```
%% Copernicus Publications Manuscript Preparation Template for LaTeX Submissions
%DIF LATEXDIFF DIFFERENCE FILE


%% --------------------------------
%% This template should be used for copernicus.cls
%% The class file and some style files are bundled in the Copernicus Latex Package, which can be downloaded from the different journal webpages.
%% For further assistance please contact Copernicus Publications at: production@copernicus.org
%% https://publications.copernicus.org/for_authors/manuscript_preparation.html


%% Please use the following documentclass and journal abbreviations for preprints and final revised papers.

%% 2-column papers and preprints
\documentclass[journal abbreviation, manuscript]{copernicus}

\graphicspath{{./Figures/}}
\usepackage[dvipsnames]{xcolor}
\usepackage{soul}


%% Journal abbreviations (please use the same for preprints and final revised papers)

%The Cryosphere (tc)

%\usepackage commands included in the copernicus.cls:
%\usepackage[german, english]{babel}
%\usepackage{tabularx}
%\usepackage{cancel}
%\usepackage{multirow}
```

```latex
%\usepackage{supertabular}
%\usepackage{algorithmic}
%\usepackage{algorithm}
%\usepackage{amsthm}
%\usepackage{float}
%\usepackage{subfig}
\usepackage{rotating}
%DIF PREAMBLE EXTENSION ADDED BY LATEXDIFF
%DIF UNDERLINE PREAMBLE %DIF PREAMBLE
\RequirePackage[normalem]{ulem} %DIF PREAMBLE
\RequirePackage{color}\definecolor{RED}{rgb}{1,0,0}\definecolor{BLUE}{rgb}{0,0,1} %DIF PREAMBLE
\providecommand{\DIFadd}[1]{{\protect\color{blue}\uwave{#1}}} %DIF PREAMBLE
\providecommand{\DIFdel}[1]{{\protect\color{red}\sout{#1}}}                    %DIF PREAMBLE
%DIF SAFE PREAMBLE %DIF PREAMBLE
\providecommand{\DIFaddbegin}{} %DIF PREAMBLE
\providecommand{\DIFaddend}{} %DIF PREAMBLE
\providecommand{\DIFdelbegin}{} %DIF PREAMBLE
\providecommand{\DIFdelend}{} %DIF PREAMBLE
%DIF FLOATSAFE PREAMBLE %DIF PREAMBLE
\providecommand{\DIFaddFL}[1]{\DIFadd{#1}} %DIF PREAMBLE
\providecommand{\DIFdelFL}[1]{\DIFdel{#1}} %DIF PREAMBLE
\providecommand{\DIFaddbeginFL}{} %DIF PREAMBLE
\providecommand{\DIFaddendFL}{} %DIF PREAMBLE
\providecommand{\DIFdelbeginFL}{} %DIF PREAMBLE
\providecommand{\DIFdelendFL}{} %DIF PREAMBLE
%DIF END PREAMBLE EXTENSION ADDED BY LATEXDIFF

\begin{document}

\title{Automated avalanche mapping from SPOT 6/7 satellite imagery\DIFdelbegin
\DIFdel{:}\DIFdelend \\ \DIFaddbegin \DIFadd{with deep learning: }\DIFaddend results, evaluation,
potential and limitations}
```

```
% \Author[affil]{given_name}{surname}

\Author[1,2,3]{Elisabeth D.}{Hafner}

\Author[3]{Patrick}{Barton}

\Author[3]{Rodrigo Caye}{Daudt}

\Author[3,4]{Jan Dirk}{Wegner}

\Author[3]{Konrad}{Schindler}

\Author[1,2]{Yves}{Bühler}

\affil[1]{WSL Institute for Snow and Avalanche Research SLF, Davos Dorf, 7260, Switzerland}

\affil[2]{Climate Change, Extremes, and Natural Hazards in Alpine Regions Research Center CERC¸
Davos Dorf, 7260, Switzerland}

\affil[3]{EcoVision Lab, Photogrammetry and Remote Sensing, ETH Zurich, Zurich, 8092, Switzerland}

\affil[4]{Institute for Computational Science, University of Zurich, Zurich, 8057, Switzerland}

%% The [] brackets identify the author with the corresponding affiliation. 1, 2, 3, etc. should be
inserted.

%% If an author is deceased, please mark the respective author name(s) with a dagger, e.g.
"\Author[2,$\dag$]{Anton}{Aman}", and add a further "\affil[$\dag$]{deceased, 1 July 2019}".

%% If authors contributed equally, please mark the respective author names with an asterisk, e.g.
"\Author[2,*]{Anton}{Aman}" and "\Author[3,*]{Bradley}{Bman}" and add a further affiliation:
"\affil[*]{These authors contributed equally to this work.}".

\correspondence{Elisabeth D. Hafner (elisabeth.hafner@slf.ch)}

\runningtitle{TEXT}

\runningauthor{TEXT}
```

\received{}

\pubdiscuss{} %% only important for two-stage journals

\revised{}

\accepted{}

\published{}

%% These dates will be inserted by Copernicus Publications during the typesetting process.

\firstpage{1}

\maketitle

\begin{abstract}

Spatially dense and continuous information on avalanche occurrences is crucial for numerous safety related applications such as avalanche warning, hazard zoning, hazard mitigation measures, forestry, risk management and numerical simulations. This information is today still collected in a non-systematic way by observers in the field. Current research has explored \DIFdelbegin \DIFdel{and proposed applying }\DIFdelend \DIFaddbegin \DIFadd{the application of }\DIFaddend remote sensing technology to fill this information gap\DIFaddbegin \DIFadd{, }\DIFaddend by providing spatially continuous information on avalanche occurrences over large regions. Previous investigations have confirmed the high potential of avalanche mapping from remote sensed imagery to complement existing databases. Currently, the bottleneck for fast data provision from optical data is the time-consuming manual mapping. In our study we deploy a slightly adapted DeepLabV3+, a state-of-the-art deep learning model, to automatically identify and map avalanches in SPOT6/7 imagery from 24 January 2018 and 16 January 2019. We relied on 24'778 manually annotated avalanche polygons split into geographically disjoint regions for training, validating and testing. Additionally, we investigate generalization ability by testing our best model configuration on SPOT 6/7 data from 6 January 2018 and comparing to avalanches we manually annotated for that purpose.

%

To assess the quality of the model results, we investigate the probability of detection (POD), the positive predictive value (PPV) and the F1-score. Additionally, we assessed the reproducibility of manually annotated avalanches in a small subset of our data. We achieved an average POD of 0.610, PPV of 0.668 and an F1-score of 0.625 in our test areas and found an F1-score in the same range for avalanche outlines annotated by different experts. Our model and approach are an important step towards a fast and comprehensive documentation of avalanche periods from optical satellite imagery in the future, complementing existing avalanche databases. This will have a large impact on safety related applications, making mountain regions safer.

\end{abstract}

%\copyrightstatement{TEXT} %% This section is optional and can be used for copyright transfers.

\introduction %% \introduction[modified heading if necessary]

Information about occurred avalanches, their location and dimensions are pivotal for many applications such as avalanche warning, hazard zoning, hazard mitigation infrastructure, forestry, risk management and numerical simulations \citep[e.g.][]{Meister.1994, RudolfMiklau.2015, Bebi.2009,BrundlMargreth.2015, Christen.2010, buhler.2022}. Currently this information is reported and collected unsystematically by observers and (local) avalanche warning services. In recent years different groups have proposed to use remote sensing to fill that gap and provide spatially continuous, complete maps of avalanche occurrences over some region of interest \citep{Buhler.2009, Lato.2012, Eckerstorfer.2016, Korzeniowska.2017}. \DIFdelbegin %DIFDELCMD <

%DIFDELCMD < %%%

\DIFdelend It has been shown that avalanches can be identified with sufficient reliably from optical data \citep[e.g.,][]{Buhler.2019} or Synthetic Aperture Radar \citep[SAR; e.g.,][]{Eckerstorfer.2016, Abermann.2019}, with varying degrees of completeness depending on the sensor and the size of the avalanches \citep{hafner.2021}.

\DIFaddbegin

\DIFadd{Both optical and SAR data have inherent advantages and disadvantages which we would like to elaborate on in the following section: For the acquisition of suitable data, SAR is independent of cloud cover, whereas for optical data a clear sky is a crucial prerequisite. Consequently, optical data may only capture occurred avalanches after a period with activity is over (except for avalanches releasing solely due to the warming during the day), whereas with SAR information may also be retrieved during an avalanche period. Due to that independence from low-visibility weather conditions, and in the case of Sentinel-1 a 12-day repeat cycle at mid latitudes, the temporal resolution is in the best case daily in northern Norway or about every 6 days in central Europe (numbers for two Sentinel-1 satellites acquiring data, currently the temporal resolution is about half as Sentinel-1B has not been acquiring since 23 December 2021). The optical satellite data currently known to be suitable for avalanche mapping need to be ordered specifically and are therefore only available at isolated dates in time. Compared to SAR, optical data is however easier to process and interpret. In our previous work \mbox{%DIFAUXCMD

\citep{hafner.2021} }\hspace{0pt}%DIFAUXCMD

we compared the performance and completeness of SAR Sentinel-1 as well as optical SPOT 6/7 and Sentinel-2 for avalanche mapping. In a detailed analysis of the manual mappings we found the following: the ground sampling distance of 10m makes Sentinel-2 unsuitable for the mapping of avalanches. The mapping from SPOT 6/7 is overall more complete compared to Sentinel-1, which is mostly caused by the inability to confidently map avalanches of size 3 and smaller in Sentinel-1 imagery, a characteristic related to the underlying spatial resolution of approximately 10-15m for Sentinel-1 and 1.5m for SPOT 6/7. Depending on the application, practitioners not only want to know when and where an avalanche occurred, but also the outlines. When analyzing which part of an avalanche can typically be identified using Sentinel-1 we found \mbox{%DIFAUXCMD

\citep[in accordance with, among others,][]{eckerstorfer.2022} }\hspace{0pt}%DIFAUXCMD

that it is mostly the deposit, but may include patches from track and release area. When only using Sentinel-1 data it is therefore neither possible to derive the number of avalanches occurred (possibly several unconnected patches for one avalanche), nor the size of the occurred avalanches (size of patches detected does not usually correspond to avalanche size). Consequently, unless unambiguous with respect to the terrain, the origin and release area of avalanche deposits detected using SAR images remain unknown. In contrast, except for shaded areas, in SPOT 6/7 avalanches can be identified from release zone to deposit in almost all cases. Additionally, research suggests SAR to be a lot less reliable for detecting dry snow avalanches compared to wet snow avalanches \mbox{%DIFAUXCMD

\citep[among others][]{hafner.2021, eckerstorfer.2022}}\hspace{0pt}%DIFAUXCMD

. The above statements made about SPOT 6/7 are transferable to optical data with similar or better spatial and spectral resolution.

}

\DIFaddend To bypass the time-consuming manual mapping, several groups have explored (semi-) automatic mapping approaches. \citet{Buhler.2009} used a processing chain that relies on directional, textural and spectral information to automatically detect avalanches in airborne optical data. \citet{Lato.2012} and \citet{Korzeniowska.2017} applied object-based classification techniques to optical high spatial resolution data (0.25 - 0.5 m). \citet{Wesselink.2017} and \citet{Eckerstorfer.2019} have introduced and consequently refined an algorithm to automatically detect avalanches in Sentinel-1 SAR imagery, via changes of the backscatter between pre- and post-event images. \citet{karbou.2018} also utilized changes in backscatter to identify avalanche debris. For avalanche detection in Radarsat-2 imagery, \citet{Hamar.2016} used supervised classification with a random forest classifier. On the contrary, the avalanche mapping from optical satellite data has so far been exclusively done manually \citep{Buhler.2019, hafner.2021, Abermann.2019}. \\

The deployment of machine learning for remote sensing image analysis has seen a surge in the last decade \citep{Ma.2019}. Modern deep learning methods often outperform competing ones in complex image understanding tasks, and have been used for example to detect rock glaciers \citep{Robson.2020}, landslides \citep{prakash.2021} and crop types in fields \citep{cai.2018}. For avalanches, the use of deep learning has so far focused on Sentinel-1 imagery: \citet{waldeland.2018} applied a pre-trained ResNet \citep{he.2016} for avalanche identification by change detection using manual reference annotations. \citet{bianchi.2021} segmented avalanches with a fully convolutional U-Net \citep{ronneberger.2015}, also relying on manual annotations for training the network. \citet{sinha.2019a} proposed a fully convolutional VGG16 network \citep{simonyan.2015} that was trained on, and compared against, an inventory of avalanche field observations. With the same inventory, \citet{sinha.2019b} also, alternatively used a Variational Autoencoder \citep{Kingma.2019} for avalanche detection.

In contrast to previous studies, our work is the first to attempt to use deep learning for the detection of avalanches in \emph{optical} satellite data. This is of major importance, as the largest avalanche mapping from remotely sensed imagery to date, with 24'778 single avalanche polygons \DIFdelbegin \DIFdel{\mbox{%DIFAUXCMD

\citep{spot.2018, spot.2019}}\hspace{0pt}%DIFAUXCMD

}\DIFdelend \DIFaddbegin \DIFadd{\mbox{%DIFAUXCMD

\citep{Buhler.2019, spot.2018, spot.2019}}\hspace{0pt}%DIFAUXCMD

}\DIFaddend , relied on optical SPOT 6/7 satellite imagery. Furthermore, there have been investigations with external data into the reliability and completeness of mappings from SPOT 6/7 \citep{hafner.2021}. Consequently, an automation of the manual mapping from this imagery would allow for a fast comprehensive documentation of future avalanche periods with background knowledge about how well it works and how much avalanche area approximately is missed. Without an automation it is not feasible to cover large regions quickly. With manual image interpretation \citep{hafner.2021} it took approximately one hour to manually delineate avalanches in SPOT images covering a region of $\approx\,$27.5~km^2. Thus, in this work we develop, describe and apply a deep learning approach for avalanche mapping based on the SPOT 6/7 sensor with the goal to automate the mapping process, so as to cover large areas and eventually operate at country-scale. We developed a variant of DeepLabV3+ \citep{chen.2018} that takes as input SPOT 6/7 images and a digital elevation model (DEM), and outputs spatially explicit raster maps of avalanches. For our DeepLabV3+ variant we made the encoder and decoder deformable \citep{Dai.2017}, thereby our convolutional kernels adapt according to the underlying terrain, which is essential in the study of avalanches. In addition to a careful description of the network architecture we evaluate results, compare to previous work, examine the reproducibility of the manually mapped avalanches, and discuss the potential and limitations of our method.

\section{Data}\label{subsub:data}

For training and validating our proposed mapping system we utilize SPOT 6/7 \DIFaddbegin \DIFadd{top-of-atmosphere reflectance }\DIFaddend images acquired on 24 January 2018 \citep[referred to as 2018 in the remainder of this paper,][]{spot.2018} and 16 January 2019 \citep[referred to as 2019 from now on,][]{spot.2019}, together with a set of 24'776 avalanche annotations delineated by manual photo-interpretation. In both cases the images were acquired after periods with very high avalanche danger, i.e., the maximum level 5 of the Swiss avalanche warning system \citep{slf.2021}. SPOT 6/7 images have a ground sampling distance (GSD) of 1.5~m and provide \DIFdelbegin \DIFdel{spectral intensities in the }\DIFdelend \DIFaddbegin \DIFadd{information in four spectral bands, namely }\DIFaddend red, green, blue, and near-infrared (R, G, B, NIR)\DIFdelbegin \DIFdel{wavelengths}\DIFdelend , at a radiometric resolution of 12 bits. The dataset covers an area of $\approx\,$12’500~km^2 in 2018 and $\approx\,$9’500~km^2 in 2019. These two areas partly overlap. \DIFdelbegin \DIFdel{Snow reflectance and atmospheric influences }\DIFdelend \DIFaddbegin \DIFadd{As both were acquired in January, the illumination conditions }\DIFaddend exhibit little variability between the two years, but they differ in terms of snow conditions: in 2019 the snow line was at a lower altitude, and consequently there was more dry snow, hardly any wet snow, and fewer glide snow avalanches. %

As additional input information we use the Swiss national DEM \emph{swissALTI3D}. To match the resolution of SPOT imagery, we resample the DEM (original GSD 2~m) to 1.5~m, aligned with SPOT 6/7. Its nominal vertical accuracy is 0.5 m below the treeline (∼2100 m a.s.l.) and 1–3 m above the treeline \citep{swisstopo.2018}. We did not apply atmospheric corrections as \DIFdelbegin \DIFdel{the water content of the atmosphere is typically low and atmospheric effects therefor relatively minor in winter \mbox{%DIFAUXCMD

\citep{nolin.2010}}\hspace{0pt}%DIFAUXCMD

. }\DIFdelend \DIFaddbegin \DIFadd{our main focus is texture and the absolute spectral values do not matter for avalanche identification.

}\DIFaddend

\DIFaddbegin \DIFadd{The 24'776 avalanches were annotated by a single person, an expert, which we define as somebody very familiar with both satellite image interpretation and avalanches. For the mapping of avalanches the visual identification of crown and release areas, track and deposit through texture and hue as well as hints of possible damage have played a role \mbox{%DIFAUXCMD

\citep[for details on the methodology see][]{Buhler.2019}}\hspace{0pt}%DIFAUXCMD

. }\DIFaddend For each mapped avalanche polygon the expert also recorded a score of how well the avalanche was visible, splitting the annotations in three groups: \emph{complete, well visible} outline; \emph{mostly well visible} outline; and \emph{not completely visible} outline, where significant parts had to be inferred with the help of domain knowledge \citep[see also][]{Buhler.2019}. \DIFdelbegin \DIFdel{The methodology for manual avalanche mapping follows \mbox{%DIFAUXCMD

\citet{Buhler.2019} }\hspace{0pt}%DIFAUXCMD

and is described in more detail there. }\DIFdelend Furthermore, we validated a subset of the initial mapping with independent ground- and \DIFdelbegin \DIFdel{helicoper-based }\DIFdelend \DIFaddbegin \DIFadd{helicopter-based }\DIFaddend photographs as reference \citep{hafner.2021}. We found that for manual mapping based on SPOT images the probability of detection (POD; \DIFaddbegin \DIFadd{see Equation~\ref{equ:PODPPV}; }\DIFaddend the probability of a true avalanche being annotated) is \DIFdelbegin \DIFdel{74\% }\DIFdelend \DIFaddbegin \DIFadd{0.74 }\DIFaddend for avalanches larger than size 1 \citep[avalanche size is categorised on a scale from 1 to 5, with size 5 the largest and most destructive ones; for more details see][]{slf.2021}. The positive predictive value (PPV; \DIFaddbegin \DIFadd{see Equation~\ref{equ:PODPPV}; }\DIFaddend probability of an annotated avalanche having a true counterpart) was \DIFdelbegin \DIFdel{88\%}\DIFdelend \DIFaddbegin \DIFadd{0.88}\DIFaddend , indicating only few false positive annotations (again for size $\geq$2).

\DIFaddbegin \DIFadd{Additionally, we used SPOT 7 imagery of the Mattertal, Val d'Hérens and Val d'Herémence in Valais, Switzerland from 6 January 2018 covering $\approx\,$660 km^2 to evaluate our model. The data were acquired for test purposes after a period with high avalanche danger and the 538 avalanches used for validation have been manually mapped with the same methodology as the others used in this work and described in \mbox{%DIFAUXCMD

\citet{Buhler.2019}}\hspace{0pt}%DIFAUXCMD

. The geographical region with additional data overlaps with data acquired on 24 January 2018, but served as test area before and did not go into training or validation (see "Generalitzation Test" areas in Figure~\ref{fig:TrainCoAreas}). The images suffer from distortion in steep terrain as they were part of a suitability study for avalanche mapping from optical data \mbox{%DIFAUXCMD

\citep }\hspace{0pt}%DIFAUXCMD

}[\DIFadd{for details see}][]{\DIFadd{Buhler.2019}} \DIFadd{and orthorectified by the satellite providers using the height information from the Shuttle Radar Topography Mission \mbox{%DIFAUXCMD

\citep[SRTM,][]{srtm.2013}}\hspace{0pt}%DIFAUXCMD

.

}

\DIFaddend \section{Method}

Many overlapping avalanches exist in the dataset whose boundaries cannot be precisely distinguished from each other even by experts. We thus restrict ourselves to identifying all pixels where avalanches have occurred, but do not attempt to group them into individual avalanche events.

In terms of image analysis this corresponds to a semantic segmentation task, where each pixel is assigned a class label: \emph{avalanche} or \emph{background} according to \DIFdelbegin \DIFdel{its class score}\DIFdelend \DIFaddbegin \DIFadd{the model confidence}\DIFaddend .

Several deep learning models have been developed for solving such problems and have achieved excellent results in various domains, such as U-Net \citep{ronneberger.2015}, HRNetV2 \citep{sun.2019} and DeepLabV3+ \citep{chen.2018}.

\subsection{Model Architecture}

On their way downwards, avalanches are constrained and guided by the local terrain. In order to accurately map avalanches from the input data, we therefore propose a deep learning architecture that adapts to the underlying terrain model. We build on state-of-the-art model DeepLabV3+ designed for semantic segmentation and add deformable convolutions that adapt their receptive field size according the input data, i.e. the terrain model in our case.

\textbf{DeepLabV3+} is a popular, fully convolutional semantic segmentation model that has been used successfully with a variety of datasets. It features a dilated ResNet \DIFaddbegin \DIFadd{\mbox{%DIFAUXCMD

\citep{he.2016} }\hspace{0pt}%DIFAUXCMD

}\DIFaddend encoder as a backbone for feature extraction, in combination with Atrous Spatial Pyramid Pooling module (ASPP). To achieve a wide receptive field able to capture multi-scale context, ASPP employs dilated convolutions at different rates. Before being fed into the decoder, the resulting features are concatenated and merged using a 1$\times$1 convolution. These high-level features are then decoded, upsampled and combined with high-resolution, low-level features from the first encoder layer. For further details about DeepLabV3+, see \citet{chen.2018}.

Our adaptions to the standard DeepLabV3+ include: \emph{deformable kernels} \citep{Dai.2017} in the encoder and decoder as well as a small network with offsets that estimates the appropriate kernel deformations in a data-driven manner, and modifies the decoder such that it can process features from \emph{all backbone layers} (Figure~\ref{fig:AvanetOverview} \DIFaddbegin \DIFadd{and Figure~\ref{fig:DefConvs}}\DIFaddend). These changes add a modest 1.9M network weights to the 22.4M weights of the standard DeepLabV3+.

The reasoning behind \emph{deformable convolution kernels in the backbone} (Figure~\ref{fig:Encoder}) is to adapt their receptive fields to the underlying terrain. To obtain deformable convolutions, we introduce an additional 18-channel tensor that encodes the 2D offset of each kernel element at each location i.e., it enables free-form deformations of the kernel, beyond dilation or rotation. The offsets are not fixed \DIFdelbegin \DIFdel{a-priori}\DIFdelend \DIFaddbegin \DIFadd{a priori}\DIFaddend , but calculated as a learned function of the DEM, separately for each feature resolution, by a small additional network branch.

By replacing the first convolution in each residual block with a deformable one, we are able to explicitly include the terrain shape encoded in the DEM, but without the need to modify other parts of the architecture, so as to benefit from the pretrained weights of the encoder.

\begin{figure}[t]

\includegraphics[width=17cm]{AvanetOverview_hori.png}

\centering

\caption{Overview of our DeepLabV3+ variant. The encoder is shown in more detail in Figure~\ref{fig:Encoder} and the Deformable Spatial Pyramid Flow (DSPF) in Figure~\ref{fig:AvanetDSPF}.}

\label{fig:AvanetOverview}.

\end{figure}

\begin{figure}[t]

\includegraphics[width=15cm]{DefConvs1.png}

\centering

\caption{For the deformable convolutions, a standard kernel (like the 3×3 as shown in a) will be adapted according to 2D offsets learned from the underlying DEM. The green dots in b, c and d exemplarily show possible final positions of the kernel elements, the displacement from the standard kernel is illustrated by the black arrows. }

\label{fig:DefConvs}.

\end{figure}

The \emph{augmented decoder} helps our DeepLabV3+ to propagate features along specific directions, in our case this is the possible downhill flow direction of avalanches which can be extracted from the DEM. Hence, we alter the ASPP such that it aggregates features from \emph{all} backbone layers, and \DIFdelbegin \DIFdel{increase }\DIFdelend \DIFaddbegin \DIFadd{increases }\DIFaddend the receptive field. The new module, which we call \emph{Deformable Spatial Pyramid Flow} (DSPF, Figure~\ref{fig:AvanetDSPF}), performs deformable convolutions at different dilation rates. The deformations are again obtained from our small network with offsets, based on the DEM. In order to propagate information along the gradient field, we also model the flow direction of an avalanche in the DSFP module of the decoder.

\begin{figure}[t]

\includegraphics[width=18cm]{Encoder.png}

\centering

\caption{Encoder of our DeepLabV3+ in detail.}

\label{fig:Encoder}.

\end{figure}

\begin{figure}[t]

\includegraphics[width=13cm]{DSPF.png}

\centering

\caption{Detailed architecture of the Deformable Spatial Pyramid Flow (DSPF) used in the Decoder of our DeepLabV3+ variant.}

\label{fig:AvanetDSPF}.

\end{figure}

\subsubsection{Sampling and Data Split} \label{subsub:dataSampling}

\DIFdelbegin \DIFdel{Satellite images are normally too large to be processed with deep learning methods all at onceand are therefore cropped to patchesof smaller size. Furthermore, a frequent challenge in }\DIFdelend \DIFaddbegin \DIFadd{Given the proposed model architecture and the available computational resources (CPU: 20 Intel Core 3.70 GHz,

GPU: 1 NVIDIA GeForce RTX 2080 Ti), we are unable to process an entire orthomosaic at once. Therefore, we process squared image subsets, called patches, of up to 512×512 pixels at training time, which translates into an area of 589'824~m^2 at the spatial resolution of SPOT 6/7 imagery. With our model and computational resources we can simultaneously process batches of 2 image patches per GPU.

}

\DIFadd{For }\DIFaddend supervised machine learning approaches \DIFaddbegin \DIFadd{it is vitally important that all desired classes are present in the patches the model learns from. As classes are usually not evenly distributed, class imbalance is a frequent challenge. Our dataset }\DIFaddend is \DIFdelbegin \DIFdel{the class imbalance. In our dataset it is }\DIFdelend very imbalanced: avalanches cover only one 1785th of the entire area covered by SPOT 6/7 imagery. Re-balancing of class frequencies is necessary to make sure our model adequately captures the variability of the avalanche class.

We use the following pragmatic strategy to ensure a training set that includes relevant examples, and with sufficient representation of both classes:

%

First, we iteratively sample patch centers inside \DIFaddbegin \DIFadd{manually annotated }\DIFaddend avalanche polygons, while avoiding overlapping patches. In this way, we obtain a set of samples that is not overly imbalanced, with $\approx$3.5$\times$ more background pixels than avalanche pixels. These patches form 95\% of our training set. \DIFdelbegin \DIFdel{The }\DIFdelend \DIFaddbegin \DIFadd{Second, the }\DIFaddend remaining 5\% are sampled randomly in areas without avalanches, to ensure also patches without avalanche pixels are seen during training. This leads to an effective ratio of 1:4 between avalanche and background pixels in the 5185 512×512 patches of the training set.

As the edges of the patches lack context, they were also given smaller weights when calculating the loss function during training starting 100 pixels from the edge, decreasing the weight linearly to 10\% of the base weight given above at the very edge. For our DeepLabV3+ we additionally used deep supervision as in \citet{simonyan.2015}, to help the model converge.

\DIFdelbegin \DIFdel{To increase the model's performance, gradients are accumulated over two iterations before weights are updated. Thereby an effective batch size of 16 (8+8) is reached and the 512×512 pixel patches may be used (see also section~\ref{subsub:ablation}). Predictions are

```latex
made for an area specified by a shapefile. To reduce artifacts at the edges of patches, the samples for the predictions overlap by 100 pixels before being cropped.

}\DIFdelend


\subsubsection{Training}\label{sub:trainingDetails}

For training and quantitative evaluation, the data were split into mutually exclusive, geographically disjoint regions for training (80\%), validation and hyper-parameter tuning (10\%) and testing (10\%), as depicted in Figure~\ref{fig:TrainCoAreas}. The test set is located completely in regions acquired either only in 2018 or only in 2019, but not in the overlap between the two acquisitions, to prevent memorization (especially of the identical topography).


\begin{figure}[t]

\includegraphics[width=16cm]{TrainTestValiAreas1a.png}

\centering

\caption{Visualization of the disjoint regions for training, validation and testing for both 2018 and 2019. Also shown are: the test region for the generalization experiments, where we had additional data from 6 January 2018, and the regions used to study reproducibility of manual avalanche maps.}

\label{fig:TrainCoAreas}.

\end{figure}


The network is trained by minimizing a weighted binary cross entropy (BCE) loss (see also ~\ref{sub:trainingDetails}), using the Adam optimizer~\citep{kingma.2017} for 20 epochs. The base learning rate was initialised to 1$\times$10$^{-4}$ and reduced by a factor of 4 after 10 epochs. A summary of the hyper-parameter settings is given in Table~\ref{tab:networkParams}.


\begin{table}[!ht]

  \centering

  \caption{Summary of training parameters}

  \label{tab:networkParams}

  \begin{tabular}{|l|c|}

  \hline

    \textbf{Parameter} & \textbf{Value} \\ \hline\hline

    Loss function & Weighted BCE \\ \hline

    Optimizer & Adam \\ \hline

    Initial learning rate & 1$\times$10$^{-4}$ \\ \hline
```

Effective batch size & 16 \\ \hline

Patch size & 512$\times$512 \\ \hline

Epochs & 20 \\ \hline

\end{tabular}

\end{table}

As a preprocessing step, the input images are normalized channel-wise using the mean and variance values of the entire dataset.

%DIF < \ks{slightly vague. Perhaps better "are normalised by mapping the value range of the image dataset to [-?..?]"}

Additionally, we flattened the peak in the image histograms caused by the shadow pixels by transforming negative values $v\rightarrow (-3\cdot v^2)$, while keeping positive values unchanged.

Even though our training dataset is large, it covers only two avalanche periods and cannot be expected to account for the whole variety of possible conditions. In order to increase the robustness of the network, we further expand the training set with synthetic data augmentation. We used randomized rotation and flipping for greater topographic variety, mean-shifting and variance-scaling to simulate varying atmosphere and lighting conditions, as well as patch shifting to increase robustness when only part of an avalanche is visible. To speed up data loading we used batch augmentation \citep{hoffer.2019}, where the same sample is read only once and used multiple times with different augmentations computed on the fly. \DIFaddbegin \DIFadd{To increase the model's performance, we additionally accumulated gradients over two iterations before weights were updated. Thereby an effective batch size of 4 (2+2) was reached and the 512×512 pixel patches may be used (see also Section~\ref{subsub:ablation}).

}\DIFaddend

As mentioned in Section~\ref{subsub:data} the avalanche polygons come with labels that quantify their visibility in the SPOT data. These labels are used to re-weight their contributions to the BCE loss as follows: pixels on \emph{complete, well visible} avalanches have weight 2, \emph{mostly well visible} avalanches as well as \emph{background} pixels not on an avalanche have weight 1, and \emph{not completely visible} avalanches have weight 0.5.

\section{Results and Discussion}

\DIFdelbegin \DIFdel{To }\DIFdelend \DIFaddbegin \DIFadd{Predictions are made for a target area specified by vector polygons in the form of shapefiles. To reduce artifacts at the edges of patches, the samples for the predictions overlap by 100 pixels before being cropped. To }\DIFaddend assess the detection performance of the network, we calculated positive predictive value (PPV, also called precision) and probability of detection (POD, also called recall) on a pixel level as well as the F1-score. PPV and POD are both based on a standard 2×2 confusion matrix \citep{Trevethan.2017}. As per pixel metrics take as input a binary mask (avalanche yes or no) and the network yields scores, we thresholded the predictions at 0.5 before calculating statistics and computed the F1-score as

\begin{equation}

\text{F1} = 2 \cdot \frac{\text{PPV} \cdot \text{POD}}{\text{PPV} + \text{POD}},

\label{equ:F1}

\end{equation}

where POD and PPV are defined as

\begin{equation}

\text{POD} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{and} \qquad \text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}},

\DIFaddbegin \label{equ:PODPPV}

\DIFaddend \end{equation}

where TP is true positive, FP is false positive and FN is false negative.

In this paper the presented pixel-wise metrics (POD, PPV and F1-score) represent the average score over all the patches we tested on. As our dataset is imbalanced and the F1 score non-symmetric, we calculated those metrics for both avalanches and the background. Additionally, we wanted to estimate how many avalanches were detected by each model. Consequently, for the object-based metrics we tested two different measures: we counted an avalanche as detected if 50\% or 80\% of all pixels within an avalanche from the manual mapping had a score of 0.5 or higher.

\subsection{Results and generalization ability}

\DIFaddbegin \DIFadd{Results were calculated for the test areas and are reported in }\DIFaddend Table~\ref{tab:networkResults}\DIFdelbegin \DIFdel{demonstrates that a }\DIFdelend \DIFaddbegin \DIFadd{. Compared to the }\DIFaddend standard DeepLabV3+\DIFdelbegin \DIFdel{has a lower POD (}\DIFdelend \DIFaddbegin \DIFadd{, our model, when run with the parameters described in Table~\ref{tab:networkParams}, has a higher POD for avalanches (0.610 vs. }\DIFaddend 0.587)\DIFdelbegin \DIFdel{than our method (0.610) }\DIFdelend \DIFaddbegin \DIFadd{, }\DIFaddend while having the same PPV\DIFdelbegin \DIFdel{, both for avalanches}\DIFdelend . This results in an F1-score of 0.612 for the standard DeepLabV3+ and 0.625 for our version\DIFdelbegin \DIFdel{with the F1-score for the background being almost identical for both models. For }\DIFdelend \DIFaddbegin \DIFadd{. For the }\DIFaddend background, the pattern is similar, the POD is slightly better for our method (0.894), compared to the standard DeepLabV3+ (0.888), while the PPV is slightly higher for the standard model (0.900 vs. 0.894). \DIFdelbegin \DIFdel{The }\DIFdelend \DIFaddbegin \DIFadd{Consequently, the }\DIFaddend F1-score is \DIFdelbegin \DIFdel{again }\DIFdelend very similar, as it only differs by one in the third decimal place between our and the standard DeepLabV3+.

For any supervised classification and deep learning methods in particular, the ability to generalize well to new datasets and regions not seen during the training phase is key. To evaluate this, we test our trained model using \DIFdelbegin \DIFdel{unseen SPOT 6 imagery of the Mattertal, Val d'Hérens and Val d'Herémence in Valais, Switzerland }\DIFdelend \DIFaddbegin \DIFadd{SPOT 7 imagery }\DIFaddend from 6 January \DIFdelbegin \DIFdel{2018 covering $\approx\,$660 km^2 (see Figure~\ref{fig:GenMattertal}). The data were acquired for test purposes after a period with high and very high avalanche danger and the avalanches used for validation have been manually mapped with the same methodology as the others used in this work and described in \mbox{%DIFAUXCMD

\citet{Buhler.2019}}\hspace{0pt}%DIFAUXCMD

. The geographical region with additional data overlaps with data acquired on 24 January 2018, but served as test area before and did not go into training or validation (see Figure~\ref{fig:TrainCoAreas}). The images suffer from distortion in steep terrain as they were part of a usability study for avalanche mapping from optical data \mbox{%DIFAUXCMD

\citep }\hspace{0pt}%DIFAUXCMD

}%DIFDELCMD < [%%%

\DIFdel{for details see}%DIFDELCMD <][]{%%%

\DIFdel{Buhler.2019}%DIFDELCMD < } %%%

\DIFdel{and orthorectified by the satellite providers using the height information from the Shuttle Radar Topography Mission \mbox{%DIFAUXCMD

\citep[SRTM,][]{srtm.2013}}\hspace{0pt}%DIFAUXCMD

.

}%DIFDELCMD <

%DIFDELCMD < %%%

\DIFdel{The test }\DIFdelend \DIFaddbegin \DIFadd{2018. The test }\DIFaddend metrics for predictions on the data from 6 January 2018 were calculated with the standard DeepLabV3+ and the adapted DeepLabV3+. As Table~\ref{tab:networkResults} shows, our version generalizes very well \DIFaddbegin \DIFadd{(see also Figure~\ref{fig:GenMattertal})}\DIFaddend , the metrics only differ from tests on the initial dataset in the fourth decimal place. The standard DeepLabV3+ on the other hand, does not generalize so well as the POD and the detection rates per avalanche are lower than for testing on the initial data\DIFdelbegin \DIFdel{(see Table~\ref{tab:networkResults})}\DIFdelend .

We also investigated object-based metrics for all model variations, when detection means 50\% of the avalanche area the models rightly capture between roughly 58\% and 69\% of all avalanches and between 38\% and 51\% when detection requires 80\% of the area (Table~\ref{tab:ObasedMetrics}). Again the the standard DeepLabV3+ performs slightly worse than our adapted DeepLabV3+, especially when run on data from a new avalanche period (6 January 2018). Therefore, our DeeplabV3+ shows better ability to generalize to new and previously unseen data. \DIFdelbegin \DIFdel{The }\DIFdelend \DIFaddbegin \DIFadd{Overall, the }\DIFaddend best performance is achieved when considering sunlit avalanche parts only, for both training and testing.

```
\begin{sidewaystable}[!ht]

  \centering

  \caption{Segmentation results for the test areas for the standard DeepLabV3+ and our DeepLabV3+, results of predicting on data from a previously unseen avalanche period and variations to our model for the ablation. The metrics shown are the averages from all tested patches. \DIFaddbegin \DIFadd{The bold fonts signify the model, data, and parameters that were varied compared to what we used as standard in our work.}\DIFaddend }

  \label{tab:networkResults}

  \begin{tabular}{ p{1.6cm} p{1.8cm} p{1.4cm} p{1.4cm} p{2cm} p{1.4cm} p{1cm} p{1cm} p{1cm} p{1cm} p{1cm} p{1cm}}

  \hline

    Model & SPOT data & training data & patch size & bands &loss function & POD & PPV & F1 & POD back & PPV back & F1 back  \\ \hline\hline

    \textbf{standard DeepLabV3+} & 24.01.2018 16.01.2019 & whole avalanches& 512$\times$512 & R, NIR, DEM & weighted BCE & 0.587 & 0.667 & 0.612 & 0.888 & 0.900 & 0.885 \\ \hline

    adapted DeepLabV3+ & 24.01.2018 16.01.2019 & whole avalanches& 512$\times$512 & R, NIR, DEM & weighted BCE & 0.610 & 0.668 & 0.625& 0.894 & 0.894 & 0.884 \\ \hline

    \textbf{standard DeepLabV3+} & \textbf{06.01.2018} & whole avalanches & 512$\times$512 & R, NIR, DEM & weighted BCE & 0.547 & 0.667 & 0.591 & 0.876 & 0.916 & 0.887 \\ \hline

    adapted DeepLabV3+ & \textbf{06.01.2018} & whole avalanches & 512$\times$512 & R, NIR, DEM & weighted BCE & 0.610 & 0.668 & 0.625 & 0.894 & 0.895 & 0.884 \\ \hline

    adapted DeepLabV3+ & 24.01.2018 16.01.2019 & whole avalanches& {\bf 256$\times$256} & R, NIR, DEM & weighted BCE & 0.723 & 0.587 & 0.645 & 0.720 & 0.796 & 0.720 \\ \hline

    adapted DeepLabV3+ & 24.01.2018 16.01.2019 & whole avalanches& {\bf 128$\times$128} & R, NIR, DEM & weighted BCE & 0.898 & 0.659 & 0.829 & 0.340 & 0.551 & 0.452 \\ \hline

    adapted DeepLabV3+ & 24.01.2018 16.01.2019 & whole avalanches& 512$\times$512 & R, NIR, DEM & \textbf{unweighted BCE} & 0.688 & 0.575 & 0.622 & 0.887 & 0.908 & 0.888 \\ \hline

    adapted DeepLabV3+ & 24.01.2018 16.01.2019 & whole avalanches& 512$\times$512 & \textbf{R, G, B, NIR, DEM} & weighted BCE & 0.559 & 0.682 & 0.613 & 0.883 & 0.913 & 0.889 \\ \hline
    adapted DeepLabV3+ & 24.01.2018 16.01.2019 & whole avalanches& 512$\times$512 & \textbf{R, G, B, NIR, DEM, Wallis-filtered} & weighted BCE & 0.610 & 0.668 & 0.597 & 0.876 & 0.906 & 0.880 \\ \hline

    adapted DeepLabV3+ & 24.01.2018 16.01.2019 & \textbf{release area only}& 512$\times$512 & R, NIR, DEM & weighted BCE & 0.053 & 0.665 & 0.183 & 0.777 & 0.992 & 0.856 \\ \hline

    adapted DeepLabV3+ & 24.01.2018 16.01.2019 & \textbf{deposits only}& 512$\times$512 & R, NIR, DEM & weighted BCE & 0.196 & 0.788 & 0.347 & 0.797 & 0.986 & 0.868 \\ \hline

    adapted DeepLabV3+ & 24.01.2018 16.01.2019 & \textbf{sunlit avalanche parts only} & 512$\times$512 & R, NIR, DEM & weighted BCE & 0.668 & 0.653 & 0.639 & 0.918 & 0.910 & 0.907 \\ \hline

  \end{tabular}

\end{sidewaystable}


\clearpage


\begin{table}[!ht]

  \centering

  \caption{Object-based metrics for selected model configurations.}

  \label{tab:ObasedMetrics}

  \begin{tabular}{p{2cm} p{1.8cm} c p{2.5cm} p{2.5cm}}

  \hline

    Model & SPOT data & Training data & Detection rate 50\% of avalanche area & Detection rate 80\% of avalanche area \\ \hline\hline

    standard DeepLabV3+ & 24.01.2018 16.01.2019 & whole avalanches & 0.63 & 0.45 \\ \hline

    adapted DeepLabV3+ & 24.01.2018 16.01.2019 & whole avalanches & 0.66 & 0.46 \\ \hline

    standard DeepLabV3+ & 06.01.2018 & whole avalanches & 0.58 & 0.38 \\ \hline

    adapted DeepLabV3+ & 06.01.2018 & whole avalanches & 0.66 & 0.46 \\ \hline

    adapted DeepLabV3+ & 24.01.2018 16.01.2019 & sunlit avalanches only & 0.69 & 0.51 \\ \hline

  \end{tabular}

\end{table}


\begin{figure}[t]

\DIFdelbeginFL %DIFDELCMD < \includegraphics[width=14cm]{20180106_Scores_Mattertal.png}

%DIFDELCMD < %%%
```

\DIFdelendFL \DIFaddbeginFL \includegraphics[width=14cm]{fig06_new.png}

\DIFaddendFL \centering

\caption{An example for the \DIFdelbeginFL \DIFdelFL{scores }\DIFdelendFL \DIFaddbeginFL \DIFaddFL{model confidence }\DIFaddendFL when predicting on data from a previously unseen avalanche period from 6 January 2018 (SPOT 6 data © Airbus DS 2018). The \DIFaddbeginFL \DIFaddFL{values closer to 1, in }\DIFaddendFL darker hues\DIFaddbeginFL \DIFaddFL{, }\DIFaddendFL indicate places where the model is more confident \DIFaddbeginFL \DIFaddFL{about the existence of an avalanche}\DIFaddendFL . In the illuminated regions those areas almost always \DIFdelbeginFL \DIFdelFL{correspond to }\DIFdelendFL \DIFaddbeginFL \DIFaddFL{overlap with }\DIFaddendFL manually mapped avalanches.}

\label{fig:GenMattertal}.

\end{figure}

\subsection{Ablation studies} \label{subsub:ablation}

To understand how our changes to the standard DeepLabV3+ affect performance we varied the model in different ways and trained, tested and compared the performance. These results can be found in Table~\ref{tab:networkResults}. First, we investigated the influence of the deformable backbone and discovered that including it outperforms the non-deformable backbone configurations of the standard DeepLabV3+. This is the case in our test areas for 2018 and 2019, but also for testing on the avalanche period from 6 January 2018. Secondly, the avalanches in our network have been weighted (see~\ref{sub:trainingDetails}) according to the quality index assigned by the manual mapper. To quantify the effects of using weights we ran training with unweighted BCE and observed a decrease in POD, a slight increase in PPV and overall a smaller F1-score. Additionally, in our adapted version of DeepLabV3+ we only considered the Red and Near-Infrared from SPOT as well as the DEM as input channels. We cannot test the adapted DeepLabV3+ without the DEM, as it is explicitly included as an integral part of the network. We analyzed however, how including all SPOT channels (additionally Blue and Green) and also adding another Wallis filtered channel (to bring out details in the shade) affects network performance (see Table~\ref{tab:networkResults}). For our model we found that including more channels did not improve the performance, rather training time was longer and metrics worse than with the initial channels.\\

We hypothesize that the proportion of potential avalanche area and context visible in the patches strongly influences network output. To investigate this, we have trained our model with varying patch size: \$512\times512\$, \$256\times256\$ and \$128\times128\$ pixels (corresponding to \$768\times768\$, \$384\times384\$ and \$192\times192\$ meters). Quantitative results in Table~\ref{tab:networkResults} show the largest patch size performs best considering metrics for both avalanches and background. When comparing them visually (Figure~\ref{fig:PatchVari}) this is further supported, as the predictions on the smallest size are patchy, dispersed over the image showing the model is unsure about the occurrence of avalanches. With increasing context through a larger patch size though, the model becomes more confident and the avalanche borders are distinctly visible.

Subsequently, in order to understand what is better for training the network, we trained on avalanche deposits or release areas only. As deposit area, we assumed the lower third (based on elevation) of each manually mapped avalanche, ignoring those avalanches were the deposit had been inferred. For the release areas, we used the zones identified by \citet{Buhler.2019}, again disregarding those avalanches where the release zone had been inferred and were therefore uncertain. As results in Table~\ref{tab:networkResults} show, performance for predicting all avalanches is a lot worse in both cases. We also observe that PPV and POD are significantly higher when the network is trained on deposits only, rather than trained only on release areas. This resulted in an increase of 0.146 in F1-score and suggests that the original model might also be learning more from texture rich avalanche deposits than from release zones.

\begin{figure}[t]

% \includegraphics[width=17.5cm]{QualitativeExamples.png}

\includegraphics[width=17.5cm]{Fig7_Qualitative_new.png}

\centering

\caption{Comparison of results for four patches when training the network with different patch sizes. The tiles depict (a) the SPOT 6 image, (b) the manually mapped annotations used as reference, (c) the predictions thresholded at 0.5, and (d) the predicted avalanche probability (SPOT 6 data © Airbus DS 2018). Visual inspections show, the model is a lot more confident the larger the patch size.}

\label{fig:PatchVari}.

\end{figure}

Finally, the experts manually mapping the avalanches generally perceived those in the sun as better visible. \citet{hafner.2021} confirmed that and found the POD to be higher roughly by a factor five for avalanches in fully illuminated terrain compared to those, at the time of image acquisition, in fully shaded terrain. In order to investigate this further, we used a Support Vector Machine (SVM) classifier to calculate a shadow mask for both 2018 and 2019. The mask also includes most forested areas due to their speckled sun/shade pattern. Subsequently, we excluded the avalanche parts being located in the shade and trained only with the remaining areas (about one fourth of the avalanche area per year). Calculating the metrics considering only avalanches in illuminated areas, we found an increase of 0.058 in POD, a slight decrease of 0.015 in PPV and consequently an increase in F1-score of 0.014. The object-based metrics (Table~\ref{tab:ObasedMetrics}) are also slightly better when only considering sunlit regions.

\subsection{Reproducibility of manually mapped avalanches}

To assess the degree of label noise in our dataset, we conducted a reproducibility experiment on the manually mapped avalanches to understand how similar the assessment of a given area by different experts would be. In other fields several comprehensive studies have already been conducted to investigate inter-observer variability, for example for contouring organs in medical images

\citep{fiorino.1998} or for manual glacier outline identification \citep{paul.2013}. For our investigation five people attempted to replicate the manual mapping with the same methodology as used before and described in detail in \citet{Buhler.2019}. All five mapping experts are very familiar with satellite imagery and/or avalanches and received the same standardized introductions. The experiment was conducted twice in an area of 90 km^2 around Flims, Switzerland, on the 2018 and 2019 SPOT 6/7 imagery (see Figure~\ref{fig:TrainCoAreas}). The area contains avalanches in the shade and in illuminated terrain as well as all outline quality classes in the initial mappings \citep{spot.2018, spot.2019}. The mapping experts did not see another mapping before having finished theirs.

```
\begin{table}[!ht]
  \centering
  \caption{F1-scores for the reproducibility investigation: the bold values in the upper right part of the table represent the scores comparing two expert mappings in illuminated terrain, the lower left values the scores in shaded terrain.}
  \label{tab:ReproResults}
  \begin{tabular}{c| c c c c c}
  \hline
     & Expert 1 & Expert 2 & Expert 3 & Expert 4 & Expert 5 \\ \hline\hline
   Expert 1 & & {\bf 0.758} & {\bf 0.623} & {\bf 0.617} & {\bf 0.653} \\
   Expert 2 & 0.401 &  & {\bf 0.711} & {\bf 0.723} & {\bf 0.724} \\
   Expert 3 & 0.232 & 0.198 &  & {\bf 0.656} & {\bf 0.782} \\
   Expert 4 & 0.188 & 0.236 & 0.205 &  & {\bf 0.786} \\
   Expert 5 & 0.123 & 0.155 & 0.204 & 0.244 &  \\ \hline
  \end{tabular}
\end{table}
```

```
\begin{figure}[t]
\DIFdelbeginFL %DIFDELCMD < \includegraphics[width=12cm]{Rep_map.png}
%DIFDELCMD < %%%
\DIFdelendFL \DIFaddbeginFL \includegraphics[width=12cm]{Rep_map_new.png}
\DIFaddendFL \centering
\caption{Heat map examplarily illustrating expert agreement on avalanche area for avalanches mapped from SPOT in January 2018 (24 January 2018, SPOT 6 © Airbus DS2018). Agreement in the shade (northern part of the study area) is generally lower than in the sunlit areas to the south.
```

\DIFdelbeginFL \DIFdelFL{The Figure also shows that agreement is }\DIFdelendFL \DIFaddbeginFL \DIFaddFL{Dark blue indicates }\DIFaddendFL very good \DIFdelbeginFL \DIFdelFL{only on few selected avalanches (depicted }\DIFdelendFL \DIFaddbeginFL \DIFaddFL{agreement or }\DIFaddendFL in \DIFdelbeginFL \DIFdelFL{dark blue)}\DIFdelendFL \DIFaddbeginFL \DIFaddFL{other words marks areas that where identified as an avalanche by all five involved experts}\DIFaddendFL . For \DIFdelbeginFL \DIFdelFL{the }\DIFdelendFL \DIFaddbeginFL \DIFaddFL{a }\DIFaddendFL more detailed location of the reproducibility study area see Figure~\ref{fig:TrainCoAreas}. }

\label{fig:compStudy}.

\end{figure}

Calculating F1-score (see \DIFdelbegin \DIFdel{Formula}\DIFdelend \DIFaddbegin \DIFadd{Equation}\DIFaddend ~\ref{equ:F1}), between all experiment mappings, we found an overall F1-score of 0.381 in illuminated and 0.018 in shaded areas \DIFaddbegin \DIFadd{(area-wise metrics)}\DIFaddend . Comparing two expert mappings at a time, the values range from 0.617 to 0.786 in the illuminated regions and from 0.123 to 0.401 in the shaded regions of our study area (Table~\ref{tab:ReproResults}). The F1-scores of the expert manual mappings with the initial mapping are in the same range (not shown). The results from 2018 (Figure~\ref{fig:compStudy}) illustrate that for some selected avalanches the agreement is very good while, especially in the shade, there is little agreement among experts on the presence of avalanches. \\

Reexamining the results from the network now in the light of this experiment, the adapted DeepLabV3+ is equally good as the experts in identifying avalanches. In other words, we cannot expect a computer algorithm to provide better scores than the average F1-score of two mapping experts. Even for the avalanches with the highest agreement, a specific boundary line will usually not match exactly. This makes it hard for any network to learn the localisation of boundaries. We do not yet know what exactly causes the differences in avalanche identification between experts. Therefore we plan on conducting a thorough analysis on imagery with different spatial resolutions in the future. This will help to better understand the inherent mapping uncertainty of avalanches and may give an indication what performance can be expected if training computational detection algorithms on different optical data.

\subsection{Limitations of this study}

The three avalanche periods for which we have SPOT imagery all occurred in January. Those images are relatively close to the winter solstice and therefore have a high percentage of shaded area. \DIFdelbegin \DIFdel{(\mbox{%DIFAUXCMD

\citet{hafner.2021} }\hspace{0pt}%DIFAUXCMD

identified for 180 km^2 around Davos that the share }\DIFdelend \DIFaddbegin \DIFadd{The amount }\DIFaddend of shaded area \DIFdelbegin \DIFdel{at the SPOT 6/7 image acquisition time ranges from }\DIFdelend \DIFaddbegin \DIFadd{depends very much on the terrain and on the season. Around Davos, Switzerland, for example, }\DIFaddend 43\% \DIFaddbegin \DIFadd{of the area is shaded }\DIFaddend at winter solstice \DIFdelbegin \DIFdel{to }\DIFdelend \DIFaddbegin \DIFadd{but only }\DIFaddend 7\% three months later \DIFdelbegin \DIFdel{. Even }\DIFdelend \DIFaddbegin \DIFadd{\mbox{%DIFAUXCMD

\citep[both at SPOT 6/7 image acquisition time;][]{hafner.2021}}\hspace{0pt}%DIFAUXCMD

. We know that the quality of the manually annotated avalanches is lower in shaded areas \mbox{%DIFAUXCMD

\citep[POD: 0.15 shade, 0.86 illuminated, 0.74 overall;][]{hafner.2021}}\hspace{0pt}%DIFAUXCMD

. Consequently, the training data have lower quality in shaded regions, which makes learning there more difficult for our model and leads to lower model confidence as well as poorer results. Based on the results when training and testing on sunlit avalanche parts only, however, we see potential for better overall metrics when a smaller portion of the area is shaded closer to the summer solstice. But regardless how much area is well illuminated, the challenges in the shade remain and make results in those areas less trustworthy. Further research to better understand and tackle that problem is needed.

}

\DIFadd{Additionally, even }\DIFaddend though 2018 includes wet snow and wet snow avalanches, the snow in January is generally colder and drier than towards \DIFdelbegin \DIFdel{spring}\DIFdelend \DIFaddbegin \DIFadd{the end of the winter}\DIFaddend . Consequently, we do not know how well our model performs under different snow \DIFdelbegin \DIFdel{and illumination }\DIFdelend conditions, for example in spring. \DIFdelbegin \DIFdel{Based on the results when training and testing on sunlit avalanche parts only, we see potential for better overall metrics as a smaller portion of the area is shaded. But whether our network }\DIFdelend \DIFaddbegin \DIFadd{Whether our model }\DIFaddend already generalizes enough or is biased towards \DIFdelbegin \DIFdel{January }\DIFdelend \DIFaddbegin \DIFadd{high winter }\DIFaddend conditions and requires retraining with \DIFdelbegin \DIFdel{spring SPOT 6/7 }\DIFdelend \DIFaddbegin \DIFadd{different snow conditions }\DIFaddend we could not yet test.

\conclusions[Conclusion and outlook]

We present a novel deep learning approach for avalanche mapping with deformable convolutions that adapts its notion of the local terrain according to the input digital elevation model (DEM). Experiments at large scale with optical, high spatial resolution (1.5 m) SPOT 6/7 satellite imagery show that our approach achieves good performance (F1-score 0.625) and generalizes well to new scenes not seen during the training phase (F1-score 0.625). As reference data for training, validating and testing our model we relied on 24'747 manually mapped and annotated avalanches from two avalanche periods on different years. With our adapted DeepLabV3+ we were able to detect 66\% of all avalanches. By varying model parameters and the input data we analyzed the impact of different configurations on the mapping result. We found that weighting the avalanches according to the perceived visibility did result in slightly better metrics than when not weighting them. By training on release areas and deposits only we demonstrated that the network learns more from deposits

(\DIFdelbegin \DIFdel{F1-score release areas 0.183; F1-score deposits 0.347}\DIFdelend \DIFaddbegin \DIFadd{Table~\ref{tab:networkResults}}\DIFaddend) and by excluding shaded areas from training we showed that in illuminated terrain both training is easier and test results are better (F1-score 0.639). Furthermore, we investigated expert agreement for manual avalanche mapping in a small reproducibility study and found that agreement on avalanche area is substantially lower than expected. Compared to the model, the agreement between experts is in the same range as the adapted DeepLabV3+ performance.

Our work is an important step towards a fast and comprehensive documentation of avalanche periods from optical satellite imagery. This could substantially complement existing avalanche databases, improving their reliability to perform hazard zoning or the planning of mitigation measures. For the future we aim at conducting a more throughout study investigating expert agreement for manual avalanche identification and its implications for automated avalanche mapping. Additionally, we intend to study the performance of our model on data from different sensors and time periods. Furthermore, we plan on improving results by masking out areas where avalanche cannot occur using for example modelled avalanche hazard indication data from~\citep{buhler.2022}.

%% The following commands are for the statements about the availability of data sets and/or software code corresponding to the manuscript.

%% It is strongly recommended to make use of these sections in case data sets and/or software code have been part of your research the article is based on.

%\codeavailability{TEXT} %% use this section when having only software code available

%\dataavailability{} %% use this section when having only data sets available

\codedataavailability{The manually mapped avalanche outlines from 24 January 2018 and 16 January 2019 used by us for training, testing and validation are available on EnviDat \citep{spot.2018, spot.2019}. The code used will be published and made available on GitHub with the final publication of this paper.} %% use this section when having data sets and software code available

%\sampleavailability{TEXT} %% use this section when having geoscientific samples available

%\videosupplement{TEXT} %% use this section when having video supplements available

%\appendix

%\section{} %% Appendix A

%\subsection{} %% Appendix A1, A2, etc.

%\noappendix %% use this to mark the end of the appendix section. Otherwise the figures might be numbered incorrectly (e.g. 10 instead of 1).

%% Regarding figures and tables in appendices, the following two options are possible depending on your general handling of figures and tables in the manuscript environment:

%% Option 1: If you sorted all figures and tables into the sections of the text, please also sort the appendix figures and appendix tables into the respective appendix sections.

%% They will be correctly named automatically.

%% Option 2: If you put all figures after the reference list, please insert appendix tables and figures after the normal tables and figures.

%% To rename them correctly to A1, A2, etc., please add the following commands in front of them:

\appendixfigures %% needs to be added in front of appendix figures

\appendixtables %% needs to be added in front of appendix tables

%% Please add \clearpage between each table and/or figure. Further guidelines on figures and tables can be found below.

\authorcontribution{EH coordinated the study, performed all initial manual mappings, expanded the neural network code originally implemented by PB, and did the statistical analysis. RD, JW and KS advised on the machine learning aspects of the project and critically reviewed the associated results. RD wrote the script for the Wallis filtering. The reproducibility investigation was initiated by KS, coordinated by EH and both YB and EH were part of its mapping team. EH wrote the manuscript with help from all other authors. EH and YB originally initiated the automation of avalanche mapping from SPOT. } %% this section is mandatory

\competinginterests{The authors declare that they have no conflict of interest.} %% this section is mandatory even if you declare that no competing interests are present

%\disclaimer{TEXT} %% optional section

\begin{acknowledgements}

We thank Leon Bührle, Benjamin Zweifel as well as Andreas Stoffel for mapping for our reproducibility investigation and Frank Techel for valuable input on its analysis\DIFaddbegin \DIFadd{. We are grateful to Ron Simenhois and Edward Bair for the critical questions, suggestions and comments in their reviews}\DIFaddend .

\end{acknowledgements}

%% REFERENCES

%% The reference list is compiled as follows:

%

%% Since the Copernicus LaTeX package includes the BibTeX style file copernicus.bst,

%% authors experienced with BibTeX only have to include the following two lines:

%%

\bibliographystyle{copernicus}

\bibliography{refs.bib}

%%

```
%% URLs and DOIs can be entered in your BibTeX file as:
%%
%% URL = {http://www.xyz.org/~jones/idx_g.htm}
%% DOI = {10.5194/xyz}


%% LITERATURE CITATIONS
%%
%% command          & example result
%% \citet{jones90}|        & Jones et al. (1990)
%% \citep{jones90}|        & (Jones et al., 1990)
%% \citep{jones90,jones93}|    & (Jones et al., 1990, 1993)
%% \citep[p.~32]{jones90}|    & (Jones et al., 1990, p.~32)
%% \citep[e.g.,][]{jones90}|   & (e.g., Jones et al., 1990)
%% \citep[e.g.,][p.~32]{jones90}| & (e.g., Jones et al., 1990, p.~32)
%% \citeauthor{jones90}|     & Jones et al.
%% \citeyear{jones90}|      & 1990


%% FIGURES

%% When figures and tables are placed at the end of the MS (article in one-column style), please add \clearpage
%% between bibliography and first table and/or figure as well as between each table and/or figure.

% The figure files should be labelled correctly with Arabic numerals (e.g. fig01.jpg, fig02.png).


%% ONE-COLUMN FIGURES

%%f
```

```
%\begin{figure}[t]
%\includegraphics[width=8.3cm]{FILE NAME}
%\caption{TEXT}
%\end{figure}
%
%%% TWO-COLUMN FIGURES
%
%%f
%\begin{figure*}[t]
%\includegraphics[width=12cm]{FILE NAME}
%\caption{TEXT}
%\end{figure*}
%
%
%%% TABLES
%%%
%%% The different columns must be seperated with a & command and should
%%% end with \\ to identify the column brake.
%
%%% ONE-COLUMN TABLE
%
%%t
%\begin{table}[t]
%\caption{TEXT}
%\begin{tabular}{column = lcr}
%\tophline
%
%\middlehline
%
%\bottomhline
%\end{tabular}
%\belowtable{} % Table Footnotes
```

```
%\end{table}
%
%%% TWO-COLUMN TABLE
%
%%t
%\begin{table*}[t]
%\caption{TEXT}
%\begin{tabular}{column = lcr}
%\tophline
%
%\middlehline
%
%\bottomhline
%\end{tabular}
%\belowtable{} % Table Footnotes
%\end{table*}
%
%%% LANDSCAPE TABLE
%
%%t
%\begin{sidewaystable*}[t]
%\caption{TEXT}
%\begin{tabular}{column = lcr}
%\tophline
%
%\middlehline
%
%\bottomhline
%\end{tabular}
%\belowtable{} % Table Footnotes
%\end{sidewaystable*}
%
```

```
%
%%% MATHEMATICAL EXPRESSIONS
%
%%% All papers typeset by Copernicus Publications follow the math typesetting regulations
%%% given by the IUPAC Green Book (IUPAC: Quantities, Units and Symbols in Physical Chemistry,
%%% 2nd Edn., Blackwell Science, available at: http://old.iupac.org/publications/books/gbook/green_book_2ed.pdf, 1993).
%%%
%%% Physical quantities/variables are typeset in italic font (t for time, T for Temperature)
%%% Indices which are not defined are typeset in italic font (x, y, z, a, b, c)
%%% Items/objects which are defined are typeset in roman font (Car A, Car B)
%%% Descriptions/specifications which are defined by itself are typeset in roman font (abs, rel, ref, tot, net, ice)
%%% Abbreviations from 2 letters are typeset in roman font (RH, LAI)
%%% Vectors are identified in bold italic font using \vec{x}
%%% Matrices are identified in bold roman font
%%% Multiplication signs are typeset using the LaTeX commands \times (for vector products, grids, and exponential notations) or \cdot
%%% The character * should not be applied as mutliplication sign
%
%
%%% EQUATIONS
%
%%% Single-row equation
%
%\begin{equation}
%
%\end{equation}
%
%%% Multiline equation
%
%\begin{align}
%& 3 + 5 = 8\\
%& 3 + 5 = 8\\
%& 3 + 5 = 8
%\end{align}
%
%
%%% MATRICES
%
%\begin{matrix}
%x & y & z\\
%x & y & z\\
%x & y & z\\
%\end{matrix}
%
%
%%% ALGORITHM
%
%\begin{algorithm}
%\caption{...}
%\label{a1}
%\begin{algorithmic}
%...
%\end{algorithmic}
%\end{algorithm}
%
%
%%% CHEMICAL FORMULAS AND REACTIONS
%
%%% For formulas embedded in the text, please use \chem{}
%
%%% The reaction environment creates labels including the letter R, i.e. (R1), (R2), etc.
%
%\begin{reaction}
```

```
%%% \rightarrow should be used for normal (one-way) chemical reactions

%%% \rightleftharpoons should be used for equilibria

%%% \leftrightarrow should be used for resonance structures

%\end{reaction}

%

%

%%% PHYSICAL UNITS

%%%

%%% Please use \unit{} and apply the exponential notation



\end{document}

% https://www.overleaf.com/project/61b988a7273c3657af457429
```