Dynamic CNNs using uncertainty to overcome domain generalization for surgical instrument localization

Markus Philipp[1,2] Anna Alperovich[3] Marielena Gutt-Will[4]
Andrea Mathis[4] Stefan Saur[1] Andreas Raabe[4] Franziska Mathis-Ullrich[2]
[1]Carl Zeiss Meditec AG, DE [2]Karlsruhe Institute of Technology, DE
[3]Carl Zeiss AG, DE [4]Inselspital Bern, CH
first.lastname@{zeiss.com, kit.edu, insel.ch}

Abstract

Due to the limited amount of available annotated data in the medical field, domain generalization for applications in computer-assisted surgery is essential. Our work addresses this problem for the task of surgical instrument tip localization in neurosurgery, which is a classical step towards computer-assisted surgery. We propose an uncertainty-based CNN approach that dynamically selects the most relevant data source by incorporating its own uncertainty into the inference. In addition, the estimated uncertainty can visualize and easily explain the network's decision. Quantitative and qualitative evaluations show that our method outperforms state of the art approaches for large domain shifts and results are on-par for in-domain applications. Further increasing domain shifts by testing on different surgical disciplines, eye and laparoscopic surgeries, proves the generalization capabilities of the proposed method.

1. Introduction

Medical computer vision algorithms form the basis for several applications in computer-assisted surgery. Towards a clinical routine, these algorithms must be robust and efficient and perform well on familiar data domains as well as on data with typical domain shift, such as different types of surgery. In the machine learning realm, such a generic solution typically requires large amount of data, ideally from various data sources. However, acquiring broad medical datasets is very difficult due to legal/administrative requirements and large annotation efforts for medical experts. Thus, in practice, only small or medium-sized datasets from few hospitals and/or single types of surgery are available. Therefore, tackling the clinical need for generalization under typical domain shifts is a major challenge for medical computer vision applications [30].

In this work, we concentrate on the problem of localizing instrument tips in surgical video data as many compu-

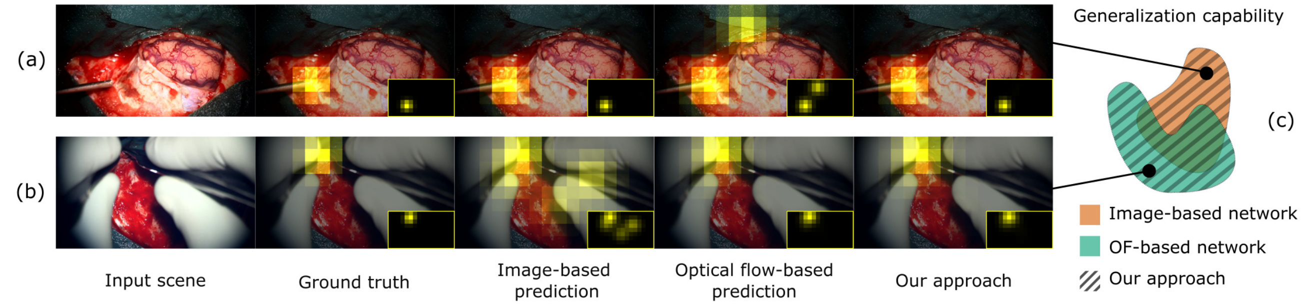


Figure 1: Instrument tip localization with the presented uncertainty-based dynamic CNN. Row (a) shows a typical neurosurgical scene with (from left to right): single frame input image; reference localization encoded as saliency overlay; saliency prediction using image information; saliency prediction using optical flow (OF); and our approach. Row (b) illustrates a neurosurgical scene that does not fit into the domain of typical training samples: large areas of the image are covered by the surgeon's hands normally not present in the training data. The image-based network fails completely on this sample. By incorporating uncertainty information, our approach ignores the incorrect parts of the prediction and relies more on OF. In (a), the opposite is observed: the image-based network correctly predicts the instrument tip, whereas the OF-based network is confused. Our approach correctly overbalances the image-based information and takes the best of both modalities (c).

tational assistance features rely on known positions of the instrument tips [1, 27]. A major challenge for instrument localization are domain shifts, where input data is different from what the algorithm saw during training [23]. By domain shift, we refer to as type of surgery (e.g. tumor, vascular, spine), illumination, level of blur, types and appearances of instruments. Collecting a large clinical dataset that contains all possible conditions is unfeasible.

Therefore, domain generalization should be achieved by algorithmic improvements. One approach to improve generalization given limited training data is to employ different information, e.g. spatial and temporal data [3]. For the problem of surgical instrument localization, previous work [22] combined spatial and temporal information by using image and optical flow modalities. While the presented method performed well for small domain shifts, it could not cope well with large domain shifts.

In this work, we also concentrate on using image and optical flow features as source of spatial and temporal information. Image data contains semantics about structures belonging to the instruments and the image background. Optical flow features contain only information about moving objects in the scene. Other static objects and background are not present in this modality. Thus, image and optical flow generalize dissimilarly on different domains. Our goal is to optimally combine these two sources of information to extend domain generalization.

Contributions. We propose an uncertainty-based dynamic convolutional neural network (CNN) for instrument tip localization that combines image and optical flow modalities. Our approach extracts relevant information from both data sources, guided by the estimated uncertainties. Given a new data sample, our network selects which information and features are employed to compute the most certain saliency map (see Fig. 1). Being trained on a single dataset, our approach generalizes on unseen data domains due the uncertainty-based fusion strategy. The estimated uncertainties are easily visualized and enable explainability of the network's decision. Our quantitative and qualitative evaluations show that we outperform state-of-the-art approaches and are on-par in-domain. We illustrate the superior generalization capability for various neurosurgical datasets. Furthermore, we observe good performance for datasets from completely different surgical disciplines, such as ophthalmology (i.e. eye surgery). This is the first approach for instrument localization that is capable of bridging such large domain gaps with a single network.

2. State of the art

Various approaches exist for surgical instrument localization. Some of them are detecting bounding boxes [21] or landmarks [2], others are segmenting the complete instrument [13]. Here, we localize the instrument tips as a coarse saliency map, due to advantages motivated in [7, 14, 22].

State-of-the-art methods towards the instrument localization problem achieve good results on benchmarking datasets [14], while domain generalization still remains a challenging task [23]. The current methods address domain generalization mainly with domain adaption techniques, e.g. fine-tuning on subsets [16] or online fine-tuning [32]. These methods assume a known test domain with access to the data. In a clinical setting, the test data is generally unknown a-priori. Thus, we need to generalize on unseen domains where no data is available.

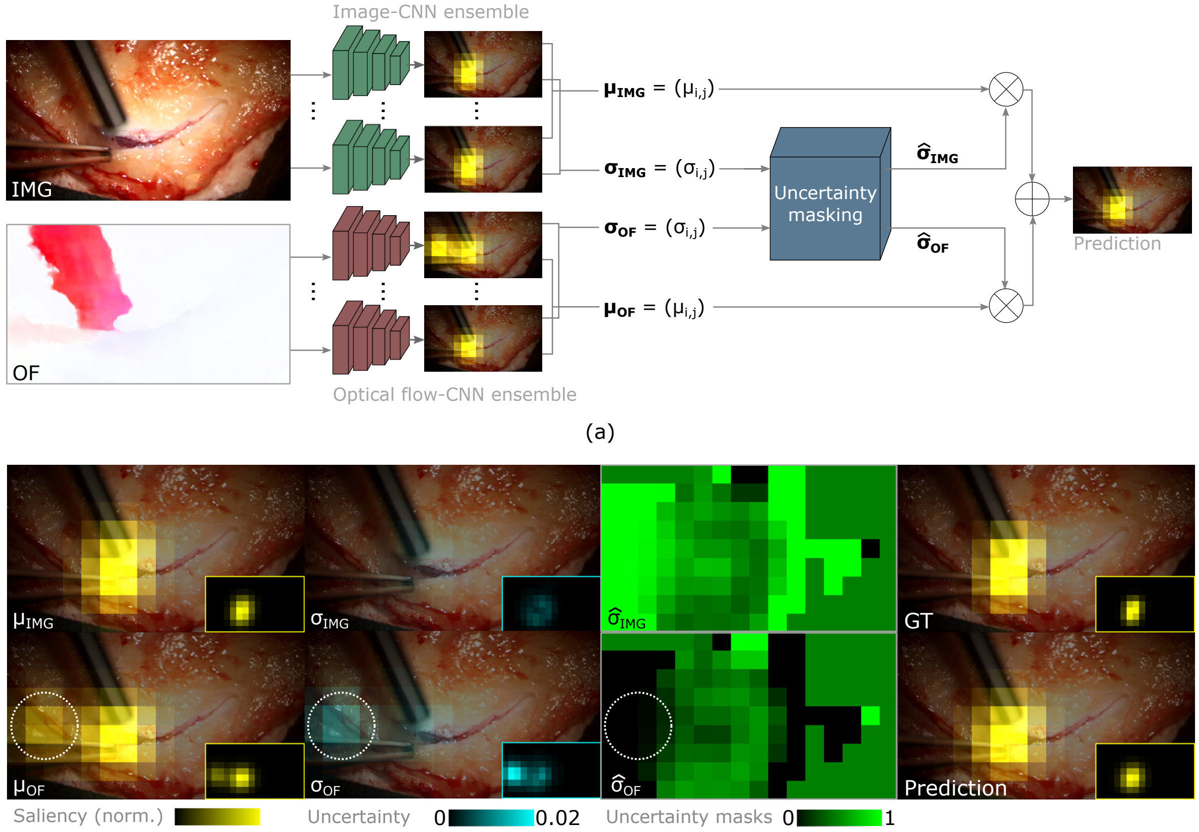
A classical approach to improve generalization is data augmentation [31]. However, it is challenging to design augmentation operations that cover all possible variations in the test domain [20]. Another method towards domain generalization is learning domain-invariant representations, which is the goal in domain alignment and meta-learning [19, 15, 29]. These approaches are beneficial when one can train on several datasets from different but related domains. For more domain generalization methods, see survey [33].

Since in the medical domain typically a lack of available datasets exists, we focus on exploring the information contained in a *single dataset*. Assuming availability of video data, one can extract both spatial and temporal information from this dataset. In [22] these modalities are fused by end-to-end learning to improve domain generalization. Although the method performs well on small domain shifts, it relies too much on spatial information and underperforms when this information varies from the training domain.

In case of large domain shifts, optimal fusion of spatial and temporal information for the prediction can improve generalization. To avoid unilateral focusing on the modality that performs best on the training domain, we propose to use dynamic neural networks. They are characterized by the ability to adapt to the input samples such that only relevant parts of the network are used for the prediction [10]. Dynamic networks can be controlled by several mechanisms, which are confidence-based [11], policy-based [6], or gating-based [5]. Our network dynamics are guided by the pixel-level uncertainty that we compute individually for spatial and temporal modalities. Furthermore, we use the uncertainty map to visualize and explain the network's decision. This addresses the need for transparency and explainability in medical computer vision applications [26]. For uncertainty estimation, most common techniques are based on Monte-Carlo drop-out [8] and ensembles [17]. Here, we utilize ensembles due to robustness and simplicity.

3. Methods

We propose a dynamic CNN, which fuses image and optical flow modalities such that for each sample the most reliable information contributes to the prediction. Our network consists of two ensemble networks, one for the image-based



(a)

(b)

Figure 2: Our method (a) dynamically combines the results from image and optical flow ensembles by means of an uncertainty estimation mechanism. For image and optical flow data, one ensemble model at a time is used to return N predictions. From these outputs, we calculate mean predictions $\boldsymbol{\mu_{IMG}}$, $\boldsymbol{\mu_{OF}}$ and pixel-wise uncertainties $\boldsymbol{\sigma_{IMG}}$, $\boldsymbol{\sigma_{OF}}$. We then determine uncertainty masks $\boldsymbol{\hat{\sigma}_{IMG}}$, $\boldsymbol{\hat{\sigma}_{OF}}$ using $\boldsymbol{\sigma_{IMG}}$ and $\boldsymbol{\sigma_{OF}}$. The final prediction is a linear combination of the outputs of the ensemble models, weighted by the uncertainty masks $\boldsymbol{\hat{\sigma}_{IMG}}$, $\boldsymbol{\hat{\sigma}_{OF}}$. In (b) we show intermediate and final outputs from our method for the scene in (a). At the highlighted location (see $\boldsymbol{\mu_{OF}}$ and $\boldsymbol{\sigma_{OF}}$), the optical flow-based ensemble model returns a false prediction but the uncertainty is increased. The third column shows the complementary uncertainty masks, where the coloring indicates how our network combines the two modalities. The black color at the highlighted region in $\boldsymbol{\hat{\sigma}_{OF}}$ indicates that our network ignores the optical flow information in the prediction. The last column shows ground truth and the prediction. Our method successfully utilizes the ensemble with low output uncertainty.

modality and one for optical flow modality. The main idea is to employ uncertainty-based criteria to assess the pixel-wise prediction quality for each modality. We obtain the final prediction as weighted linear combination of the individual predictions from the image and optical flow modalities. For each pixel, the weights are based on the uncertainty. The complete workflow of our method is illustrated in Fig. 2.

3.1. Ensembles

We describe the individual CNNs and how we organize them in ensembles. The general architecture includes two independent CNN types, following the identical architecture as the single-stream network in [22], which is derived from DenseNet [12]. The image-based network `IMG` takes only image information as input and outputs a saliency map $\boldsymbol{q_{IMG}}$, $\texttt{IMG} : (r, g, b) \rightarrow \boldsymbol{q_{IMG}}$. The optical flow-based network `OF` receives only optical flow and returns a saliency map $\boldsymbol{q_{OF}}$, $\texttt{OF} : (u, v) \rightarrow \boldsymbol{q_{OF}}$.

We independently train two ensemble models, $(\texttt{IMG}_i)_{i=1}^{N}$ and $(\texttt{OF}_i)_{i=1}^{N}$, where in our experiments $N = 10$. We use random weight initialization to achieve diversity within an ensemble, which is superior to e.g. bagging strategies [18] or Monte-Carlo dropout [17]. Figure 2(a) shows the arrangement of ensemble models, implemented in our approach.

3.2. Uncertainty estimation and masking

We combine the outputs for each ensemble by computing the mean of the prediction maps,

$$\boldsymbol{\mu} = (\mu_{i,j}) = \left(\frac{1}{N} \sum_{k=1}^{N} q_{i,j,k} \right), \tag{1}$$

where i, j are the pixel coordinates in the predicted saliency map and k is the index for the ensemble individual.

Following state of the art [9], we calculate the pixel-wise uncertainty map as standard deviation along the ensemble individuals:

$$\boldsymbol{\sigma} = (\sigma_{i,j}) = \left(\sqrt{\frac{1}{N} \sum_{k=1}^{N} (q_{i,j,k} - \mu_{i,j})^2} \right). \tag{2}$$

Thereby we obtain pairs $(\boldsymbol{\mu_{IMG}}, \boldsymbol{\sigma_{IMG}})$ and $(\boldsymbol{\mu_{OF}}, \boldsymbol{\sigma_{OF}})$ for image and optical flow ensembles, respectively.

To compute the final prediction, one would ideally combine the most certain parts from the image and optical flow outputs $\boldsymbol{\mu_{IMG}}$, $\boldsymbol{\mu_{OF}}$ in the final saliency map. To achieve this, the uncertainty map for each modality should serve as basis for a weighting function with entries between 0 and 1. The most confident areas should have values close to 1, while the uncertain regions are 0.

To model this behavior, we compute *uncertainty masks* by performing normalization of the uncertainty maps:

$$\boldsymbol{\hat{\sigma}_{IMG}} = \left(1 - \frac{\sigma_{IMG_{i,j}} + \epsilon}{\sigma_{OF_{i,j}} + \sigma_{IMG_{i,j}} + 2\epsilon} \right), \tag{3}$$

$$\boldsymbol{\hat{\sigma}_{OF}} = \left(1 - \frac{\sigma_{OF_{i,j}} + \epsilon}{\sigma_{OF_{i,j}} + \sigma_{IMG_{i,j}} + 2\epsilon} \right), \tag{4}$$

where we choose $\epsilon = 10^{-15}$. The uncertainty masks have the following properties:

- The uncertainty masks are complementary, i.e. $\boldsymbol{\hat{\sigma}_{IMG}} + \boldsymbol{\hat{\sigma}_{OF}} = \mathbf{1}$.
- If the image-based network is more uncertain than the optical flow, then our network relies the optical information, i.e. $\sigma_{IMG_{i,j}} \gg \sigma_{OF_{i,j}} \Rightarrow \hat{\sigma}_{IMG_{i,j}} = 0$, $\hat{\sigma}_{OF_{i,j}} = 1$. For $\sigma_{OF_{i,j}} \gg \sigma_{IMG_{i,j}}$ vice versa.
- If the image-based and optical flow-based networks are equally certain, they are averaged, i.e. $\sigma_{IMG_{i,j}} \approx \sigma_{OF_{i,j}} \Rightarrow \hat{\sigma}_{IMG_{i,j}} = \hat{\sigma}_{IMG_{i,j}} = 0.5$.

3.3. Fusion method and explainability

Finally, we use the uncertainty masks to dynamically join the ensembles for the two modalities. We compute the prediction by weighted linear combination of the two ensemble outputs,

$$q_{\text{pred}_{i,j}} = \hat{\sigma}_{OF_{i,j}} \times \mu_{OF_{i,j}} + \hat{\sigma}_{IMG_{i,j}} \times \mu_{IMG_{i,j}}, \tag{5}$$

where the weights are represented by the uncertainty masks.

We formulate our approach in pseudocode:

Algorithm 1: Dynamic, uncertainty-based CNN

```
Function predict_saliency(image, optical_flow):
    compute ensemble individuals $(\mathtt{IMG}_i)_{i=1}^{N}, (\mathtt{OF}_i)_{i=1}^{N}$
    compute $(\mu_{IMG}, \sigma_{IMG})$ and $(\mu_{OF}, \sigma_{OF})$
    compute uncertainty masks $\hat{\sigma}_{IMG}, \hat{\sigma}_{OF}$
    $q_{pred} \leftarrow \hat{\sigma}_{IMG} \times \mu_{IMG} + \hat{\sigma}_{OF} \times \mu_{OF}$
    return $q_{pred}$
```

Explainability. Based on the uncertainty mask properties, our fusion approach can be explained by assessment of intermediate results of the ensembles and the estimated uncertainty masks. Figure 2(b) illustrates the workflow for combination of the outputs of the two ensembles. Image and optical flow are fed into the separate ensembles and mean predictions are computed. In the illustrated example, the estimated uncertainty mask shows that the optical flow fails locally and thus, presents large uncertain areas. The image-based network shows higher confidence in most of the areas. The final prediction ignores the false saliency areas in the optical flow-based output and relies on the image information. Based on this decision, the final result matches the ground truth.

4. Experiments

In the following, we describe our experimental set-up and the base-line methods included in our comparison.

4.1. Base-line methods

For the remaining sections, we refer to the uncertainty-based fusion approach as `u-FUS`. For the comparison, we use two single-stream base-line methods. One uses image information as input, `IMG`, and the other uses optical flow as input, `OF`. We also consider a simple fusion approach, `p-FUS`, where we average the output of `IMG` and `OF`,

$$q_{\texttt{p-FUS}_{i,j}} = 0.5 \times q_{\texttt{IMG}_{i,j}} + 0.5 \times q_{\texttt{OF}_{i,j}}. \tag{6}$$

This approach assumes that both input modalities are equally important irrespective of the situation. Furthermore, we include an end-to-end fusion approach, `l-FUS`, where the final prediction is learned by a single network. According to [22], `l-FUS` outperforms the single-stream base-line

Dataset name	Dataset type	Domain shift severity	Ground truth data available
NeuroSurg-Tumor	clinical	in-domain	yes
NeuroSurg-Vascular	clinical	low-medium	yes
NeuroSurg-Spine	clinical	low-medium	yes
NeuroSurg-Phantom	phantom	large	yes
Cataract-101	clinical	large	no
SurgicalActions160	clinical	large	no

Table 1: Datasets used for evaluation.

methods on test data similar to the training domain, while it underperforms in the presence of a large domain shift.

4.2. Datasets and evaluation metric

During experiments, we use various surgical datasets, described in Tab. 1.

For training, we use eight tumor surgeries from the neurosurgical dataset NeuroSurg, introduced in [22]. We refer to the remaining independent tumor surgeries as in-domain examples. Other clinical datasets, NeuroSurg-Vascular and NeuroSurg-Spine, feature low to medium domain shift. These surgeries are visually similar to NeuroSurg-Tumor, but include differences in instrument set, tissue type and visual appearance. The phantom surgeries, NeuroSurg-Phantom, introduce a large domain shift compared to the tumor cases. This dataset is recorded on artificial tissue, while using commercial surgical instruments. Figure 3, top row, illustrates example scenes from the NeuroSurg dataset.

To test the limits of our approach, we include datasets from completely different surgical disciplines. We include the domain of eye surgery (Cataract-101 [25]) and laparoscopy (SurgicalActions160 [24]). Both datasets differ strongly w.r.t. surgical instruments and biological tissue visible in the video data (for visual examples, see Section 5.2, Fig. 5). We pre-processed both datasets from raw video data by resampling and cropping (256 × 144 px). As there are no saliency ground truth annotations for these two datasets, we focus on qualitative observations.

For numerical evaluation we use the SIM score [4],

$$\text{SIM} = \sum_{(i,j)} \min(q_{\text{Ground truth}_{i,j}}, q_{\text{pred}_{i,j}}), \tag{7}$$

whereas $\sum q_{\text{Ground truth}_{i,j}} = \sum q_{\text{pred}_{i,j}} = 1$.

4.3. Training and implementation

For training we use 22,315 samples (6 surgeries) from the training surgeries in NeuroSurg-Tumor. The remaining 2 surgeries are used for validation (5,093 samples). We use Adam optimizer with an initial learning rate = 0.01 and decay factor = 0.1. As loss function, we take MSE. We employ an early stopping strategy based on the validation metric (with max. 500 epochs). Training is performed on a system with a NVIDIA T4, Intel Xeon Gold 6242R, 64 GB. Run-time for prediction on this system ≈ 700 ms. Optical flow is computed with PWC-Net [28].

5. Results

We present quantitative and qualitative evaluations for in-domain and out-domain datasets.

5.1. Quantitative results

We compare base-line approaches, two single-stream networks IMG and OF, and fusion approaches p-FUS and l-FUS with our uncertainty-based u-FUS method. We compute the mean values of the SIM (Eq. (7)) for all test datasets with available ground truth and calculate statistics (see Tab. 2). In-domain, when training and testing on tumor surgeries, the performance of the uncertainty-based approach is on-par with the end-to-end network l-FUS and outperforms all other methods (Tumor 1, $M_{\texttt{l-FUS}} = 0.84$ vs. $M_{\texttt{u-FUS}} = 0.84$). In case of a minor domain shift, when testing on vascular or spinal surgeries, we observe similar behavior to the in-domain scenario. In case of a large domain shift, when testing on phantom data, we observe a performance improvement for our approach compared to other networks (Phantom 2, $M_{\texttt{l-FUS}} = 0.71$ vs. $M_{\texttt{u-FUS}} = 0.80$).

In Tab. 2, one can easily select the best base-line method for a particular dataset. However, among those methods there is no clear winner that performs equally well in all scenarios. Our approach consistently shows competitive or superior performance on all datasets, regardless of the amount of domain shift.

To understand how our approach u-FUS reacts to the domain shift, we analyze the correlation between the SIM distribution and the distributions of the uncertainty masks (see Fig. 3). In-domain, our network heavily prefers the image-information (see Fig. 3, first column). With increasing domain shift, our method starts to prefer optical flow-information (see Fig. 3, remaining columns). This explains the superior performance of u-FUS because it understands which modality performs best in every particular situation.

Statistical analysis on the complete datasets shows that our approach is superior to the base-line methods w.r.t. explainability and robustness. In the next step, we look closer to the performance of u-FUS on individual examples.

5.2. Qualitative results

We compare fusion-based base-line networks with our approach u-FUS. The end-to-end fusion approach l-FUS performs better on the clinical cases than the single-stream network. However, it could not benefit from both modalities on phantom data. From Tab. 2 we conclude that the simple averaging approach p-FUS performs better on large domain shifts than the pure end-to-end solution.

Figure 4(a) illustrates the results for an in-domain sample, where our approach and end-to-end learning perform

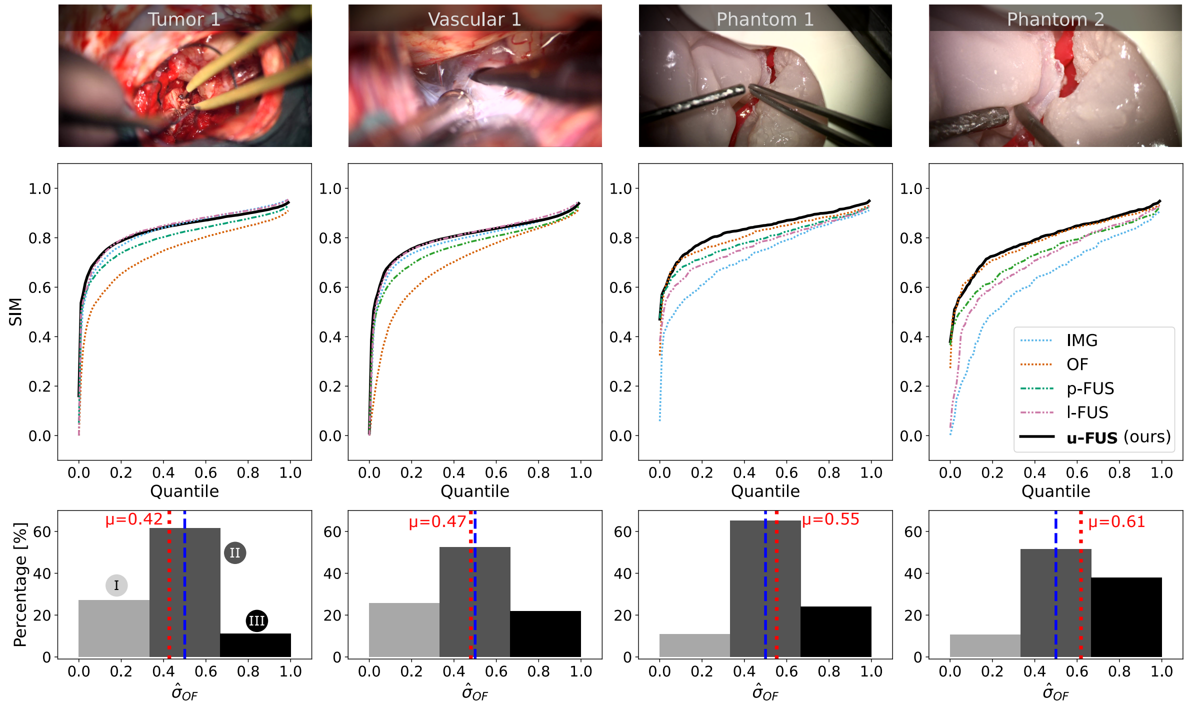


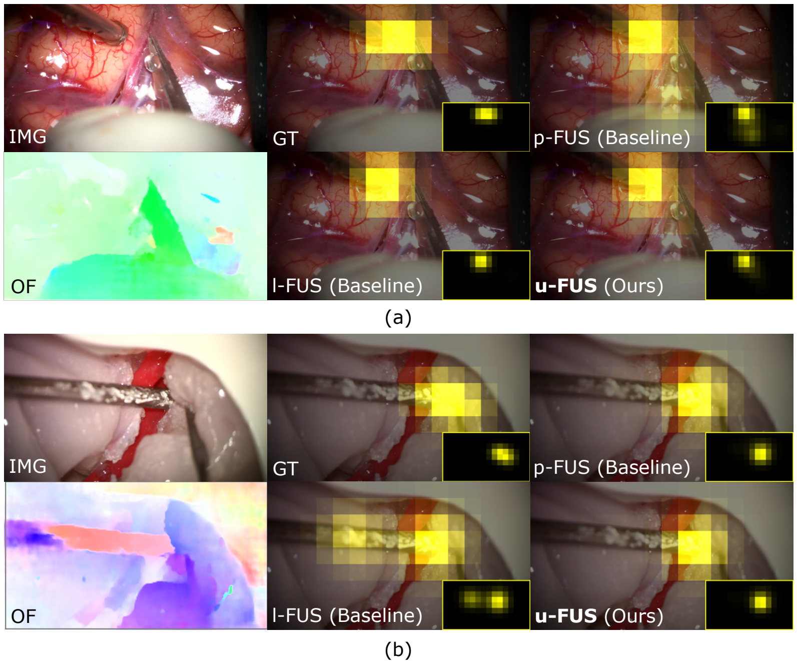
Figure 3: Quantitative evaluation of the SIM quantile distribution and uncertainty masks of test-surgeries, spanning from in-domain to large domain shifts. The first row illustrates a sample image for each type of surgery. The second row shows a comparison of the SIM distribution for all base-line methods and our approach `u-FUS`. Our method outperforms or is on-par with all base-line methods for each surgical domain. In the third row, we illustrate how `u-FUS` combines image- and optical flow-based predictions depending on the domain. For this, we compute the uncertainty masks for all video frames within each test case. We then calculate the histogram w.r.t. the entries of all masks $\hat{\boldsymbol{\sigma}}_{\boldsymbol{OF}}$ within a test case, where we choose $n_{\text{bins}} = 3$. The first bin (I) collects saliency pixels where our network is certain about the image-modality. For bin (II), our network is approximately equally confident in both modalities. In (III), our network is more certain about the optical flow modality. For the in-domain tumor surgery, we observe that our network is mostly certain about the image information, while for the out-of-domain phantom surgery, the network is more confident about optical flow.

well, while the simple fusion underperforms due to false positives in the optical flow prediction. Figure 4(b) shows a sample from the phantom data (large domain shift). Learning-based fusion `l-FUS` fails to predict a correct saliency map, since it is provided unseen image content. The simple fusion `p-FUS` and our approach `u-FUS` do not fully rely on the image information and thus produce better results. We observe that our approach adapts consistently better to the input domain compared to base-line methods.

We focus on large domain shifts and explainability for datasets `NeuroSurg-Phantom`, `Cataract-101`, `SurgicalAction160` that are significantly different from the training data. We investigate which type and amount of information contribute to the final prediction of `u-FUS`. Figure 5(a) illustrates another example from the phantom data, where we observe that the most certain and correct information are obtained from the optical flow-based branch. Figure 5(b) shows a sample from a laparo-

	`NeuroSurg-Tumor`		`NeuroSurg-Vascular`		`NeuroSurg-Spine`		`NeuroSurg-Phantom`	
	Tumor 1	Tumor 2	Vascular 1	Vascular 2	Spine 1	Spine 2	Phantom 1	Phantom 2
IMG	$0.83^{(**)}$	$0.81^{(**)}$	$0.78^{(**)}$	$0.72^{(**)}$	$0.78^{(**)}$	$0.72^{(**)}$	$0.73^{(**)}$	$0.63^{(**)}$
OF	$0.71^{(**)}$	$0.73^{(**)}$	$0.70^{(**)}$	$0.65^{(**)}$	$0.73^{(**)}$	$0.67^{(**)}$	$0.81^{(**)}$	$0.79^{(**)}$
p-FUS	$0.80^{(**)}$	$0.79^{(**)}$	$0.76^{(**)}$	$0.71^{(**)}$	$0.77^{(**)}$	$0.73^{(**)}$	$0.79^{(**)}$	$0.74^{(**)}$
l-FUS	$\mathbf{0.84}^{(**)}$	$\mathbf{0.83}^{(\ )}$	$\mathbf{0.80}^{(\ )}$	$0.74^{(**)}$	$\mathbf{0.81}^{(\ )}$	$\mathbf{0.77}^{(**)}$	$0.77^{(**)}$	$0.71^{(**)}$
u-FUS (ours)	0.84	**0.83**	**0.80**	**0.75**	0.80	0.76	**0.83**	**0.80**

Table 2: Mean values (M) of SIM score for clinical data and phantom data. Legend for Bonferroni-corrected pairwise t-test $(H_0 : M_{\texttt{u-FUS}} = M_{\text{X}})$: $p \geq 0.05 = ^{(\ )}$, $p < 0.05 = ^{(*)}$, $p < 0.01 = ^{(**)}$



(a)

(b)

Figure 4: Qualitative results on individual scenes for the NeuroSurg dataset. (a) shows an in-domain example from a clinical case, together with the estimated optical flow. The second column illustrates the ground truth (GT) and the results from end-to-end fusion. The third column shows the results for the simple fusion by averaging and our uncertainty-based fusion method. We observe that l-FUS and our method u-FUS perform well, while p-FUS produces false positive predictions. (b) shows an example from the phantom data, where w.r.t. base-lines methods we observe opposite behavior compared to the in-domain example. l-FUS fails, while p-FUS performs well. Our approach u-FUS copes well with both scenarios.

scopic surgery scene. Although the instrument set is different from the neurosurgical domain, the visual appearance of the scene seems similar to the training data. Thus, the most certain information comes from the image-based branch. Figure 5(c) provides a sample from a cataract surgery. Here, the instrument set and the visual appearance is vastly different from anything the network saw during training. In addition to this, eye motion is captured by the optical flow, which is not present in other surgery types. Nevertheless, our u-FUS method correctly finds the exact instrument tip localization. From the uncertainty masks we conclude that both modalities are utilized although the network is not certain in either of them. However, both networks are equally certain about the areas where there are no instruments. This suggests that our approach still focuses on the correct image area while this area is slightly fuzzier than in the other scenarios due to extremely large domain shift.

6. Discussion and conclusion

We investigated a dynamic and explainable CNN for surgical instrument tip localization. This method is specifically designed for use in the medical domain, where only few training datasets are available and explainability is of high importance. While the network is trained with a limited amount of data, it generalizes on a variety of different domains, ranging from in-domain data (tumor surgeries) towards large domain shifts (phantom, laparoscopic and eye surgeries). Our quantitative evaluation proves the superior performance of our approach compared to state of the art baseline-methods. Addressing explainability, our network's decision is based on the uncertainty masks, which is easy to visualize and interpret. They clearly illustrate which part of the input modalities contribute to the final prediction.

Being dynamic, our network successfully adapts to new domains by selecting and utilizing the relevant information from the input modalities. These properties are essential towards application in a clinic, where the surgical conditions and the setting of the operation are not a-priori known by the algorithm.

To test the limits of our approach, we went *beyond* expected data variations. Normally, training and test surgeries are conducted under similar conditions (instruments, tissue, recording conditions). To account for the possible large variations, for instance change of instrumentation set or illumination, we test our algorithm on completely different surgical disciplines. Even in these challenging conditions, when neither instruments nor tissue has previously been seen by the network, it successfully localizes the in-

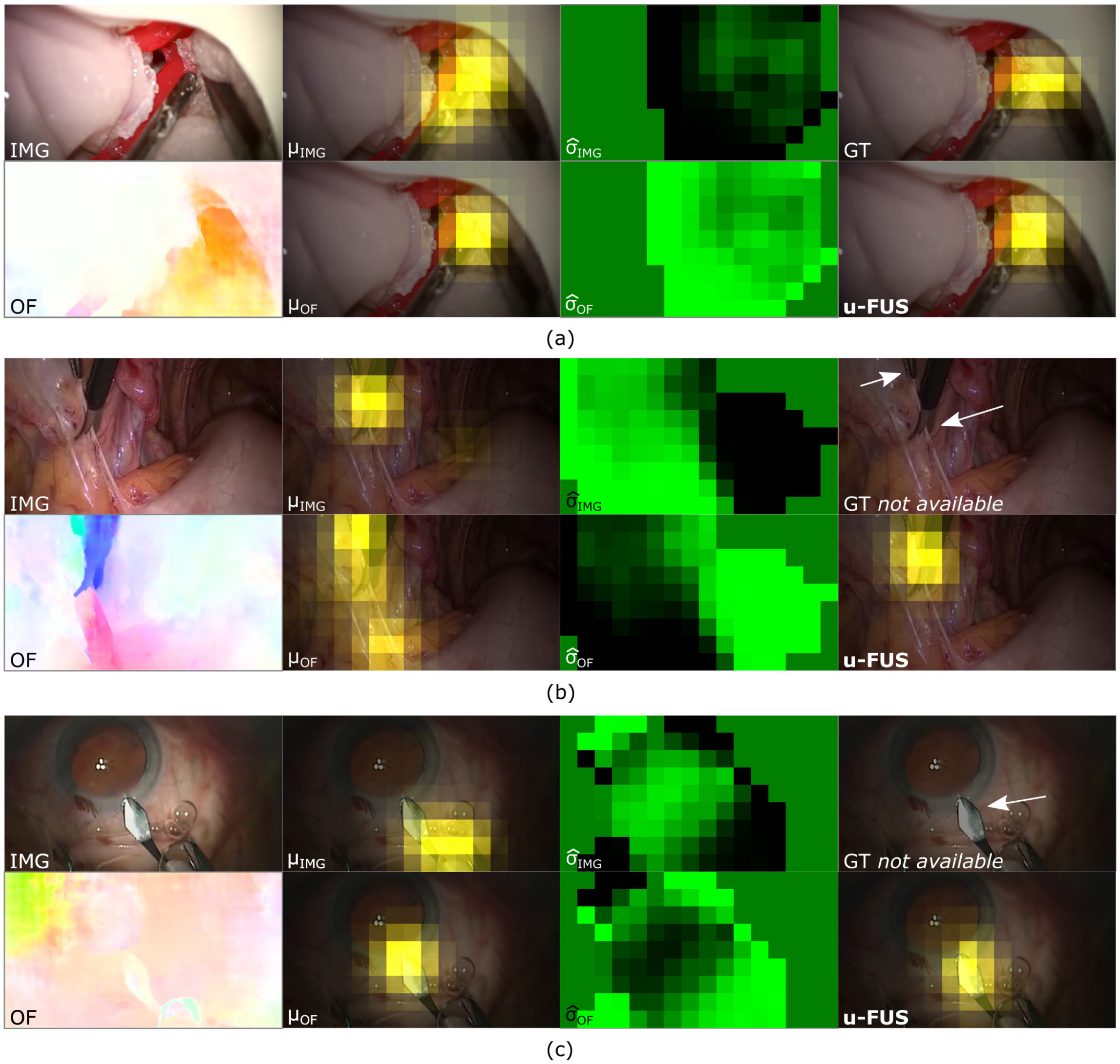


Figure 5: Qualitative results on single scenes from phantom data and other surgical disciplines, laparoscopy and cataract surgery. (a) illustrates an example of the phantom data, where the first column depicts the input image and optical flow. The second column shows the intermediate prediction of the image- and optical flow-based ensembles. The third column displays the uncertainty masks. The last column shows ground truth (GT) vs. our result. (b) displays a scene from a laparoscopic surgery (`SurgicalActions160`) with the same arrangement of figures as in (a). The ground truth for this dataset is not available. Thus, we evaluate only qualitatively. (c) shows a scene from a cataract surgery (`Caract-101`). This is an edge case since both content and visual appearance differ strongly from the neurosurgical data. In all cases, our network achieves correct predictions, and its decision can be explained by assessing the uncertainty masks.

strument tips and remains explainable. Compared to state-of-the-art end-to-end learning, our uncertainty-based network presents a step towards clinical application given its generalization and explainability properties.

In order to bring the algorithm to a level ready for clinical routine, large-scale evaluations are required. Once the algorithm is used in a clinical setting, an additional calibration step of the uncertainties can be performed to further improve the localization performance. While we demonstrate the applicability of our approach to the problem of instrument tip localization, our method does not require any explicit domain knowledge. Instead, the domain knowledge is purely gained by the uncertainty mask, which is generally available through ensembling. Therefore, our method can be applied to further computer vision tasks where multiple input modalities are available for network training.

References

[1] Anthony Agustinos, Rémi Wolf, Jean-Alexandre Long, Philippe Cinquin, and Sandrine Voros. Visual servoing of a robotic endoscope holder based on surgical instrument tracking. In *5th IEEE RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics*, pages 13–18. IEEE, 2014.

[2] Mohamed Alsheakhali, Abouzar Eslami, Hessam Roodaki, and Nassir Navab. Crf-based model for instrument detection and pose estimation in retinal microsurgery. *Computational and mathematical methods in medicine*, 2016, 2016.

[3] C. Bak, A. Kocak, E. Erdem, and A. Erdem. Spatio-temporal saliency networks for dynamic saliency prediction. *IEEE Transactions on Multimedia*, 20(7):1688–1698, 2018.

[4] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018.

[5] Jinting Chen, Zhaocheng Zhu, Cheng Li, and Yuming Zhao. Self-adaptive network pruning, 2019.

[6] Zhourong Chen, Yang Li, Samy Bengio, and Si Si. You look twice: Gaternet for dynamic filter selection in cnns, 2019.

[7] Shahram Eivazi, Roman Bednarik, Markku Tukiainen, Mikael von und zu Fraunberg, Ville Leinonen, and Juha E Jääskeläinen. Gaze behaviour of expert and novice microneurosurgeons differs during observations of tumor removal recordings. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 377–380, 2012.

[8] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016.

[9] Ayana Ghosh, Bobby G Sumpter, Ondrej Dyck, Sergei V Kalinin, and Maxim Ziatdinov. Ensemble learning-iterative training machine learning for uncertainty quantification and automated experiment in atom-resolved microscopy. *npj Computational Materials*, 7(1):1–8, 2021.

[10] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey, 2021.

[11] Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Q. Weinberger. Multi-scale dense networks for resource efficient image classification, 2018.

[12] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018.

[13] Mobarakol Islam, Daniel Anojan Atputharuban, Ravikiran Ramesh, and Hongliang Ren. Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning. *IEEE Robotics and Automation Letters*, 4(2):2188–2195, 2019.

[14] Mobarakol Islam, Yueyuan Li, and Hongliang Ren. Learning where to look while tracking instruments in robot-assisted surgery, 2019.

[15] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Feature alignment and restoration for domain generalization and adaptation, 2020.

[16] Xiaowen Kong, Yueming Jin, Qi Dou, Ziyi Wang, Zerui Wang, Bo Lu, Erbao Dong, Yun-Hui Liu, and Dong Sun. Accurate instance segmentation of surgical instruments in robotic surgery: model refinement and cross-dataset evaluation. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–8, 2021.

[17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2017.

[18] Stefan Lee, Senthil Purushwalkam, Michael Cogswell, David Crandall, and Dhruv Batra. Why m heads are better than one: Training a diverse ensemble of deep networks, 2015.

[19] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization, 2017.

[20] Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex C. Kot. Domain generalization for medical imaging classification with linear-dependency regularization, 2020.

[21] Yeqing Li, Chen Chen, Xiaolei Huang, and Junzhou Huang. Instrument tracking via online learning in retinal microsurgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 464–471. Springer, 2014.

[22] M. Philipp, A. Alperovich, M. Gutt-Will, A. Mathis, S. Saur, A. Raabe, and F. Mathis-Ullrich. Localizing neurosurgical instruments across domains and in the wild. *Proceedings of Machine Learning Research*, 143:581—595, 2021.

[23] Tobias Ross, David Zimmerer, Anant Vemuri, Fabian Isensee, Manuel Wiesenfarth, Sebastian Bodenstedt, Fabian Both, Philip Kessler, Martin Wagner, Beat Müller, et al. Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. *International journal of computer assisted radiology and surgery*, 13(6):925–933, 2018.

[24] Klaus Schoeffmann, Heinrich Husslein, Sabrina Kletz, Stefan Petscharnig, Bernd Muenzer, and Christian Beecks. Video retrieval in laparoscopic video recordings with dynamic content descriptors. *Multimedia Tools and Applications*, 77(13):16813–16832, Jul 2018.

[25] Klaus Schoeffmann, Mario Taschwer, Stephanie Sarny, Bernd Münzer, Manfred Jürgen Primus, and Doris Putzgruber. *Cataract-101: Video Dataset of 101 Cataract Surgeries*, page 421–425. Association for Computing Machinery, New York, NY, USA, 2018.

[26] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(6), 2020.

[27] Kai-Tai Song and Chun-Ju Chen. Autonomous and stable tracking of endoscope instrument tools with monocular camera. In *2012 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, pages 39–44. IEEE, 2012.

[28] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume, 2018.

[29] Shujun Wang, Lequan Yu, Kang Li, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. Dofe: Domain-oriented feature

embedding for generalizable fundus image segmentation on unseen datasets. *IEEE Transactions on Medical Imaging*, 39(12):4237–4248, 2020.

[30] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Holger Roth, Andriy Myronenko, Daguang Xu, and Ziyue Xu. When unseen domain generalization is unnecessary? rethinking data augmentation. 2019.

[31] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Bradford J. Wood, Holger Roth, Andriy Myronenko, Daguang Xu, and Ziyue Xu. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Transactions on Medical Imaging*, 39(7):2531–2540, 2020.

[32] Zixu Zhao, Yueming Jin, Bo Lu, Chi-Fai Ng, Qi Dou, Yun-Hui Liu, and Pheng-Ann Heng. One to many: Adaptive instrument segmentation via meta learning and dynamic online adaptation in robotic surgical video, 2021.

[33] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization in vision: A survey, 2021.