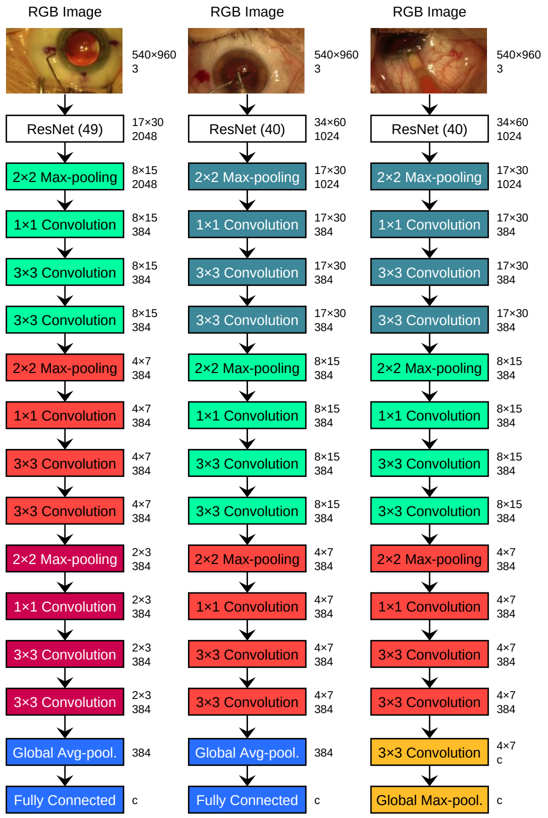

Fig. 3. Three exemplary instances of the FFE network family with different choices for k and the classification head. Left: $k = 49$, *avg-fc*. Center: $k = 40$, *avg-fc*. Right: $k = 40$, *conv-max*.

positive and negative examples. The set of positive examples for class i contains all frames that show tool i (potentially together with other tools), while the set of negative examples contains all other frames.

Each of the binary problems is trained using the cross-entropy loss function

$$H(p, q) = -(1 - p)\log(1 - q) - p\log(q) \ , \qquad (1)$$

where $p \in \{0, 1\}$ is the ground-truth probability for the positive class and $q \in [0, 1]$ is the predicted probability (network output) for the positive class. While the binary problems are conceptually separate, this makes no difference during training. For each example in the training set, the loss is calculated for all c network outputs and then backpropagated through the network.

IV. Dataset

A. Dataset description

During 50 cataract surgeries the view of the surgeon's microscope has been video recorded in a resolution of 1920×1080 pixels at 30 frames per second. The viewpoint is mostly static but the camera shakes and occasionally the zoom level is changed. On average, the surgery duration is 11 minutes so that over 9 hours of video are available. A total of 21 different surgical tools are used in the videos and up to 3 tools can be visible at a time. However, this is extremely rare (0.04 %) and most frames show no tools (45 %), one tool (38 %) or two tools (17 %). The videos have been independently annotated by two experts, which allows to ignore frames where the experts disagree during the evaluation. For each frame and surgical tool, a label indicates if the tool touches the eyeball. Note that this means a tool can be visible but still not annotated because it does not yet touch the eyeball. Also, multiple tools can be present in a single frame but bounding box information to distinguish the tools in the frame is not available. This is essentially a weakly supervised classification setting or more specifically multi-label classification. Finally, all 50 videos have been evenly split into a training and test set but labels are only provided for the training set.

B. Dataset challenges and preprocessing

The dataset poses some challenges that have to be addressed before being usable for training. First of all, the video resolution is very high at 1920×1080 pixels. This is a problem for CNNs as the required processing time and especially memory is directly influenced by the image resolution. However, the images cannot be scaled down arbitrarily because most tools are elongated objects of only about 30 pixels width. A resolution of 960×540 pixels was found to be a good compromise between resource demands and object size.

Due to the nature of video, subsequent frames are extremely similar to one another. They are heavily correlated and using neighboring frames during supervised learning yields almost no information gain. Therefore, only every sixth frame of each video is used during training which leads to 200 ms intervals between processed frames. No significant information is lost this way, but the required training time is reduced considerably.

Additionally, the similarity between neighboring frames has implications on the choice of validation data. Consider a random split of all frames into a large training and small validation set. It is highly likely that for each frame in the validation set either the predecessor or successor of that frame is part of the training set. The validation error would therefore be a significant underestimation of the test error. Instead, the training-validation-split is performed on the video level: 5 of the 25 training videos are set aside for validation purposes.

The dataset also exhibits a strong class imbalance. Because almost half of all frames do not show any tools, this subset of the data is undersampled to 40 % of its original size. While the number of available training examples for each tool also varies considerably, the more important consideration is how many videos contain a sequence showing each tool. This is relevant because frames showing the same tool in different videos are a lot more varied than frames showing the same tool in a single video. In consequence, the available videos are distributed between training and validation set so that both contain some videos showing each tool. The distribution of the eventual split can be seen in Figure 4. Unfortunately, it