jozefowicz, R, Kaiser, L, Kudlur, M, Levenberg, J, Mané, D, Monga, R, Moore, S, Murray, D, Olah, C, Schuster, M, Shlens, J, Steiner, B, Sutskever, I, Talwar, K, Tucker, P, Vanhoucke, V, Vasudevan, V, Viégas, F, Vinyals, O, Warden, P, Wattenberg, M, Wicke, M, Yu, Y, Zheng, X (2015) TensorFlow: large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. https://www.tensorflow.org/
26. Simonyan, K, Zisserman, A (2015) Very deep convolutional networks for largescale image recognition. In: Bengio, Y, LeCun, Y (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. arXiv:1409.1556
27. Howard, AG, Zhu, M, Chen, B, Kalenichenko, D, Wang, W, Weyand, T, Andreetto, M, Adam, H (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. CoRR arXiv:1704.04861
28. Szegedy, C, Vanhoucke, V, Ioffe, S, Shlens, J, Wojna, Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818-2826
29. He, K, Zhang, X, Ren, S, Sun, J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778
30. Huang, G, Liu, Z, Maaten, LVD,Weinberger, KQ (2017) Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261-2269. IEEE Computer Society, Los Alamitos, CA, USA. https://doi.org/10.1109/CVPR.2017.243

31. Selvaraju, RR, Das, A, Vedantam, R, Cogswell, M, Parikh, D, Batra, D (2016) Grad-cam: why did you say that? Visual explanations from deep networks via gradient-based localization. CoRR arXiv:1610.02391

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Jacopo Dapueto is a Ph.D. student in Computer Science at University of Genova. He received his B.S. degree in Computer Science at University of Genova in 2019 and M.S. in Computer Science at University of Genova in 2021. His research interests include computer vision, machine learning. His graduate studies include collaborations with the University of Genoa (Italy) and Esaote S.p.A. (Italy) in utilizing machine learning systems designed to be installed in a real ultrasound device.

Luca Zini received a Laurea degree in Information Sciences in 2009 and a PhD in Computer Science in 2013 both from the University of Genova. His research interests include machine learning, and computer vision applied to medical devices to improve their usability and to speedup the workflow of the exam. Since 2018 he joined Esaote where is currently developing cutting edge technologies that intertwine artificial intelligence and medical imaging.

Francesca Odone is a Professor of Computer Science at the University of Genova and a founding member of MaLGa (Machine Learning Genoa) center. She received a Laurea degree in Information Sciences and a PhD in Computer Science both from the University of Genova. In 1999-2000, she was a visiting student at Heriot-Watt University (Edinburgh UK) with a EU Marie Curie research grant. Her research interests are in the fields of computer vision and machine learning, including multi-resolution signal processing, feature extraction, feature selection and data-driven representations for visual data. Often times her research is linked to applied tasks, in the fields robotics, ambient-assisted living, rehabilitation, video-surveillance, medical image analysis. She authored over 100 papers on international conferences and journals. She has been involved in various research projects and acted as a scientific coordinator of technology transfer contracts with SMEs, large companies and hospitals. More information can be found at http://malga.unige.it