University of St Andrews

School of Mathematics and Statistics

On the Generalizability of Iterative Patch Selection for High-Resolution Image Classification

---

I hereby certify that this dissertation, which is approximately 9,684 words in length, has been composed by me, that it is the record of work carried out by me, and that it has not been submitted in any previous application for a degree. This project was conducted by me at the University of St Andrews from 05/2024 to 08/2024 towards fulfillment of the requirements of the University of St Andrews for the degree of MSc Statistics under the supervision of Dr. Chrissy Fell.

Max Thibault Riffi-Aslett

Abstract

Classifying large images with small or tiny regions of interest (ROI) in computer vision is challenging due to computational and memory constraints. Efforts to optimize memory consumption have achieved comparable results to strongly supervised methods by integrating custom patch selectors. Nonetheless, challenges occur in scenarios with low signal-to-noise ratios and low entropy attention, which have been associated with overfitting. We seek to contribute to exploring these shortcomings by leveraging a novel testbed on a previously introduced memory-efficient cross-attention-based transformer. Our testbed extends the megapixel MNIST benchmark to different O2I (Object-to-image) ratios while keeping the canvas size fixed and introducing a novel noise generation component. We present extensive experimental results indicating that the O2I threshold below which patch-based classifiers fail to generalize is affected by both the task of the megapixel MNIST benchmark ("Maj", "Top", "Max", "Multi") and the training dataset size. Contributing to these shortcomings, we perform 142 model runs with 41 configurations across 2 benchmark datasets, the megapixel MNIST, and Swedish traffic signs datasets. We first experiment with two previously introduced regularization strategies, namely Stochastic Top-k Instance Masking (STKIM) and Semantic and Diversity regularization, and show that the latter offers clear improvements for the Swedish traffic signs dataset on a 25% subset of the original data. Secondly, we find that the patch size relative to the average region of interest has a distinct impact on generalization whereby the classifier benefits from having a patch size equal to or lower than the smallest region of interest. Additional experiments on the noise generation's effect on convergence, show that the model fails to converge for low O2I ratios as the distribution of the noise component gets closer to the region of interest. Our code is made available at https://github.com/MRiffiAslett/ips_MaxRiffiAslett.git.

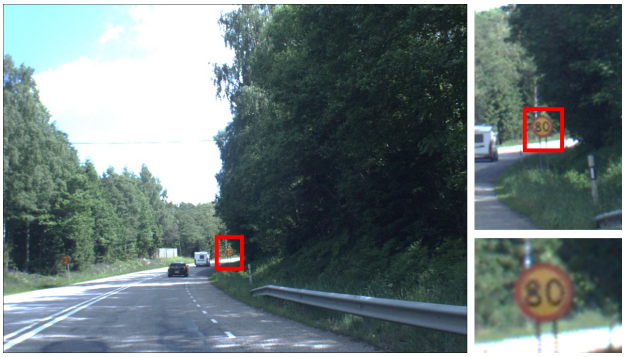


Figure 1: Visual representation of image "000036" from the Swedish traffic signs dataset with label "80". On the left, the original image in full resolution 960×1280. On the right, the image is incrementally magnified to approximately 75×45 (top) and 35×15 (bottom).

1 Introduction

Advancements in Convolutional Neural Networks (CNN) such as AlexNet **(Krizhevsky, Sutskever, and Hinton 2012)** and ResNet **(He et al. 2015)** have been successful in classifying natural images with resolutions below one megapixel on datasets such as ImageNet **(Deng et al. 2009)**. Yet, in multiple practical applications such as aerial imagery **(Ofli et al. 2016)**, traffic monitoring **(LaLonde, D. Zhang, and Shah 2018)**, and automatic industrial inspection **(Abouelela et al. 2005)**, the label correlates with only a small part of the input, leading to a low signal-to-noise ratio (refer to Figure 1 for a visual representation of the Swedish traffic signs dataset). This is especially true for computational methods for pathological diagnosis, which analyze whole-slide images (WSIs) of cancerous tissue, to provide computer-assisted diagnosis and prognosis **(Sun, Meng, and Liu 2022)**. In the CAMELYON 17 challenge, for instance, a standard whole-slide image is 200,000 × 100,000 pixels with a 3-byte RGB pixel format resulting in 600GB of data for 899 WSIs.

A common solution is to use strongly supervised learning which utilizes local region-level annotations from domain experts **(Sun, Meng, and Liu 2022)**. In digital pathology, however, region-specific information is expensive to collect while labels for whole-slide images are frequently captured **(Gadermayr and Tschuchnig 2024)**. Weakly supervised methods on the other hand can be applied to gigapixel images without fully annotated data, only including slide level labels. A common weakly supervised learning approach is to use attention-based Multiple Instance Learning (MIL). This is achieved by aggregating patches via an attention-based weighted average (pooling operator) thus providing a way to quantify the importance of each patch (region) in the image (refer to section 2.1 for a breakdown of MIL). These methods, however, process all tissue patches at high magnification, which means a large number of uninformative patches are aggregated, increasing memory usage. This makes it impractical to scale these methods to larger histopathology images, like Prostatec-

tomy which can use slides measuring 300,000×400,000 pixels at 40× magnification **(Thandiackal et al. 2022)**.

To promote the clinical deployment of these methods, a disjoint but related line of work that also falls under the weakly supervised paradigm, focuses on the fact that it is unnecessary to process the whole input image as relevant information is often unevenly distributed. Some methods use one or more lower-resolution versions of the input image to select salient patches. **Katharopoulos** and **Fleuret** 2019 and **Kong** and **Henao** 2021 for instance sample patches based on their attention values at low magnification to be processed at higher resolutions. The attention values are differentiable, with gradients of the loss computed using the Monte Carlo approximation. Alternatively, **Cordonnier** et al. 2021 create a discrete ranking of the most informative patches to select and aggregate the top K most salient **(Cordonnier et al. 2021)**. As rankings are not differentiable they add Gaussian noise to each ranking (refer to section 2.6). In some datasets, however, the informative regions are not visible at lower magnification. This is the case for the Needle MNIST dataset where the goal is to detect the presence of the digit "3" on a large canvas containing cluttered digits. At lower magnifications, the digits are not discernible from each other.

The method implemented and experimented upon in our work overcomes this limitation. Iterative Patch Sections (IPS), **(Bergner, Lippert, and Mahendran 2023)**, processes full resolution patches in batches and retains the top M most salient after each iteration. The most informative patches are then aggregated by a cross-attention-based pooling operator reminiscent of Multiple Instance Learning. The cross attention pooling operator, follows the original transformer setup of **Vaswani** et al. 2023, using a network architecture based only on learning a representation of each patch weighted by their attention values. Patches are initially scored with the same transformer as the pooling operator running in no gradient mode. This allows them to retain only the top M most salient after each iteration, reducing memory consumption (Refer to section 2.7 for a full breakdown of IPS). IPS has achieved state-of-the-art results on the megapixel MNIST, Swedish traffic signs, and CAMELYON16 dataset while boasting lower memory consumption compared to its predecessors **(Cordonnier et al. 2021; Katharopoulos and Fleuret 2019)**. For these reasons, we decide to center our experiments around IPS to build upon a state-of-the-art memory efficient patch-based classifier.

While memory-efficient classifiers have achieved comparable results to weakly and strongly supervised state-of-the-art methods, they struggle with generalizability in low signal-to-noise scenarios and tend to over-concentrate on a small subset of informative patches, which is intrinsically linked to overfitting. We contribute to exploring these limitations by extending the megapixel MNIST dataset to different O2I (object-to-image) ratios and introducing a novel Bézier curve-based noise generation strategy. Our results show that performance worsens in low data scenarios as the O2I ratio drops. To explore this shortcoming, we perform a series of experiments to evaluate previously introduced regularization and pertaining strategies and study the effects of patch size and noise on generalization at low O2I ratios. Specifically, we alter the pipeline of IPS in the following ways:

1. We extend the megapixel MNIST to various object-to-image ratios with a novel noise-generation strategy.
2. Implement Stochastic Top-K Instance Masking **(Y. Zhang et al. 2024)** aimed at masking out salient regions of the top K most informative regions to avoid the dominance of a subset of informative patches in the attention distribution.
3. We alter the loss to include Semantic and Diversity regularisation to penalize the cosine similarity between attention heads of the Multiple Head Attention (MHA) aggregation module **(Y. Zhang et al. 2024)** (Refer to section 4.5 for a full breakdown).
4. We extend the current encoding strategy to larger residual networks (ResNet-50) by treating it as a patch encoder where the weights are frozen.
5. We evaluate the effects of patch size on generalization at low O2I ratios.
6. We study the effects of noise that gradually gets closer to the distribution of the region of interest (ROI) on convergence.

2 Literature Review

High-dimensional images are too demanding for existing CNN architectures due to their computational and memory constraints. A typical solution overcomes this challenge by splitting the images into patches to frame the task as a strongly supervised learning problem; however, these methods require ground truth annotations **(Nazeri, Aminpour, and Ebrahimi 2018)**. Furthermore, patch-level annotations assign a definitive label to each patch, even if some regions are ambiguous which can provide further context during inference **(Anand et al. 2021)**. In the following, we address weakly supervised methods that operate without region-level annotations, aiming to cover memory-efficient patch-based image classifiers and the methods upon which they are built.

2.1 Multiple Instance Learning

A common solution to the weakly supervised problem in high-dimensional image classification is Multiple Instance Learning (MIL). It involves assigning labels to collections of unordered instances, called bags. The unordered instances in this case are square regions of the image called patches that are assigned the same label. Patches are then processed separately by a feature extractor that usually consists of a CNN. Their outputs are then aggregated by a pooling function consisting of any differentiable function **(Gadermayr and Tschuchnig 2024)**. The pooling operator aggregates instance embeddings into a bag-level representation which can then be used directly or processed further to obtain a bag-level label. Attention-based pooling, where instance embeddings are weighted by their informativeness and then averaged has been shown to be superior to traditional max and mean pooling operations **(Ilse, Tomczak, and Welling 2018)**.

More related to our work on IPS, transformer-based MIL methods were introduced to capture the correlations between different instances. Correlated MIL **(Shao et al. 2021)**, employs self-attention **(Vaswani et al. 2023)**, to aggregate instance-level features into a bag-level representation for classifying breast cancer WSI. Self-attention learns similarity scores between each pair of patches which is then used to weight a learnable bag level representation **(Vaswani et al. 2023)**. Alternatively, Attention-based Deep Multiple Instance Learning **(Ilse, Tomczak, and Welling 2018)** employs a Gated Attention (GA) mechanism as the weighting scheme for their attention pooling operator for grading breast cancer and colon cancer whole-slide images. While traditional attention mechanisms attend to the entire sequence of patches, GA attends to only a subset of the sequence via a network that generates binary gates to mask informative patches **(Xue, Li, and N. L. Zhang 2019)**.

2.2 Self supervised learning

The successes of self-supervised learning (SSL) in computer vision have prompted the development of SSL methods for high-dimensional images. The objective of self-supervised pre-training is to learn a transferable representation of data without manual supervision. This is achieved by training an in-domain feature extractor beforehand that leverages contrastive loss. Contrastive loss works by minimizing the distance between positive samples containing augmented versions of the same patches while maximizing the distance between negative samples. By using Momentum Contrast v2 (MoCo v2), a previously introduced self-supervised learning algorithm, **Dehaene** et al. 2020 showed that using self-supervision for training the feature extractor improves the performance of Multiple Instance Learning for grading breast cancer whole-slide images. Because histopathology images often contain patches with similar tissue composition, **Wang** et al. 2022 suggests modifying the traditional contrastive learning approach with a cosine similarity metric to encourage similar pairs of tiles to be close in the embedding space.

2.3 Recurrent visual attention models

The aforementioned weakly supervised methods are not intended to lower the computation and memory footprint that is inherent in all patch-based models. Recurrent visual attention models on the other hand improve computational efficiency by only processing some parts of the full image. The research within this framework sequentially processes patches, focusing on selecting specific regions for further processing. **Mnih** et al. 2014 first used a recurrent neural network to identify regions of interest in high-resolution images using reinforcement learning to train their model as it is non-differentiable. Meanwhile, **Ranzato** 2014 proposed using a downsampled input image to provide additional spatial context to the RNN. These works however solve an optimization

problem that is non-differentiable which makes them difficult to train **(Katharopoulos and Fleuret 2019)**.

2.4 Gradient checkpointing

Another line of work intended to lower the computation and memory footprint uses gradient checkpointing, which manages memory by only storing specific activations. The method deletes all activations after the forward pass, except for a few in checkpoint layers. In the backward pass, intermediate activations are recomputed with respect to the checkpoints to reduce memory **(Marra et al. 2019)**. StreamingCNN **(Pinckaers, Van Ginneken, and Litjens 2022)** uses gradient checkpointing to process images as large as 8k×8k pixels sequentially through a CNN. **Hou** et al. 2016 segments the input image into n overlapping tiles. Each tile is then processed through the network's initial layers up to a certain checkpoint. Gradients are then recomputed separately for each tile during backpropagation. These approaches, however, fail to make use of a common feature amongst gigapixel images, where only a small region of the canvas correlates with the label.

2.5 Attention Sampling

Katharopoulos and **Fleuret** 2019 introduced an attention network to sample important areas in a downsampled view of the original image to process a fraction of the original image. The sampled patches are then aggregated by computing the expectation of the patches over their attention distribution. This attention-based sampling method is fully differentiable, with gradients of the loss with respect to the attention parameters computed using the Monte Carlo approximation to estimate the expected gradients. **Kong** and **Henao** 2021 extend **Katharopoulos** and **Fleuret** 2019's work to further lower memory requirements by splitting the attention-based sampling process into two stages, starting with the lowest resolution and then progressing to the higher resolution. Moreover, they incorporate a contrastive learning objective into their Zoom-In network by reversing the label of sub-tiles with low attention weights ($y = 1 \rightarrow 0$) to create negative contrasting samples.

One significant limitation of downlsampling the image occurs when there is no discernible discriminate information at a lower scale. **Kong** and **Henao** 2021 for instance assessed their Zoom-In network's performance on the Needle MNIST dataset **(Pawlowski et al. 2020)**, which consists of randomly positioned digits on an image canvas and the task is to detect the presence of the digit "3" (Refer to Section 3.1.2). Their network failed to handle this dataset as down-sampling the image washes out discriminative information, which prevents the attention network from finding regions of interest. Interestingly, **Kong** and **Henao** 2021 experimented by incorporating pixel-level annotations from the CAMELYON16 dataset which resulted in better model performance.

2.6 Differentiable Top-K

Cordonnier et al. 2021 also builds upon the work of **(Katharopoulos and Fleuret 2019)** by introducing a differentiable Top-K operator to select the most relevant patches in the input. Their approach leverages a shallow scorer network that operates on a downsampled image to assign a relevance score to each patch used to select the K most relevant patches in the image. As discrete rankings are not differentiable, they utilize the perturbed maximum approach from **Blondel** et al. 2020, which incorporates Gaussian noise to each rank to make them differentiable. An aggregation network, which can accommodate various pooling methods, then combines this information into the final model output.

Thandiackal et al. 2022 builds on top of **Cordonnier** et al. 2021 by implementing multi-level zooming and adapting the technique to gigapixel-sized whole slide images (WSIs), treating it as a Multiple Instance Learning problem. Attention scores are computed for each patch at every magnification level, and the patches with the highest scores are chosen for additional processing at a higher magnification. For feature aggregation, they implement an attention-pooling operation, particularly the Gated Attention (GA) mechanism introduced by **Ilse**, **Tomczak**, and **Welling** 2018 (refer to Section 2.1).

2.7 Iterative patch selection

The method implemented and experimented upon in our work is closely related to Multiple Instance Learning. **Bergner**, **Lippert**, and **Mahendran** 2023's Iterative Patch Selection iterates through each patch in the image to only maintain the top M most informative in memory. The image is split into patches and fed through IPS in batches which loads a fixed number of patches I at a time. At each iteration, patches are encoded first with a CNN and then with a cross attention module running in no gradient mode to extract attention values. These attention values in turn allow IPS to only keep the top M most informative patches after each iteration. Importantly the cross attention module is shared with the patch aggregation module (refer to Equation 3). Once the top M patches of dimension D are selected, $X^* \in \mathbb{R}^{M \times D}$ they are aggregated by a weighted average as shown in Equation 1.

$$z = \sum_{m=1}^{M} a_m (X_m^* W^v) \tag{1}$$

where a_m is the attention score for the m-th patch and $X_m^* W^v$ is a linear projection of the patch embeddings X^* and $\mathbf{W}^v \in \mathbb{R}^{D \times D_v}$ is a learnable weight matrix , where D_v is the dimension of the value embeddings. In transformer notation, $X_m^* W^v$ corresponds to the values (V). Each attention score a_m is the result of a multiple head cross attention layer which follows the original setup of transformers **(Vaswani et al. 2023)**. The attention for each patch is given by Equations 2, 3 and 4:

$$a_{1,\ldots,M} = \text{Softmax}(f_\theta(X_1^*), ..., f_\theta(X_M^*)) \tag{2}$$

$$f_\theta(X_m^*) = \frac{QK^T}{\sqrt{D_k}} \tag{3}$$

$$Q = q_\phi W^q, \quad K = X_m^* W^k \tag{4}$$

Here, Q (queries) and K (keys) are learned linear transformations of the patch embeddings, where q_ϕ is a learnable query token, and W^q, W^k are learnable weight matrices of dimensions $\mathbb{R}^{D \times D_k}$ (D_k is the dimension of the Key embeddings and D the dimensions of the patch embeddings). The learnable query token q_ϕ allows the model to focus on different aspects of the patch embedding. The Queries (Q) and Keys (K) learn a similarity score between each pair of patch embeddings which is used to weight the values (V) which contain a learned projection of the patch embeddings. This results in learned bag level representation. A single weighted average z, corresponds to a single head. To enable the model to learn different attention maps, 8 attention heads are used $(z_1, z_2, \ldots, z_8)$. All heads are then concatenated and down-projected (refer to Equation 5).

$$z = \text{concat}(z_1, z_2, \ldots, z_H) W^o \tag{5}$$

where z_h is the output of head h and W^o is a learnable weight matrix used to down project the concatenated outputs. For handling multiple tasks, the transformer uses multiple sets of heads, each specific to a task. Finally, these global representations are passed through classification heads that consist of a multi-layer perceptron (MLP) (1 for each task). The transformer module therefore includes a multi-head cross-attention (MCA) layer, a multi-layer perceptron (MLP), and a layer normalization layer (LN) (refer to Equation 8) where $X_{\text{pos}} \in \mathbb{R}^{M \times D}$ is an optional positional encoding:

$$z = \text{MCA}(X^* + X_{\text{pos}}) \tag{6}$$

$$z' = \text{LN}(z + q_\phi) \tag{7}$$

$$z_o = \text{LN}(\text{MLP}(z')) \tag{8}$$

Importantly, at each iteration of IPS, patches are scored by the same multi-head cross-attention module in no gradient mode. The corresponding attention scores are then used to select to top M most salient patches at each iteration before being fed through the same module again, this time in gradient mode to achieve a bag-level representation as in equation 1. A key advantage that IPS holds over Attention Sampling **(Kong and Henao 2021; Katharopoulos and Fleuret 2019)** and Differentiable Top-K **(Cordonnier et al. 2021)** is that it does not rely on lower resolution views of the original image to detect salient patches. It is therefore advantageous in that it benefits from mostly better efficiency while bypassing down sampling the canvas which can blur discriminate information for very low O2I ratios.

2.8 Object-to-image (O2I) ratios

While Attention Sampling **(Katharopoulos and Fleuret 2019)**, Differentiable Top-K **(Cordonnier et al. 2021)**, and Iterative Patch Selection **(Bergner, Lippert, and Mahendran 2023)** have reached performance levels comparable to fully supervised techniques, difficulties are attributed to scenarios with low object-to-image (O2I) ratios. For instance, **Kong** and **Henao** 2021 found that their Zoom-In network had poor performance on the Needle MNIST dataset as information is lost when the image is down-sampled. Additionally, **Thandiackal** et al. 2022 found that their Top-K module tends to overlook extremely small metastases, resulting in the misclassification of WSIs due to low attention. Furthermore, **Bergner**, **Lippert**, and **Mahendran** 2023 found that performance can decline in IPS, if the signal-to-noise ratio decreases, even when using full-resolution images in the patch selection module. This is demonstrated in their experiments with the megapixel MNIST datasets where the image size was scaled from 1k to 10k pixels. Their results showed that IPS performs with high accuracy up to 8k pixels, and begins to decrease in accuracy from 9k pixels. Importantly, these difficulties are attributed to either the patch selector failing to discern discriminate information or failing to assign attention values to very small but informative regions.

Robustness to small object-to-image ratio was explored previously by **Pawlowski** et al. 2020, who found that convolutional neural networks (CNNs) have poor performance on very low object-to-image ratio problems with limited dataset size. These findings highlight the value of developing data-efficient solutions for gigapixel image classification. Our work seeks to follow theirs by reproducing the intuition behind their experiments on memory-efficient patch-based classifiers and experimenting with different avenues to contribute to robustifying IPS in scenarios with limited data and low object-to-image ratios.

2.9 Overlapping patches

To find a solution to the challenges outlined in the previous section, we turned our attention to a disjoint but related challenge that is amplified in low object-to-image ratios. This is the trade-off of searching for new informative patches within an image while utilizing patches already identified as informative. The smaller the object of interest, the more attention maps will tend to over-concentrate on small regions of the input image **(Y. Zhang et al. 2024)**. We recognize that this issue is common across multiple domains in Machine Learning however for simplicity we only treat it as it relates to patch-based image analysis. In **Katharopoulos** and **Fleuret** 2019, an entropy regularizer is used in the attention distribution to penalize low entropy and prevent the attention network from over-concentration on regions that directly optimize the loss. Additionally, **Cordonnier** et al. 2021 reported that there differentiable Top-K operator tended to select patches that directly optimize the training loss resulting in overlapping patches around the region of interest, causing overfitting.

This problem was addressed by **Y. Zhang** et al. 2024 who made similar observations during the application of Multiple Instance Learning (MIL) to whole slide images, where they found that overfitting can be caused by the attention module over-concentrating on a small set of discriminate instances. To address this, they propose a multibranch cross-attention mechanism similar to Multiple-Head Attention, but with the key difference being that MBA includes Diversity regularization to ensure different branches learn different patterns. Additionally, they present a new masking technique known as Stochastic Top-K Instance Masking. We show that these regularization methods apply to IPS's multiple-head cross-attention pooling operator in sections 3.3 and 4.6 where we experiment with both in an attempt to address the shortcomings of IPS in scenarios with low object-to-image ratios.

3 Methodology

3.1 Object-to-image ratio

3.1.1 Background

CNNs have been tested under different noise conditions either by experimenting with artificially introduced label noise **(C. Zhang et al. 2017)**, or by using datasets with noisy labels **(Mahajan et al. 2018)**. Large amounts of label noise have been shown to negatively impact generalization **(C. Zhang et al. 2017)**. Further research efforts however found that CNNs can counteract this label-corrupting noise by increasing the training dataset size **(Mahajan et al. 2018)** and fine-tuning the optimization hyperparameters such as the learning rate and batch size **(Jastrzebski et al. 2018)**.

3.1.2 Needle MNIST

Unlike the aforementioned works, **Pawlowski** et al. 2020 address an adverse scenario with noiseless labels and low object-to-image ratio by introducing the Needle MNIST dataset, inspired by the Cluttered MNIST dataset **(Ba, Mnih, and Kavukcuoglu 2015)**. The task is to predict whether the digit "3" appears on the canvas amongst cluttered digits with labels {0, 1, 2, 4, 5, 6, 7, 8, 9} sampled with replacement (Refer to Figure 2, (left)). The object-to-image ratio is then changed by maintaining the digits at 28 × 28 pixels and increasing the canvas resolution. This results in O2I ratios of {19.1, 4.8, 1.2, 0.3, and 0.075}%, with canvas sizes of 64×64, 128×128, 256×256, 512×512, and 1024×1024 pixels. Their findings show that CNNs fail to generalize below a certain signal-to-noise ratio, and the dataset size influences this ratio.

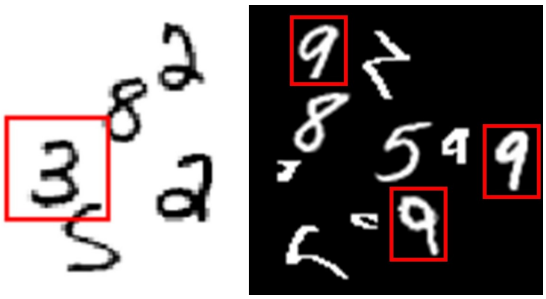


Figure 2: Visual representation of the Needle MNIST dataset (left) and the megapixel MNIST dataset (right) with five noise digits using Bézier curves.

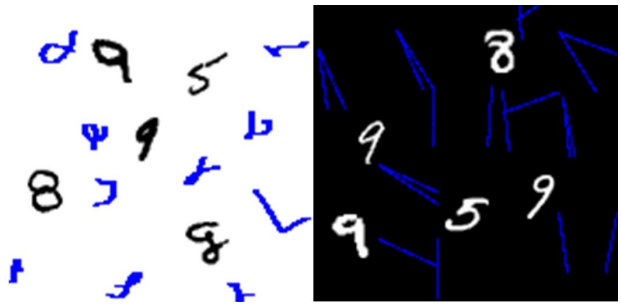


Figure 3: Visual representation of the megapixel MNIST dataset (150 × 150) with 10 noise digits from our method (left) and the original megapixel MNIST (right).

3.1.3 Megapixel MNIST

While **Pawlowski** et al. 2020's Needle MNIST tasks can be solved with just one informative patch, megapixel MNIST requires the recognition of multiple patches and their relationships. The megapixel MNIST dataset introduced by **Katharopoulos** and **Fleuret** 2019 features 5 MNIST digits placed randomly on a canvas. Of these, 3 digits belong to the same class, while the remaining 2 are from different classes. The task is to detect the majority class (refer to Figure 2, (right)). **Bergner**, **Lippert**, and **Mahendran** 2023 found that this problem is well-solved by most baselines. To introduce more complexity, they extended the setup with three additional tasks: detecting the maximum digit, identifying the topmost digit, and recognizing the presence or absence of all classes.

3.1.4 O2I experimental setup

This first part of our work seeks to extend the experiments of **Bergner**, **Lippert**, and **Mahendran** 2023 in two ways. 1: by varying both the O2I ratio and the size of the training data, unlike **Bergner**, **Lippert**, and **Mahendran** 2023, who only changed the O2I ratio, and 2: by changing the noise component of the megapixel MNIST dataset to mimic a more adverse setting. Importantly, our setup keeps the canvas size fixed and scales the images upward, unlike the approaches in **Bergner**, **Lippert**, and **Mahendran** 2023 and **Pawlowski** et al. 2020, which scaled the image size.

We scale the original megapixel MNIST from 1500 by 1500 to 3000 by 3000 pixels and increase the size of the digits. The digit sizes are as follows: 28×28, 56×56, 84×84, and 112×112 pixels, which correspond to the following O2I ratios: {0.01, 0.034, 0.078, 0.13}%. The aforementioned O2I ratios were selected to obtain the largest array of very small object-to-image ratios while fitting within our memory constraints. Additionally, we linearly decrease the noise as a function of the digit sizes as follows: noise = digit size $\times(-7.14)+1000$, resulting in 800 noise digits for the lowest O2I ratio and 200 for the highest. Each set of O2I ratios is run 4 times with different amounts of training data, specifically {4000, 2000, 1000, 800} samples, chosen as 4000 samples solves the task while 800 is the limit below which the model fails to converge.

3.1.5 Noise generation

Here we present the noise generation strategy. In the original megapixel MNIST dataset **(Bergner, Lippert, and Mahendran 2023)**, 50 line patterns are created by sampling angles θ_i from a uniform distribution. The slopes m_i are calculated, and line coordinates (x_j, y_j) are generated based on these slopes, as shown on the right of Figure 3. Given that DNNs tend to prioritize capturing simpler patterns to minimize training loss, we instead seek to create nonlinear curves that mimic the structure of digits using Bézier curves (refer to Figure 3, on the right).

Bézier curves are parametric curves used to model a smooth surface which are defined by the relative positions of a set number of control points **(Baydas and Karakas 2019)**. They have been used extensively in computer-aided geometric design for instance where a popular area of research seeks to model surfaces with Bézier curves using a shape parameter **(Qin et al. 2013)**. We suggest randomly sampling control points from a set with predefined probabilities that match the control point counts observed in the digits 1 through 9 as illustrated in Table 1.

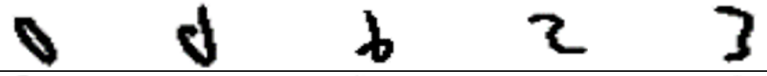
Control Points	P	Noise Digits				
4	**0.33**					
6	**0.56**					
8	**0.11**					

Table 1: Illustration of noise components using Bézier curves with $\{4, 6, 8\}$ control points, sampled with probabilities $\left\{\frac{3}{9}, \frac{5}{9}, \frac{1}{9}\right\}$. The two last columns on the right present rare noise components chosen for their close resemblance to digits 1, 2, 4, 5, and 6.

We begin by sampling control points from a uniform distribution on an $N \times N$ canvas. For each curve, we sample p control points, where p is randomly chosen from the set $\{4, 6, 8\}$ with probabilities $\left\{\frac{3}{9}, \frac{5}{9}, \frac{1}{9}\right\}$. We analyzed the number of control points required for each digit to ensure that the sampled number of control points reflects the observed values as close a possible. For instance, our observations indicated that digit "0" required 4 control points, digit "2" required 6, and digit "8" required 8 control points. By aligning the sampled control points with these values, we can generate curved lines with a distribution of control points inspired by

what is observed in digits. The Bézier curve is defined in Equation 9.

$$\mathbf{B}(t) = \sum_{i=0}^{n} \binom{n}{i} (1-t)^{n-i} t^i \mathbf{P}_i, \quad t \in [0,1] \tag{9}$$

where $\mathbf{P}_i$ are the control points and n is the number of control points minus one. We discretize t into 100 points and draw the curve on an $N \times N$ grid (refer to Algorithm 1).

Algorithm 1 Proposed noise generation module using Bézier Curves with a custom sampling scheme to create noise components closer to the distribution of original digits.

Input: Canvas size $N \times N$, set of control point probabilities $\left\{\frac{3}{9}, \frac{5}{9}, \frac{1}{9}\right\}$, number of points $p \in \{4, 6, 8\}$
Output: Bézier curves on an $N \times N$ grid
Initialize an empty canvas of size $N \times N$
for each curve **do**
    Randomly choose p control points from $\{4, 6, 8\}$ with probabilities $\left\{\frac{3}{9}, \frac{5}{9}, \frac{1}{9}\right\}$
    Sample control point positions $\mathbf{P}_i$ from a uniform distribution on the $N \times N$ canvas
    The Bézier curve is defined by the control points:

$$\mathbf{B}(t) = \sum_{i=0}^{n} \binom{n}{i} (1-t)^{n-i} t^i \mathbf{P}_i, \quad t \in [0,1]$$

    Discretize t into 100 points $\mathbf{B}(t_j)$
end for

3.2 Swedish traffic signs

In the interest of running experiments across domains, we use the Swedish traffic signs dataset, where the task is to classify traffic signs in road images. The original dataset **(Larsson and Felsberg 2011)** contains 3,777 images with dimensions 960×1280 and 20 different traffic sign classes (refer to Figure 1). We employ the setup used in previous works **(Katharopoulos and Fleuret 2019)**, where the task is limited to detecting speed limits of 80, 70, and 50. This results in using only a subset of the original data, including 747 training and 684 validation instances. The subset includes 100 images for each speed limit class and 400 road images that do not contain a speed limit sign. After reproducing the results of IPS, we found that the dataset can be solved easily. To be able to stress test the following regularisation methods, we train on a subset of the original data (25%) stratified by the categories: *no sign* (50%), *80* (16.6%), *70* (16.6%), *50* (16.6%). Comprehensive experiments evaluating the effects of downsizing the dataset size are presented in Section 4 as well as the rationale for the down-scaling.

3.3 Diversity regularisation

In scenarios with limited data and low object-to-image (O2I) ratios, we observed that the generalizability of the model suffers, which aligns with findings from **Pawlowski** et al. 2020 on CNNs. To address this, we experiment with the regularization methods proposed by **(Y. Zhang et al. 2024)**, as these methods are also applicable to IPS **(Bergner, Lippert, and Mahendran 2023)**. As noted by **Y. Zhang** et al. 2024, the key difference between Multiple-Head Attention (MHA) and their Multiple Branch Attention (MBA) is the implementation of Diversity and Semantic loss. Diversity loss penalizes the cosine similarity of each head's attention map while Semantic loss attaches an MLP to each head to make an individual prediction. Both are intended to work in tandem where each head is constrained to learn diverse yet effective patterns. The Diversity loss, as presented by **Y. Zhang** et al. 2024, is defined as:

$$L_d = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} \cos(a_i, a_j) \tag{10}$$

where M is the number of attention heads, and a_i and a_j are the attention maps of the i-th and j-th heads.

Following the setup from **Y. Zhang** et al. 2024, we implement Semantic regularization by feeding each head through a distinct MLP to make a prediction and adding the cross-entropy loss to ensure that each head learns discriminative patterns. We extend IPS by placing a Multi-Layer Perceptron (MLP) before the heads are concatenated and linearly transformed as in Equation 11:

$$z = \text{concat}(z_1, z_2, \ldots, z_H)W^o \tag{11}$$

where each head z_i is processed through the following MLP (refer to Equation 12):

$$\text{MLP}_h(z_i) = W_{h3}(\sigma(W_{h2}(\sigma(W_{h1}z_i + b_{h1})) + b_{h2})) + b_{h3} \tag{12}$$

where W_{h1}, W_{h2}, W_{h3} are learnable weight matrices, b_{h1}, b_{h2}, b_{h3} are the biases, and σ is the activation function. After concatenation, W^o (Equation 11) projects the combined heads to the same dimensionality as an individual head. We can therefore conveniently use the current MLP implementation used after the heads are concatenated (refer to Equation 7) and place one after each head to make an individual prediction. Specifically, this MLP consists of two layers each with 128 neurons corresponding to the patch embedding size D, with ReLU activation and a dropout rate of 0.1. The total loss function combines the bag classifier loss L_b, the Semantic loss L_p, and the Diversity loss L_d as demonstrated in Equation 13.

$$L = L_b + L_p + L_d \tag{13}$$

3.4 Attention masking

In the realm of natural image classifications, masking salient regions or activations has proven effective in enhancing generalization. Cutout **(DeVries and Taylor 2017)** for instance, randomly masks out square regions of the input during training, and RSC **(Huang et al. 2020)**, discards dominant features and activates remaining features that correlate with the label.

In the context of patch-based high-resolution image classification, MHIM-MIL **(Tang et al. 2023)** masks easier instances using attention scores, pushing the model to concentrate on more challenging instances by using a teacher-student framework. Another method, named Hard Positive Instance Mining (HPM) **(Qu et al. 2022)**, forces the teacher network to focus on more challenging positive instances by training initially both teacher and student networks and then using the student classifier to exclude instances with high prediction probabilities.

Y. Zhang et al. 2024 on the other hand experimented with Stochastic Top-K Instance Masking (STKIM). Specifically, their strategy, randomly sets the attention values of the Top-K instances to 0, with a probability of P, where P and K are two hyperparameters that control the intensity of masking. Unlike methods like WENO **(Qu et al. 2022)** and MHIM-MIL **(Tang et al. 2023)**, which require pre-trained teacher models and mask a larger number of instances, STKIM doesn't rely on a teacher-student framework or pre-training process. As the method provides no additional memory constraints it is applicable to our scope, we implement STKIM to IPS, experimenting with multiple configurations of P and K.

4 Results

For simplicity, our results are broken down into 7 questions that are work seeks to answer:

1. Is IPS **(Y. Zhang et al. 2024)**, robust to small object-to-image ratios? Does the nature of the classification task in the megapixel MNIST affect the O2I threshold below which IPS fails to generalize?
2. Following the work and observations made by **Pawlowski** et al. 2020, does the dataset size influence the minimum object-to-image (O2I) ratio required for the patch-based classifier to generalize?
3. To what extent can adapting the patch size to be smaller/equal/larger than the Region of Interest (ROI) be useful in a limited data scenario with a low O2I ratio?
4. To improve generalizability, can vanilla residual models in the patch encoding stage be leveraged, either for fine-tuning with pre-initialized weights or with frozen weights, effectively using them as feature extractors?
5. Can we improve generalization by encouraging the multiple heads of the cross-attention aggregation module to learn disjoint but effective patterns through our implementation of Semantic and Diversity loss?
6. Can a more tailored masking strategy that promotes the search for new instances while using existing salient ones be useful in the case or our implementation of STKIM **(Y. Zhang et al. 2024)**?
7. Lastly, how does the noise generation component affect performance for low O2I ratios? Is convergence affected by the resemblance of the noise to the region of interest (ROI)?

In **Pawlowski** et al. 2020's O2I experiments, each model was run 6 times over random seeds and then averaged to control for random effects that appear in the parameter initialization phase. Due to this work's, time and hardware constraints, however, not all experiments are run across random seeds. The number of runs will be made evident throughout. For the megapixel MNIST dataset, each model configuration was trained with the following hyperparameters unless specified otherwise: 100 epochs, a batch size of 16, a memory size of 100, an iteration size of 100, a patch size of 50×50, patch stride of 50 and a ResNet-18 encoder.

In experiments with the Swedish traffic signs dataset, each configuration was run 5 times across random seeds for 150 epochs with a memory and iteration size of 10 and 32 patches respectively. The patch size and batch size remain fixed at 100×100 and 16 respectively. Additionally, a ResNet-18 pre-trained on IMAGENET1K V1 in gradient mode is used as the patch encoder. Here we formulate the reason for using a subset of the original size of the Traffic sign dataset. We reproduced the results of IPS for 3 subset sizes across 5 model runs, namely {25%, 50%, 100%} resulting in the following training set sizes {744, 372, 184}. Each instance is randomly sampled without replacement from the original set stratified by label: *no sign* (50%), *80* (16.6%), *70* (16.6%), *50* (16.6%). Results in Table 2 show that with 25% of the data, the validation accuracy sits at 80% as opposed to 95% and 98% for 50% and 100% of the training data. Additionally, the standard deviation for 25% of the data sits at 5.18% indicating that the model converged to a similar accuracy for each run. We therefore choose 25% as it struggles with some harder instances in the validation set while consistently converging giving us space to experiment with the regularisation methods and saving computation resources.

Subset	Validation			Training			Time (s)/ Batch
	n-val	Mean (%)	Std (%)	n-train	Mean (%)	Std (%)	
25%	169	79.745	5.18	184	99.88	0.24	3.25s
50%	341	95.2228	3.31	372	99.946	0.10	7.15
100%	684	98.3126	0.24	747	99.9172	0.06	14.9

Table 2: Mean and standard deviation of the 18 model runs with different subsets of the Swedish traffic signs data stratified to the following proportions: *no sign* (50%), *80* (16.6%), *70* (16.6%), *50* (16.6%).

For the cross-attention transformer, we followed the hyperparameters of **Y. Zhang** et al. 2024, who in turn used the default values in **Vaswani** et al. 2023. The optimization strategy is also unchanged, with a warm-up period of 10 epochs where the learning rate decreases linearly. When fine-tuning pre-trained networks, the learning rate is adjusted to 0.0003, whereas it is set to 0.001 for networks trained from scratch. Throughout the training process, a cosine schedule is then used to gradually decrease the learning rate by a factor of 1,000.

4.1 Object-to-image ratio

Changing the object-to-image ratio on the megapixel MNIST dataset (refer to Figure 4) indicates some interactive relationships between the O2I threshold for generalization, the dataset size, and the task of the megapixel MNIST dataset. Notably, for the task of identifying the majority digit ("Maj") and the top-most digit ("Top"), the object-to-image ratio greatly affects the amount of data needed to generalize. For the task Majority for instance (Top left), at the lowest O2I ratio (0.01%), 2000 samples are needed to achieve a validation accuracy of 88% while at the highest O2I ratio (0.14%) 1000 samples yield a validation accuracy of 89%. The same relationship between the dataset size and O2I ratio is also observable for the task "Top" (bottom left). The tasks "Max" and "Multi" however showcase that the task affects the rate by which the object-to-image ratio affects the number of instances needed to generalize. This rate is much greater for tasks "Maj" and "Top" than it is for tasks "Max" (top right) and "Multi" (bottom left). For the task "Max" for instance, at the lowest O2I ratio (0.01%), 800 training samples achieve a validation accuracy of 49% while at the highest O2I ratio (0,14%), 800 training instances yield a validation accuracy of 64%. This effect is less present as the number of training instances increases. With 2000 training instances, for example, the validation accuracy of the task "Max" is 83% for an O2I of 0.01% and 79% for an O2I of 0.14%. For the task "Multi" the rate decreases further. For each number of training instances, no trend is apparent between the O2I ratio and validation accuracy. For instance, at 2000 samples the validation accuracy is $\{79\%, 71\%, 87\%, 83\%\}$ for the following O2I ratios $\{0.01\%, 0.03\%, 0.08\%, 0.14\%\}$. **Pawlowski** et al. 2020 found that the number of training instances affects the O2I threshold for generalization. We contribute to their work by finding that this effect is observable in the megapixel MNIST dataset on a patch-based classifier and that there are interactions between the task of the megapixel MNIST and the rate by which the size of the training data affects the O2I threshold for generalization.

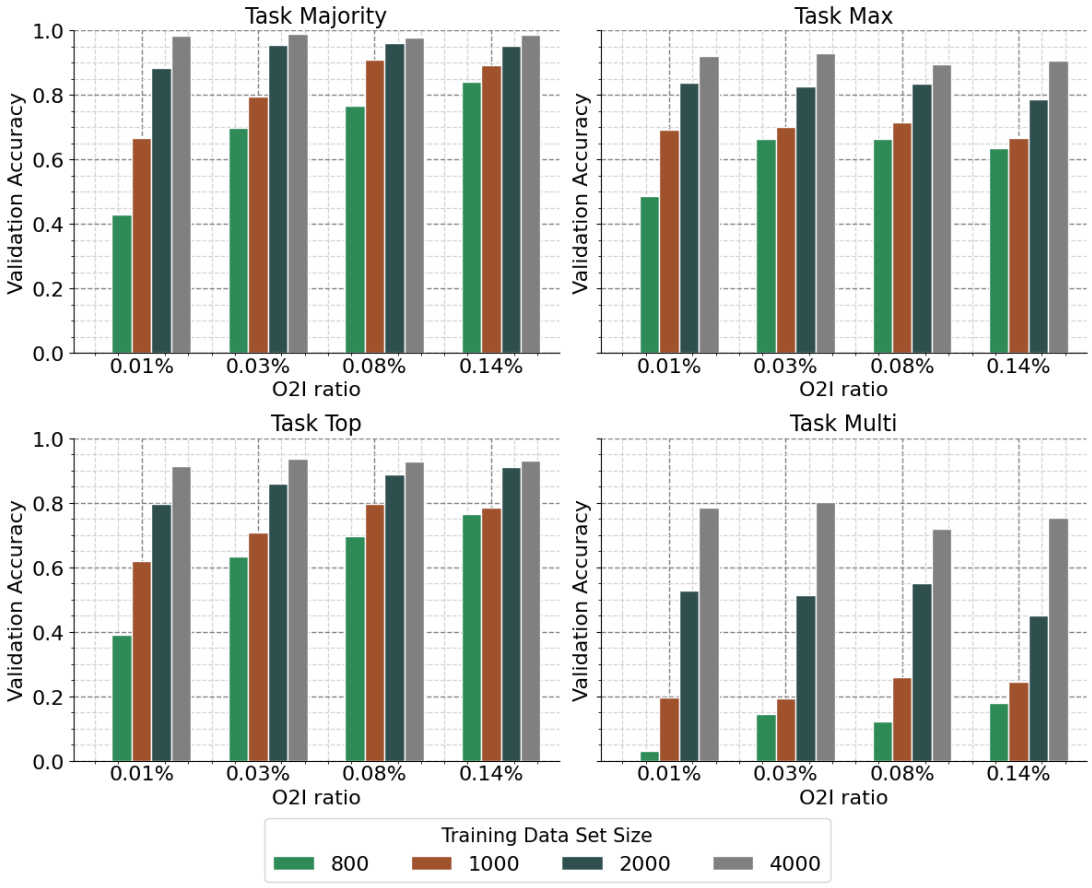


Figure 4: Results of the experiments on megapixel MNIST with a novel noise generation component. Four object-to-image ratios were tested: across four training dataset sizes. Canvas size and patch size remain fixed at 3000×3000 and 50×50, respectively, and the O2I changes by varying the digit resolutions to 28×28, 56×56, 84×84, and 112×112.

4.2 Effect of dataset size

In the previous section, we found that the megapixel MNIST task affects how the dataset size influences the object-to-image threshold for generalization. Here we focus on the lowest object-to-image ratio (0.01%) and show that tasks "Maj", "Max" and "Top" generalize at the same rate with increasing dataset sizes compared to task "Multi" which requires much more data to generalize. Figure 5 illustrates an experiment where we incrementally train IPS with a larger array of training dataset sizes, namely $\{200, 300, 600, 800, 1000, 2000, 3000, 4000, 5000\}$ while maintaining the object-to-image ratio at it's lowest (0.01%). The results clearly illustrate that the task of recognizing the presence or absence of all classes ("Multi") requires significantly more data (3000 samples) to achieve a validation accuracy of 86% compared to the tasks, "Maj", "Max" and "Top" which require 2000 samples to achieve a validation accuracy of 89%, 84%, and 80% respectively. This is due to the nature of the task. The task "Multi" has significantly more possible combinations (10 possibilities for each of the 5 digits) compared to the other tasks which have 10 possible labels. This showcases that in a low data setting (2000 training samples) IPS can comfortably make predictions on the canvas as a whole by identifying, the maximum, majority, and topmost digit but falls short in detecting each digit individually on the canvas.

4.3 Effect of patch size

To improve IPS's ability to generalize in low data settings with tiny object-to-image ratios, we start by experimenting with varying the patch size to be higher or lower than the regions of interest (ROI). The main finding of this section is that by tuning the patch size in low data settings we can achieve higher validation accuracy compared to the original patch sizes used by **Bergner**, **Lippert**, and **Mahendran** 2023 to solve the full megapixel MNIST and Swedish traffic signs dataset. For the megapixel MNIST dataset, a canvas size of 1500×1500 was used with a fixed digit size of 84×84. The patch sizes were changed to the following values: $\{25 \times 25, 50 \times 50, 100 \times 100, 150 \times 150\}$ and the validation accuracy averaged over three random seeds. The range of patches were chosen to be larger and smaller than the size of the region of interest (ROI). Figure 6 illustrates that the accuracy for all tasks is higher when the patch size is smaller than the ROI for the megapixel MNIST. When the patch size sits at 50, the validation accuracy for tasks "Maj", "Max", and "Top" are 76%, 72%, 69%, while with a patch size of 150, the accuracy falls to 43%, 47%, and 40% respectively. Interestingly, the highest validation accuracy for all tasks is observed when the patch size is set to 25×25, except for the task "Multi" where a patch size of 50×50 yields the highest testing accuracy (25%).

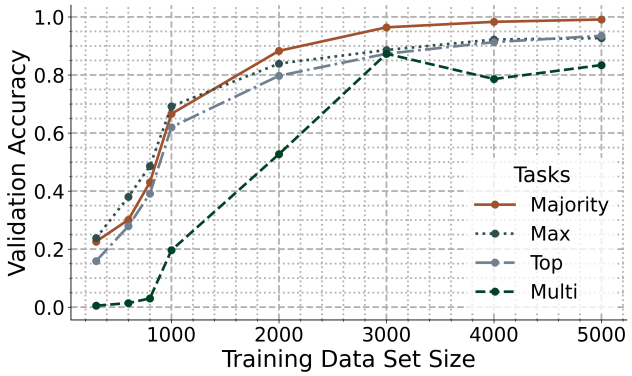


Figure 5: Validation accuracy for different training data size. Canvas size and digit size remain fixed at 3000×3000 and 28×28 respectively resulting in an object-to-image ratio of 0.01%. Model trained for 100 epochs with 800 noise digits.

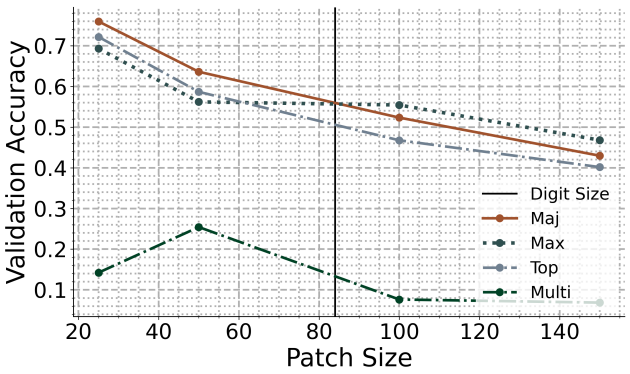


Figure 6: Validation accuracy for different patch sizes. Each iteration in this table was run for 100 epochs with an object-to-image ratio of 0.078% using 1000 training samples on a 1500×1500 canvas. The digit size is set at 84×84.

In the megapixel MNIST implementation of IPS with 5000 training instances, the patch size was set to 50×50 and the digit size to 28×28, resulting in a 31% object-to-patch ratio (O2P). This is equivalent to our setup when the object size is 84×84 and the patch size is 150×150 as the object-to-patch ratio is $84^2/150^2 = 31\%$. However, an O2P ratio of 110% corresponding to a 25×25 patch size (refer to Figure 6), is preferred over 31%. The validation accuracy for an O2P ratio of 110% (Patch size; 25×25) is 76%, 75%, 69% and 15% for tasks "Maj", "Top", "Max" and "Multi" and falls to 43%, 40%, 48% and 0% respectively with an O2P ratio of 31% (Patch size; 150×150). An alternative way to set up the experiment could have been to keep the digit size the same as in IPS, however, this would have limited our ability to scale the patch size well below the original object size of 28×28.

We extend the experiment to the Swedish traffic signs dataset by evaluating the following patch sizes: $\{25 \times 25, 50 \times 50, 75 \times 75, 100 \times 100\}$, chosen as the range contains the generalization maxima. Each patch size was run for 5 random seeds, and the mean and standard deviation of the validation accuracy were reported in Table 3. Generalization is highest when the patch size is 75 (85%) (refer to Table 3) and at its lowest when it is 25 (71%) and 50 (78%). For reference, the original setup of IPS set the patch size to 100×100, yet within a low datasetting with 25% of the original size of the dataset, a patch size of 75 offers clear improvements in validation accuracy with a marginal decrease in standard deviation. Specifically, for a patch size of 100×100, the average validation accuracy stands at 79% (Std: 3%), while it climbs to 84% for a patch size of 75×75 (Std: 2.8%).

Patch Size	Validation		Training		Time (ms)/Batch	Peak memory (GB)
	Mean (%)	Std (%)	Mean (%)	Std (%)		
25	71.12	12.74	99.02	0.63	449.72	0.97 GB
50	77.75	4.77	100.00	0.00	302.31	1.11 GB
75	84.62	2.83	100.00	0.00	299.31	1.33 GB
100	79.29	3.07	100.00	0.00	322.67	1.65 GB

Table 3: Mean and standard deviation of the validation and training accuracy for the Swedish traffic signs data for different patch sizes. Each model configuration was run 5 times for 150 epochs on a 25% subset of the data. Average time per epoch was taken by averaging the run time of each 16 batches in epoch 0.

An additional benefit is the lower memory and runtime requirements of a smaller patch size. This is expected as the number of patches in memory M remains fixed in IPS. When the patch size decreases, fewer pixels are kept in memory. A patch size of 100×100 consumes 1.6 GB of peak memory usage with an average time per batch of 323 ms. A patch size of 75×75 on the other hand consumes only 1.3 GB of peak memory usage with an average time per batch at a lower 299 ms. The implications for this are clear, in low data scenarios we can improve validation accuracy for low object-to-image ratio problems by tuning the patch size. By using the same object-to-image (O2I), and object-to-patch (O2P) ratio we were able to get substantial gains in validation accuracy compared to the origin setup in IPS **(Bergner, Lippert, and Mahendran 2023)**.

4.4 Effect of pre-training

In this section, we extend the setup of IPS by changing the encoding strategies. We employ both ResNet-50 and ResNet-18 architectures with three sets of pre-trained weights. Residual networks were introced by **He** et al. 2015, to allow the training of deeper neural networks. Particularly, they employ skip connections to avoid the vanishing gradient problem. Skip connections bypass the nth layer by applying ReLU to the sum of the convolution of the $(n-1)$th layer and the $(n+1)$th layer. ResNet-18 is composed of "BasicBlocks," with two convolutional layers, whereas ResNet-50 employs "BottleneckBlocks," which include three convolutional layers aimed at reducing the dimensionality of the weights and restoring them.

For ResNet-18, the encoder maintains the original implementation of IPS **(Bergner, Lippert, and Mahendran 2023)**, where the weights are initialized with pre-trained values and are fully trainable. In our implementation of the ResNet-50 backbone, the encoder includes layers up to the last global average pooling layer, but all parameters are frozen apart from the last fully connected layer effectively using it as a feature extractor. The original output dimension of 2048 is further down-projected to 512 through a linear projection layer to fit our hardware constraints. We used both the pre-trained ResNet-18 and ResNet-50

weights from PyTorch's `torchvision.models` base class implementation. Additionally, we experimented with `ResNet50.IMAGENET1K_V2`, which improves upon the results of `ResNet50.IMAGENET1K_V1` by using a novel training recipe that includes additional features such as Label Smoothing **(Szegedy et al. 2015)** to prevent the model from becoming too overconfident, and improved weight decay and learning rate configurations.

For our experiments with the megapixel MNIST, Each configuration was run for 50 epochs for the lowest object-to-image ratios of 0.01% and 0.03% on 1000 training samples. The number of epochs was shortened in this case by 50% as the runtime was greater than our resource limit and the results are equally as revealing. Result for the megapixel MNIST in Table 4 show that there is no improvement in validation accuracy when pre-trained ResNet weights are leveraged in the encoding phase for the megapixel MNIST dataset. In the original implementation of IPS, the weights are randomly initialised on a ResNet-18 backbone (first two rows of Table 4). At the lowest object-to-image ratio, (0.01%), ResNet-18 with random initialization yielded a higher validation accuracy (Majority: 46%) than when initializing with pre-trained weights on IMAGENET V1 (Majority: 27%) or with a ResNet-50 patch encoder using either IMAGENET V1 (15%) or V2 (22%). The superiority of the initial setup in IPS is further demonstrated at a higher object-to-image ratio of 0.03%. In this case the original setup of IPS yielded a validation performance of 75% for the task "Maj" while it falls to 17% with pre-initialised ResNet-18 weights and to 30% (V1) and 17% (V2) with Frozen ResNet-50 weights.

Backbone	Version	Regime	O2I (%)	Majority (%)	Max (%)	Top (%)	Multi (%)
ResNet-18	-	Random init	0.01%	46.4	61.1	47.0	3.9
ResNet-18	-	Random init	0.03%	75.2	78.3	76.4	8.5
ResNet-18	V1	Trainable	0.01%	27.4	30.4	25.7	1.1
ResNet-18	V1	Trainable	0.03%	17.2	29.7	16.8	0.0
ResNet-50	V1	Frozen	0.01%	15.9	24.1	13.7	0.1
ResNet-50	V1	Frozen	0.03%	30.8	36.3	21.8	0.3
ResNet-50	V2	Frozen	0.01%	22.2	25.4	25.4	0,7
ResNet-50	V2	Frozen	0.03%	17.0	23.7	14.1	0.3

Table 4: Validation accuracy for different encoding strategies on the megapixel MNIST dataset. Each iteration in this table was run for 50 epochs with 1000 training samples.

In the case of the Swedish traffic signs dataset, we employed the same encoding strategies including a pre-initialised ResNet-18 (original setup is IPS) as well as ResNet-50 with frozen weights. Each configuration was run 5 times across random seeds for 150 epochs on a 25% subset of the data stratified by label. Results in Table 5 reveal that the original setup in IPS again results in superior validation performance as well as substantial runtime improvements. By using a ResNet-18 backbone (row 1 of table 5) the original configuration of IPS yields a validation performance of 80% while it falls to 61% (V1) and 65%(V2) for our ResNet-50 setup. Importantly the average time per batch is 345 milliseconds for ResNet-18 and 1112 milliseconds (V1) for ResNet-50. The peak memory usage is also lower for ResNet-18 at 1.6 GB compared to ResNet-50 (V1) at 1.9 GB.

Backbone	Version	Regime	Mean (%)	Std (%)	Time(ms) / batch	Peak Memory
ResNet-30	V1	Trainable	79.74	5.17	349.0929	1.6463 GB
ResNet-50	V1	Frozen	61.41	2.69	1110.5184	1.9469 GB
ResNet-50	V2	Frozen	65.47	1.11	1119.48	1.9469 GB

Table 5: Validation accuracy for different encoding setups on the Swedish traffic signs dataset. Average time per epoch was taken by averaging the run time of each 16 batches in epoch 0.

The drop in performance is the result of freezing ResNet-50's weights. Further experiments could have been conducted by unfreezing some layers; however, the memory and runtime requirements of such scaling would fall outside of our scope, which is to robustly memory-efficient patch-based methods trainable on a single GPU. These findings imply that larger residual networks offer no improvement in the patch encoder of IPS when frozen. It is important to remark however that ResNet-50 has been successfully implemented in training mode as the feature extractor for several patch-based classifiers **(Shao et al. 2021; Cordonnier et al. 2021)**.

4.5 Diversity regularisation

Here we present the results of the Semantic and Diversity loss implementation inspired by **Y. Zhang** et al. 2024. Our findings indicate that the Diversity loss can offer some improvement which is not generalizable across domains. For the megapixel MNIST, Each model was run for both the lowest O2I ratios of 0.01% and 0.034%, corresponding to the digit sizes 28×28 and 56×56. To explore different weights for the Diversity loss, we multiplied it by 1, 2, and 3 for the megapixel MNIST. We find that the Diversity loss offers no improvement across all 4 tasks, 2 object-to-image ratios, and weights on megapixel MNIST.

Table 6 shows that for the lowest object-to-image ratio (0.01%), the validation performance with Diversity set to 0 for task "Max" is 66%. As Diversity increases to 1× Div loss, 2× Div loss, and 3× Div loss, the performance decreases to 39%, 39%, and 38%, respectively. We posit that the constraint of the cosine similarity in the loss function is restricting the model from generalizing. This trend is further illustrated in Figure 12 in Section 9.4, where the dynamic validation accuracy across the 100 epochs shows that different Diversity loss weighting schemes slow down convergence.

The validation accuracy for the O2I ratio of 0.03% gets progressively smaller as the Diversity loss increases. Specifically with no Diversity, the accuracy of the task "Maj" is 79% and then falls to 71%, 39%, 8% for 0× Div loss, 1× Div loss 2× Div loss and 3× Div loss respectively. We argue that penalizing the cosine similarity between attention heads results in a more challenging optimization problem for the megapixel MNIST dataset, as it assumes a fixed number of attention maps with different yet equally discriminating information.

Weight	Digit Size	O2I	Diversity	Maj (%)	Max (%)	Top (%)	Multi (%)
0xD	"28x28"	0.01%	0.000	66.6	69.2	62.0	19.6
0xD	"56x56"	0.03%	0.000	79.4	70.2	70.9	19.4
1xD	"28x28"	0.01%	0.140	39.6	54.4	35.1	0.4
1xD	"56x56"	0.03%	0.140	71.2	69.2	74.2	0.59
2xD	"28x28"	0.01%	0.254	39.4	58.0	44.3	0.08
2xD	"56x56"	0.03%	0.276	39.9	60.4	59.1	0.46
3xD	"28x28"	0.01%	0.426	37.6	39.4	30.4	0.0
3xD	"56x56"	0.03%	0.429	8.59	27.4	10.1	0.0

Table 6: Validation accuracy for the megapixel MNIST dataset with varying Diversity loss weights. This range was selected as no improvement was observed beyond each boundary. Each iteration was run for digits of size 28×28 and 56×56, resulting in the following object-to-image ratios: {0.01%, 0.034%}.

Our experiments with the Swedish traffic signs dataset on the overhand suggest that the Semantic and Diversity loss additions offer some improvement in a low data setting. We experimented with 6 Diversity loss weightings, namely 0× Div loss, 0.25× Div loss, 0.5× Div loss, 0.75× Div loss, and 0.1× Div loss. The range was chosen as performance drops beyond both boundaries. Each configuration was run for 150 epochs, across 5 random seeds, and the validation accuracy reported in Table 7. When the Diversity loss is set to 0× Div loss (original setup in IPS **(Bergner, Lippert, and Mahendran 2023)**) the mean validation accuracy is 80% and it increases to 84% for 0.1× Div and 82% for 0.25× Div. This improved performance however is at the expense of significantly higher variability with a standard deviation of 0.05 without Diversity and 0.11 with 0.1× Div and 0.25× Div.

Diversity Loss weighting	0xD	0.1xD	0.25xD	0.5xD	0.75xD	1xD
Mean	79.75%	83.68%	82.94%	74.90%	58.58%	68.78%
Std	5.18%	11.15%	10.98%	17.54%	19.48%	17.81%

Table 7: Mean and standard deviation of the validation accuracy for the Swedish traffic signs data for different Diversity loss weighting schemes. Each model configuration was run 5 times for 150 epochs on a 25% subset of the data.

The implications of these results are nuanced. Although some improvement is evident when tuning the weight of the Diversity loss, this came at the cost of much higher variability and was not beneficial across domains. We believe that the Diversity loss likely performs well for some in-distribution scenarios (as was the case with

the Swedish traffic signs dataset) but would struggle with out-of-distribution scenarios dissimilar to the training and validation sets. This is due to the assumptions of Diversity loss which constrain the attention maps of the 8 attention head to be equally as far from one another.

4.6 Stochastic Top-K Instance Masking

Here we detail the results of our implementation of Stochastic Top-K Instance Masking inspired by **Y. Zhang** et al. 2024. Given the two hyperparameters, K and P, where K controls the number of instances to mask and P the proportion of instances to mask in the subset K, we split our experiments into two parts for the megapixel MNIST dataset. 1) We experiment with changing the value of P while keeping K fixed at 10, where the canvas size is 1500×1500 with a digit resolution of 28×28 and 50 noise digits (O2I: 0.0035%). 2) We keep P fixed at a conservative 0.2 and increase the value of K, where the canvas size stays fixed at 3000×3000 with a digit resolution of 28×28 (O2I: 0.01%). The difference in canvas size and noise was unintentional. No correction was made as the O2I ratios are reported and later findings indicate that the number of noise digits on the canvas does not affect performance (refer to section 9.3). The results are thus equally as valid.

Table 8 (top) shows that masking for different values of P while K remains fixed at 10 shows no improvement on megapixel MNIST. Specifically, for the task "Max", the validation accuracy with no masking was 94% and fell to 71%, 61%, 48%, 37%, and 27% for the values of $P = \{0.1, 0.2, 0.4, 0.6, 0.8\}$. For the experiments where P is fixed at 0.2 and K varies (refer to Table 8 (bottom)), we found no improvement. Without masking, the validation accuracy is 80% for task "Max" and falls to 49%, 47%, and 45% for the values of $K = \{20, 30, 40\}$.

dataset size	O2I (%)	Noise	K	P	Maj (%)	Max (%)	Top (%)	Multi (%)
5000	0.03	50	-	-	99.1	93.7	92.7	81.9
			10	0.1	74.5	71.5	62.9	33.5
			10	0.2	64.4	61.0	54.6	15.8
			10	0.4	50.2	48.8	39.5	2.3
			10	0.6	31.1	37.0	28.1	0.4
			10	0.8	17.7	27.4	14.6	0.0
dataset size	O2I (%)	Noise	K	P	Maj (%)	Max (%)	Top (%)	Multi (%)
1000	0.01	800	-	-	88.0	80.7	75.0	45.1
			20	0.2	57.2	49.2	45.0	5.6
			30	0.2	54.4	47.2	42.0	4.4
			40	0.2	34.4	51.7	45.0	6.6

Table 8: Validation accuracy for Stochastic Top-K Instance Masking on megapixel MNIST. On the top, IPS was trained for 150 epochs on a canvas size of 1500×1500 with a digit size of 28×28 with 5000 training samples (O2I: 0.0035%). On the bottom, IPS was trained on a 3000×3000 canvas with a digit size of 28×28 with 1000 training samples (O2I: 0.01%).

In our experiments with STKIM on the Swedish traffic signs dataset, we run 6 model configurations with 3 values of K $\{10, 20, 30\}$, each run for 2 values of P $\{0.1, 0.05\}$ across 5 random seeds. The results are then averaged and the mean and standard deviation are reported in Table 9. Results show that there is no improvement in performance across the parameter space covered in this experiment. Without STKIM, performance after 150 epochs on 25% of the data is 80% however it falls to 57%, 57%, and 53% respectively when P is fixed at 0.05 and K varies to {10, 20, 30}. Similarly, when P is fixed at 0.1 and K varies to {10, 20, 30}, performance drops to {48%, 45%, 44%}.

One caveat is that this setup fails to consider some other interactions between parameters K and P; however, we deem the parameter space is large enough to warrant no further searching. Furthermore, we present additional experimental results in Section 9.5, where Figure 13 displays the shape of the attention distributions for each masking configuration in Table 8. It illustrates that when 2 to 5 instances are informative, as is the case with both the megapixel MNIST and Swedish traffic signs datasets, masking an instance results in a drastic change in

attention. This is due to the lack of ambiguous patches that would allow the model to learn an alternative means to solve the task with STKIM (refer to section 9.6.1 for attention maps of the megapixel MNIST and Swedish traffic signs dataset which showcases that no more than 4 patches occupy most of the attention distribution).

{k & P}	{0 & 0}	{10 & 0.1}	{10 & 0.05}	{20 & 0.1}	{20 & 0.05}	{30 & 0.1}	{30 & 0.05}
Mean	79.74%	47.81%	57.04%	0.45.32%	57.04%	43.66%	53.60%
Std	5.17%	8.70%	11.13%	3.16%	8.61%	11.81%	10.06%

Table 9: Mean validation accuracy across 5 random seeds for Stochastic Top K Instance Masking on a 25% subset of the Swedish traffic signs dataset. Each configuration was run for 150 epochs following the setup of IPS.

4.7 Preliminary experiments

In this section, we justify our decision to scale the canvas size from 1500×1500 to 3000×3000 for the object-to-image ratio experiments on the megapixel MNIST benchmark. As part of our preliminary experiments, we scaled the digit resolution both downward and upward from the original resolution of 28×28 with the following resolutions $\{14 \times 14, 21 \times 21, 28 \times 28, 46 \times 46, 64 \times 64\}$ with a canvas size of 1500×1500 (refer to Figure 7 for a visual representation of the down-sampled digits). We ran these configurations for 150 epochs, resulting in the results presented in Figure 8. We can see that there is a stark decrease in performance when the digit size is down-sampled below its original size $\{28 \times 28\}$.

When the digit size is set to 14×14 (O2I: 0.01%), the validation performance for the task "Maj" sits at 66% for 5000 samples and 150 epochs. On the other hand, the same O2I of 0.01% without scaling the digit size down, with a 28×28 digit on a 3000×3000 canvas, needs 1000 training samples to achieve a validation score of 66% (refer to Section 4). Unsurprisingly, when the image is downsampled below its original size, the loss of information requires more data to achieve comparable validation performance. This observation motivates our choice to scale the image to 3000×3000 and ensure that we only scale the digits and noise upwards.

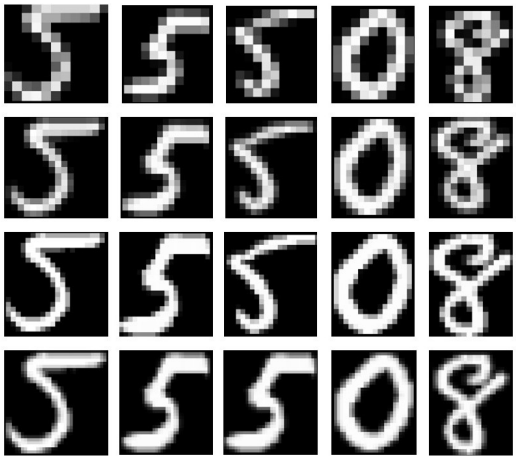


Figure 7: Five MNIST digits (left to right) from the megapixel MNIST dataset, downsampled to the following resolutions in order from top to bottom: $\{14 \times 14, 21 \times 21, 28 \times 28, 46 \times 46, 64 \times 64\}$.

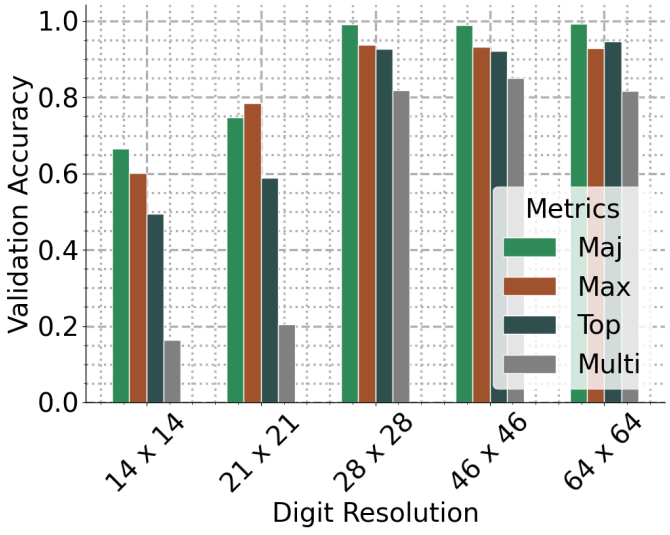


Figure 8: Validation accuracy using lower resolution digits to change the object-to-image (O2I) ratio. The model was run for 150 epochs on a 1500×1500 canvas with 5000 training samples and the following object-to-image ratios: $\{0.01\%, 0.0196\%, 0.34\%, 0.094\%, 0.18\%\}$.

4.8 Effect of noise on convergence

Here we present side findings on the failure of IPs to converge in low O2I ratios when the noise relates more closely to the region of interest. In **Y. Zhang** et al. 2024's experiments, they find that some model configurations fail to converge on both the training and validation sets on the Needle MNIST, with training accuracy close to random. To further understand this phenomenon, they replace the MNIST digits with Gaussian noise, following the setup in **C. Zhang** et al. 2017. It was found that although their setup could memorize the Gaussian noise, it failed to recognize the digit "3" from other digits on the canvas, and convergence became increasingly difficult as the object-to-image ratio decreased. They hypothesize that structured noise, such as digits, may be harder for CNNs to process than Gaussian isotropic noise.

Contributing to these observations, we find that as the thickness of the noise digits approaches the thickness of the digits in the megapixel MNIST dataset, the model fails to converge on both validation and training sets. We empirically demonstrate this by incrementally changing the thickness of the noise digits to $\{1.4, 1.6, 1.8, 2, 2.2, 2.4\}$ (refer to Figure 10 for a visual comparison between the thickness of original MNIST digits and the thicknesses of the noise in the experiment). Each setup was run for 50 epochs across 3 random seeds on a 1500×1500 canvas with a 28×28 digit size (O2I ratio: 0.03%) and noise size of 800. Figure 9 illustrates that the validation accuracy for all tasks decreases in a non-linear fashion from thickness 1.4 to 1.6 before reaching random accuracy at digit thickness 2. Specifically, this threshold shows random accuracy for each task. Tasks "Maj" and "Top" don't converge with a random accuracy of 10% as 10 digits could be chosen. The Task multi doesn't converge at 0% as there are 10 possibilities for each of the 5 digits which is $10^5 = 100,000$ total combinations resulting in a 0% random accuracy. In Section 9.2 we show that the nature of the task "Max" is such that by consistently choosing the digit 9, the random accuracy falls to 30%.

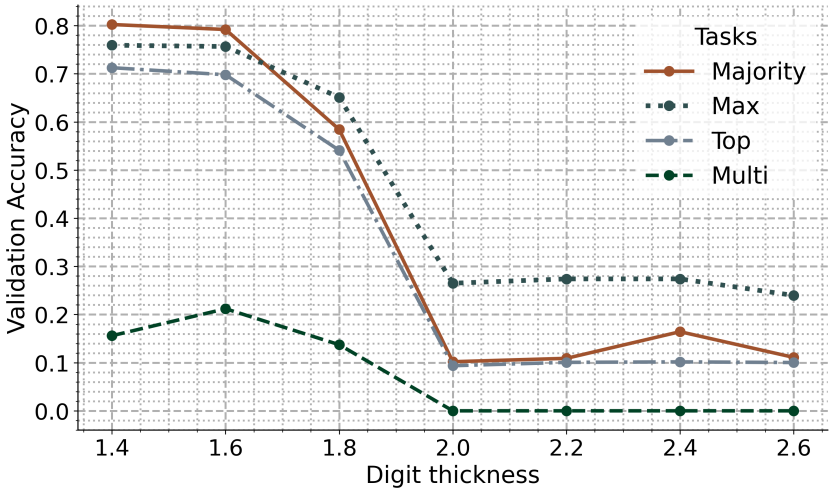


Figure 9: Validation accuracy average over three random seeds varying the thickness of the noise digit used in the megapixel MNIST dataset. Experiments were run on 2000 training samples for 50 epochs and the lowest object-to-image ratio of 0.01%.

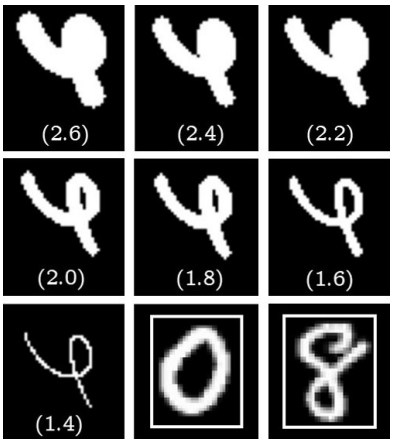


Figure 10: Visualisation of 7 noise digits with the thickness in parenthesis as well as two MNIST digits (bottom right)

This further supports the observations in **C. Zhang** et al. 2017 that at lower object-to-image ratios, the model struggles to converge with noise that more closely resembles the distribution of the region of interest. For these reasons, in all our experiments, the noise digit thickness was set to 1.925 to strike a balance between being able to converge and being as close to the threshold as possible. During our experiments, we found that as the noise thickness became higher than the digits (ex: 5), the model would occasionally converge however the inconsistency at which it did did not allow us to include a structured observation in our results.

Lastly, we experimented with changing the number of noise digits on the canvas with a digit thickness of 1.925 with $\{100, 200, 300, 400, 600, 800\}$ noise digits and found that there was no discernible difference in performance on the training and validation sets (refer to Table 11 in the supplementary material). We argue that the number of digits on the canvas doesn't affect performance as the main driving factor to adversity lies in the resemblance of the noise to the object of interest.

5 Discussion

In this section, we discuss our findings as a whole as well as their overall implications for our work. Our experiments with changing the object-to-image ratio extends **Pawlowski** et al. 2020's experiments with Needle MNIST on CNNs to patch-based classifiers on the megapixel MNIST benchmark. Specifically, our work indicates that the O2I threshold below which IPS **(Bergner, Lippert, and Mahendran 2023)** fails to generalize is affected by both the task ("Maj", "Top", "Max", "Multi") and the training dataset size. There is a positive correlation between the object-to-image ratio and validation accuracy. The training data size interacts negatively with this effect where the larger the amount of instances used for training, the less generalizability suffers from low object-to-image ratios. We contribute to the existing literature by finding that this relationship is different for each task. For Tasks "Maj" and "Top" we found that the positive rate at which the dataset size affects the minimum O2I ratio threshold for generalization was at its highest, followed by tasks "Max" and "Multi" where in the latter, this rate is almost at 0.

Our attempts to robustify adverse scenarios with limited data and low O2I ratios revealed some interesting dynamics on the nature of the problem and future directions to consider. Firstly, we found that in a low-datasetting, tuning the patch size results in an improvement of validation performance of +30% for the megapixel MNIST and + 5% for the Swedish traffic signs dataset compared to the original patch sizes in IPS. In both datasets, a patch size that is smaller than the original implementation is preferred when a subset of the training data is used. For the Swedish traffic signs, we also empirically demonstrate that the smaller patch size results in less memory consumption (1.6 GB of peak memory for the original patch size of 100 and 1.3 GB for our tuned patch size of 75).

Our implementation of Diversity Loss to the multi-head cross-attention transformer of IPS successfully improved validation performance for the Swedish traffic signs dataset by +4% at the expense of 4× increase in standard deviation. Even with thorough tuning, we were not able to extend these gains to the megapixel MNIST. We argue that penalizing the cosine similarity of the attention maps for each head assumes that there exists a predefined number of equally discriminative and uniformly spaced attention maps. Specifically, our results on the megapixel MNIST dataset showed that each formulation tended to slow down converge. This outcome suggests that tackling the low data and O2I problem by altering the dynamics of the patch aggregation module to be constrained to assumptions related to the distance between saliency maps can result in a harder optimization problem.

Our additional implementation of Stochastic Top-K Instance Masking revealed that although STKIM was shown to be effective for breast cancer WSI, it is not well suited to high-dimensional images that contain no ambiguous patches. After experimenting with a wide parameter space and finding no improvement, further investigation (refer to Section 9.5) revealed that ambiguous patches are missing in the megapixel MNIST and Swedish traffic signs dataset to enable the model to learn alternative attention configurations to solve the task. These outcomes indicate that STKIM is dependent on the domain and likely an unviable direction for tackling the low data and O2I problem which is observable across domains.

Our work with scaling the patch encoder from a ResNet-18 to a ResNet-50 revealed that the increased memory requirements are hard to manage. Our work attempted to bypass the increased usage by freezing the layers of the ResNet-50 except for the last fully connected layer. We found that this configuration results in no improvement in validation performance across domains. Although other CNNs could be useful, we are not informed enough to make an educated recommendation. An educated recommendation would involve finding a hypothetical alternative that is superior to Residual models, deviating from the successful implementation of ResNet-50 as a feature extractor in multiple works **(Shao et al. 2021; Cordonnier et al. 2021)**.

6 Conclusion & future efforts

Our work identifies that as the object-to-image ratio decreases, performance suffers in low data settings for IPS **(Bergner, Lippert, and Mahendran 2023)**. We find that this vulnerability can be marginally mitigated by tuning the patch size and using Diversity loss for the Swedish traffic signs dataset. Future works should consider other memory-efficient patch-based classifiers such as Attention sampling **(Katharopoulos and Fleuret 2019)** and Differentiable top k **(Cordonnier et al. 2021)** to use as baselines. Further efforts should consider integrating a contrastive learning objective in the patch encoding phase. Similarly to **Kong** and **Henao** 2021, one could reverse the label of patches with low attention to 0 ($y = 1 \rightarrow 0$). This would in turn allow the contrasting loss to push positive instances closer in the embedding space while maximizing the distance between negative samples. The main advantage of contrastive loss is that it doesn't increase memory and runtime requirements. Additionally, **Kong** and **Henao** 2021 were successful in implementing contrastive learning across multiple domains including the Swedish traffic signs dataset and digital histopathology.

7 Acknowledgements

We would like to thank Dr. Chrissy Fell (cmf21@st-andrews.ac.uk) for her involvement and guidance throughout the dissertation, including providing ideas on how to structure experiments and stress test the architecture under more adverse noise conditions. We also extend our gratitude to Dr. Stuart Norcross (stuart.norcross@st-andrews.ac.uk) for his help with the GPU setup and connection to the university server. Additionally, we would like to thank Yunlong Zhang (zju@westlake.edu.cn), first author of "Attention-Challenging Multiple Instance Learning for Whole Slide Image Classification" **(Y. Zhang et al. 2024)**, for his help with our implementation of his Diversity regularisation and Stochastic Top-K Instance Masking.

8 Reproducibility

To reproduce our results and play around with the added features in IPS, the best level of detail is the code available at `https://github.com/MRiffiAslett/ips_MaxRiffiAslett.git`. Additionally, all model outputs for each 142 runs including the hyperparameters and full accuracy are available in the result_library folder in the repository.

References

Abouelela, **Ahmed** et al. (July 2005). "Automated vision system for localizing structural defects in textile fabrics". In: *Pattern Recognition Letters* 26.10, pp. 1435–1443. ISSN: 0167-8655. DOI: 10.1016/j.patrec.2004.11.016. URL: https://www.sciencedirect.com/science/article/pii/S0167865504003794 (visited on 06/16/2024).

Anand, **Deepak** et al. (2021). "Weakly supervised learning on unannotated H&E-stained slides predicts BRAF mutation in thyroid cancer with high accuracy". en. In: *The Journal of Pathology* 255.3, pp. 232–242. ISSN: 1096-9896. DOI: 10.1002/path.5773. (Visited on 06/23/2024).

Ba, **Jimmy**, **Volodymyr Mnih**, and **Koray Kavukcuoglu** (Apr. 2015). *Multiple Object Recognition with Visual Attention*. en. arXiv:1412.7755 [cs]. URL: http://arxiv.org/abs/1412.7755 (visited on 07/05/2024).

Baydas, **Senay** and **Bulent Karakas** (Dec. 2019). "Defining a curve as a Bezier curve". In: *Journal of Taibah University for Science* 13.1. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/16583655.2019.1601913, pp. 522–528. ISSN: null. DOI: 10.1080/16583655.2019.1601913. URL: https://doi.org/10.1080/16583655.2019.1601913 (visited on 07/23/2024).

Bergner, **Benjamin**, **Christoph Lippert**, and **Aravindh Mahendran** (Mar. 2023). *Iterative Patch Selection for High-Resolution Image Recognition*. arXiv:2210.13007 [cs, eess]. DOI: 10.48550/arXiv.2210.13007. URL: http://arxiv.org/abs/2210.13007 (visited on 05/29/2024).

Blondel, **Mathieu** et al. (Nov. 2020). "Fast Differentiable Sorting and Ranking". en. In: *Proceedings of the 37th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, pp. 950–959. URL: https://proceedings.mlr.press/v119/blondel20a.html (visited on 06/01/2024).

Cordonnier, **Jean-Baptiste** et al. (2021). "Differentiable Patch Selection for Image Recognition". en. In: pp. 2351–2360. URL: https://openaccess.thecvf.com/content/CVPR2021/html/Cordonnier_Differentiable_Patch_Selection_for_Image_Recognition_CVPR_2021_paper.html (visited on 06/01/2024).

Dehaene, **Olivier** et al. (Dec. 2020). *Self-Supervision Closes the Gap Between Weak and Strong Supervision in Histology*. arXiv:2012.03583 [cs, eess]. DOI: 10.48550/arXiv.2012.03583. URL: http://arxiv.org/abs/2012.03583 (visited on 05/30/2024).

Deng, **Jia** et al. (June 2009). "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. ISSN: 1063-6919, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848. URL: https://ieeexplore.ieee.org/document/5206848 (visited on 07/17/2024).

DeVries, **Terrance** and **Graham W. Taylor** (Nov. 2017). *Improved Regularization of Convolutional Neural Networks with Cutout*. arXiv:1708.04552 [cs]. DOI: 10.48550/arXiv.1708.04552. URL: http://arxiv.org/abs/1708.04552 (visited on 06/22/2024).

Gadermayr, **Michael** and **Maximilian Tschuchnig** (Mar. 2024). "Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations & future potential". In: *Computerized Medical Imaging and Graphics* 112, p. 102337. ISSN: 0895-6111. DOI: 10.1016/j.compmedimag.2024.102337. URL: https://www.sciencedirect.com/science/article/pii/S0895611124000144 (visited on 06/20/2024).

He, **Kaiming** et al. (Dec. 2015). *Deep Residual Learning for Image Recognition*. arXiv:1512.03385 [cs]. DOI: 10.48550/arXiv.1512.03385. URL: http://arxiv.org/abs/1512.03385 (visited on 07/17/2024).

Hou, **Le** et al. (Mar. 2016). *Patch-based Convolutional Neural Network for Whole Slide Tissue Image Classification*. en. arXiv:1504.07947 [cs]. URL: http://arxiv.org/abs/1504.07947 (visited on 06/19/2024).

Huang, **Zeyi** et al. (July 2020). *Self-Challenging Improves Cross-Domain Generalization*. arXiv:2007.02454 [cs]. DOI: 10.48550/arXiv.2007.02454. URL: http://arxiv.org/abs/2007.02454 (visited on 06/22/2024).

Ilse, **Maximilian**, **Jakub M. Tomczak**, and **Max Welling** (June 2018). *Attention-based Deep Multiple Instance Learning*. arXiv:1802.04712 [cs, stat]. DOI: 10.48550/arXiv.1802.04712. URL: http://arxiv.org/abs/1802.04712 (visited on 06/01/2024).

Jastrzebski, **Stanisław** et al. (Sept. 2018). *Three Factors Influencing Minima in SGD*. arXiv:1711.04623 [cs, stat]. DOI: 10.48550/arXiv.1711.04623. URL: http://arxiv.org/abs/1711.04623 (visited on 07/10/2024).

Katharopoulos, **Angelos** and **François Fleuret** (July 2019). *Processing Megapixel Images with Deep Attention-Sampling Models*. en. arXiv:1905.03711 [cs, stat]. URL: http://arxiv.org/abs/1905.03711 (visited on 05/31/2024).

Kong, **Fanjie** and **Ricardo Henao** (Dec. 2021). *Efficient Classification of Very Large Images with Tiny Objects*. arXiv:2106.02694 [cs]. DOI: 10.48550/arXiv.2106.02694. URL: http://arxiv.org/abs/2106.02694 (visited on 05/29/2024).

Krizhevsky, **Alex**, **Ilya Sutskever**, and **Geoffrey E. Hinton** (2012). "ImageNet classification with deep convolutional neural networks". In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'12. Red Hook, NY, USA: Curran Associates Inc., pp. 1097–1105. (Visited on 07/17/2024).

LaLonde, **Rodney**, **Dong Zhang**, and **Mubarak Shah** (June 2018). "ClusterNet: Detecting Small Objects in Large Scenes by Exploiting Spatio-Temporal Information". en. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA: IEEE, pp. 4003–4012. ISBN: 978-1-5386-6420-9. DOI: 10.1109/CVPR.2018.00421. URL: https://ieeexplore.ieee.org/document/8578519/ (visited on 06/16/2024).

Larsson, **Fredrik** and **Michael Felsberg** (2011). "Using Fourier Descriptors and Spatial Models for Traffic Sign Recognition". en. In: *Image Analysis*. Ed. by **Anders Heyden** and **Fredrik Kahl**. Berlin, Heidelberg: Springer, pp. 238–249. ISBN: 978-3-642-21227-7. DOI: 10.1007/978-3-642-21227-7_23.

Mahajan, **Dhruv** et al. (May 2018). *Exploring the Limits of Weakly Supervised Pretraining*. arXiv:1805.00932 [cs]. DOI: 10.48550/arXiv.1805.00932. URL: http://arxiv.org/abs/1805.00932 (visited on 07/10/2024).

Marra, **Francesco** et al. (Sept. 2019). *A Full-Image Full-Resolution End-to-End-Trainable CNN Framework for Image Forgery Detection*. en. arXiv:1909.06751 [cs]. URL: http://arxiv.org/abs/1909.06751 (visited on 05/29/2024).

Mnih, **Volodymyr** et al. (June 2014). *Recurrent Models of Visual Attention*. arXiv:1406.6247 [cs, stat]. DOI: 10.48550/arXiv.1406.6247. URL: http://arxiv.org/abs/1406.6247 (visited on 06/22/2024).

Nazeri, **Kamyar**, **Azad Aminpour**, and **Mehran Ebrahimi** (2018). "Two-Stage Convolutional Neural Network for Breast Cancer Histology Image Classification". In: vol. 10882. arXiv:1803.04054 [cs], pp. 717–726. DOI: 10.1007/978-3-319-93000-8_81. URL: http://arxiv.org/abs/1803.04054 (visited on 07/13/2024).

Ofli, **Ferda** et al. (Mar. 2016). "Combining Human Computing and Machine Learning to Make Sense of Big (Aerial) Data for Disaster Response". In: *Big Data* 4.1. Publisher: Mary Ann Liebert, Inc., publishers, pp. 47–59. ISSN: 2167-6461. DOI: 10.1089/

big.2014.0064. URL: https://www.liebertpub.com/doi/abs/10.1089/big.2014.0064 (visited on 06/07/2024).

Pawlowski, **Nick** et al. (Jan. 2020). *Needles in Haystacks: On Classifying Tiny Objects in Large Images*. (Visited on 05/31/2024).

Pinckaers, **Hans**, **Bram Van Ginneken**, and **Geert Litjens** (Mar. 2022). "Streaming Convolutional Neural Networks for End-to-End Learning With Multi-Megapixel Images". en. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.3, pp. 1581–1590. ISSN: 0162-8828, 2160-9292, 1939-3539. DOI: 10.1109/TPAMI.2020.3019563. URL: https://ieeexplore.ieee.org/document/9178453/ (visited on 05/29/2024).

Qin, **Xinqiang** et al. (Oct. 2013). "A novel extension to the polynomial basis functions describing Bezier curves and surfaces of degree n with multiple shape parameters". In: *Applied Mathematics and Computation* 223, pp. 1–16. ISSN: 0096-3003. DOI: 10.1016/j.amc.2013.07.073. URL: https://www.sciencedirect.com/science/article/pii/S0096300313008230 (visited on 07/23/2024).

Qu, **Linhao** et al. (Oct. 2022). *Bi-directional Weakly Supervised Knowledge Distillation for Whole Slide Image Classification*. arXiv:2210.03664 [cs]. DOI: 10.48550/arXiv.2210.03664. URL: http://arxiv.org/abs/2210.03664 (visited on 07/21/2024).

Ranzato, **Marc'Aurelio** (Apr. 2014). *On Learning Where To Look*. arXiv:1405.5488 [cs]. DOI: 10.48550/arXiv.1405.5488. URL: http://arxiv.org/abs/1405.5488 (visited on 07/18/2024).

Shao, **Zhuchen** et al. (Oct. 2021). *TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification*. arXiv:2106.00908 [cs]. DOI: 10.48550/arXiv.2106.00908. URL: http://arxiv.org/abs/2106.00908 (visited on 06/24/2024).

Sun, **Tianbo**, **Tong Meng**, and **Yutong Liu** (Sept. 2022). "CAMELYON 17 Challenge: A Comparison of Traditional Machine Learning (SVM) with the Deep Learning Method". en. In: *Wireless Communications and Mobile Computing* 2022. Ed. by **Nawab Muhammad Faseeh Qureshi**, pp. 1–9. ISSN: 1530-8677, 1530-8669. DOI: 10.1155/2022/9910471. URL: https://www.hindawi.com/journals/wcmc/2022/9910471/ (visited on 05/25/2024).

Szegedy, **Christian** et al. (Dec. 2015). *Rethinking the Inception Architecture for Computer Vision*. arXiv:1512.00567 [cs]. DOI: 10.48550/arXiv.1512.00567. URL: http://arxiv.org/abs/1512.00567 (visited on 07/30/2024).

Tang, **Wenhao** et al. (Dec. 2023). *Multiple Instance Learning Framework with Masked Hard Instance Mining for Whole Slide Image Classification*. arXiv:2307.15254 [cs]. DOI: 10.48550/arXiv.2307.15254. URL: http://arxiv.org/abs/2307.15254 (visited on 07/21/2024).

Thandiackal, **Kevin** et al. (2022). "Differentiable Zooming for Multiple Instance Learning on Whole-Slide Images". en. In: *Computer Vision – ECCV 2022*. Ed. by **Shai Avidan** et al. Cham: Springer Nature Switzerland, pp. 699–715. ISBN: 978-3-031-19803-8. DOI: 10.1007/978-3-031-19803-8_41.

Vaswani, **Ashish** et al. (Aug. 2023). *Attention Is All You Need*. arXiv:1706.03762 [cs]. DOI: 10.48550/arXiv.1706.03762. URL: http://arxiv.org/abs/1706.03762 (visited on 07/22/2024).

Wang, **Xiyue** et al. (Oct. 2022). "Transformer-based unsupervised contrastive learning for histopathological image classification". eng. In: *Medical Image Analysis* 81, p. 102559. ISSN: 1361-8423. DOI: 10.1016/j.media.2022.102559.

Xue, **Lanqing**, **Xiaopeng Li**, and **Nevin L. Zhang** (Dec. 2019). *Not All Attention Is Needed: Gated Attention Network for Sequence Data*. arXiv:1912.00349 [cs, stat]. DOI: 10.48550/arXiv.1912.00349. URL: http://arxiv.org/abs/1912.00349 (visited on 08/12/2024).

Zhang, **Chiyuan** et al. (Feb. 2017). *Understanding deep learning requires rethinking generalization*. (Visited on 06/22/2024).

Zhang, **Yunlong** et al. (Apr. 2024). *Attention-Challenging Multiple Instance Learning for Whole Slide Image Classification*. en. arXiv:2311.07125 [cs]. URL: http://arxiv.org/abs/2311.07125 (visited on 06/17/2024).

9 Supplementary material

In this section, we present extra results and visualization aimed at furthering our understanding of the shortcomings of our regularisation methods as well as additional insights into the outcomes of our O2I experiments.

9.1 Objet-to-image ratio

Table 10 displays the full validation and training accuracy of our object-to-image ratio experiments in Section 4.1. The training performance illustrates that regardless of the object-to-image ratio and varying training data sizes $\{800, 1000, 2000, 4000\}$, the model manages to memorize the training data. In the case of the tasks "Maj", "Max", and "Top", the training accuracy converges to 0.99 or higher regardless of the size of the training data. The task of identifying each digit, however ("Multi"), only learns the full training data when the training dataset size is equal to or greater than 4000. This indicates that task "Multi" doesn't only need extra training data to generalize but also to memorize the training data. These results suggest that the model is adept at finding regions of interest as shown by the performance of the tasks "Top", "Max" and "Maj", but falls short in identifying each number individually (task: "Multi").

				Validation (%)					Training (%)				
Data	Digit	Noise	O2I	Maj	Max	Top	Multi	loss	Maj	Max	Top	Multi	loss
800	28	800	0.01%	43.0	48.6	39.1	3.0	2.539	99.0	99.1	98.4	20.0	0.120
	56	600	0.03%	69.8	66.3	63.5	14.6	1.308	99.3	99.9	99.3	35.0	0.075
	84	400	0.08%	76.6	66.5	69.8	12.1	0.934	99.5	99.0	99.5	22.0	0.104
	112	200	0.14%	84.2	63.6	76.4	17.9	0.864	99.8	99.6	99.8	35.4	0.073
Data	Digit	Noise	O2I	Maj	Max	Top	Multi	loss	Maj	Max	Top	Multi	loss
1000	28	800	0.01%	66.6	69.2	62.0	19.6	1.400	99.5	99.5	99.4	35.1	0.077
	56	600	0.03%	79.4	70.2	70.9	19.4	0.967	99.0	99.1	99.2	33.5	0.083
	84	400	0.08%	91.0	71.5	79.7	25.8	0.698	99.9	99.9	100.0	54.6	0.045
	112	200	0.14%	89.3	66.8	78.4	24.6	0.729	99.6	99.4	99.5	45.5	0.065
Data	Digit	Noise	O2I	Maj	Max	Top	Multi	loss	Maj	Max	Top	Multi	loss
2000	28	800	0.01%	88.3	83.9	79.7	52.7	0.652	99.9	99.8	100.0	79.1	0.023
	56	600	0.03%	95.5	82.7	85.9	51.5	0.432	99.8	99.8	99.8	71.0	0.034
	84	400	0.08%	96.1	83.6	88.7	55.0	0.411	100.0	99.8	100.0	87.8	0.015
	112	200	0.14%	95.2	78.7	91.1	45.1	0.428	100.0	99.9	99.8	83.6	0.019
Data	Digit	Noise	O2I	Maj	Max	Top	Multi	loss	Maj	Max	Top	Multi	loss
4000	28	800	0.01%	98.3	92.2	91.3	78.6	0.250	100.0	100.0	100.0	97.5	0.004
	56	600	0.03%	98.9	93.0	93.8	80.2	0.213	100.0	100.0	100.0	97.9	0.003
	84	400	0.08%	97.9	89.6	92.8	72.1	0.272	100.0	100.0	100.0	93.3	0.008
	112	200	0.14%	98.7	90.7	93.2	75.3	0.243	100.0	100.0	100.0	95.8	0.006

Table 10: Full results of the experiments on megapixel MNIST with a novel noise generation component (refer to Section 4.1). Four Object-to-image (O2I) ratios were tested: $\{0.01\%, 0.034\%, 0.078\%, 0.13\%\}$ across four training dataset sizes $\{800, 1000, 2000, 4000\}$. Canvas size and patch size remain fixed at 3000×3000 and 50×50, respectively, and the O2I changed by varying the digit resolutions to 28×28, 56×56, 84×84, and 112×112 pixels. The model was trained for 100 epochs following the setup of IPS **(Bergner, Lippert, and Mahendran 2023)**.

Figure 11 displays the validation accuracy over all 100 epochs for the tasks 'Majority' (right) and 'Multi' (left), trained on 1000 samples across different object-to-Image (O2I) ratios. It illustrates that the object-to-image ratio is affecting the model's ability to tackle a subset of harder instances for task "Maj" but not for task "Multi". For task "Maj" (right of Figure 11), all O2I ratios show similar performance during the first 30 epochs. After this period, their convergence rates begin to diverge before eventually plateauing. Specifically, at an O2I of 0.01%, validation accuracy plateaus at 40% after epoch 50. For an O2I of 0.03%, it plateaus at 65% at epoch 80, while at O2Is of 0.078% and 0.14%, accuracy plateau at 75% and 85% respectively after epoch 90. These variations suggest that different O2I ratios impact the model's ability to learn from harder examples, with distinct bottlenecks observed at each ratio. An interesting experiment would be to further weigh these harder instances to uncover whether it would improve generalization.

In contrast, for the task 'Multi' (left) the O2I ratio has a lesser effect on performance (refer to Figure 11). A higher O2I ratio of 0.08% results in slower convergence and a lower final validation accuracy of 12%, while a lower O2I ratio of 0.01% leads to faster convergence and a higher validation accuracy of 15%. This demonstrates that for task 'Multi', the O2I ratio does not significantly impact the model's training dynamics compared to task 'Maj'.

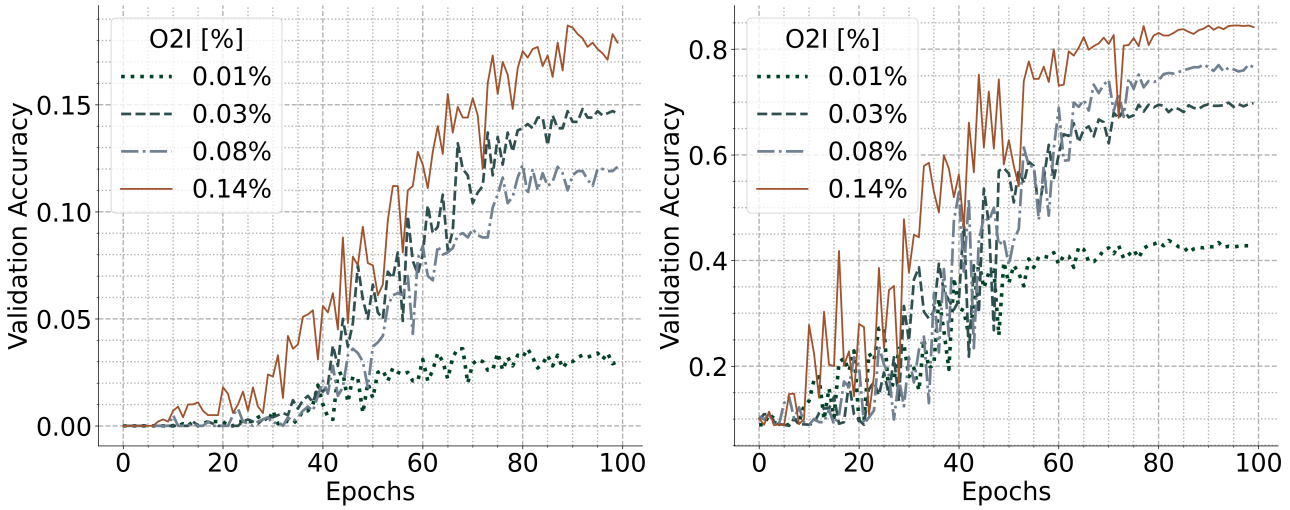


Figure 11: Training dynamics for tasks "Multi" (left) and "Maj" (right) on the validation set over 100 epochs with 1000 training samples and the following object-to-image ratios: $\{0.01\%, 0.034\%, 0.078\%, 0.13\%\}$.

9.2 Task: Max (random accuracy)

In Section 4.8 we found that validation accuracy for all tasks was random when the noise thickness reached a threshold of 2. After this threshold, validation accuracy was 10%, 10%, and 0% for tasks "Maj", "Top" and "Multi" which aligns with the number of random outcomes for each task which is 10, 10, and 100,000 respectively. The random accuracy for task Max was 30% however which is less straightforward to interpret. Here we detail the intuition that leads us to believe that by consistently choosing the digit 9 for the task "Max" in the megapixel MNIST, it will achieve a random accuracy close to 30% (refer to Section 4.8). For task "Max", where the goal is to find the highest digit out of five digits with three of them being the same, we observed that the number 9 is predicted most often, resulting in a random accuracy of about 30%. We frame the task as finding the probability that a number is the maximum in a set of three different randomly chosen digits. The total number of ways to choose 3 different numbers from 10 can be computed by the binomial coefficient $\binom{10}{3} = 120$. For a number x to be the highest, the other 2 numbers must be less. The number of ways to choose 2 numbers from $(x-1)$ numbers is $\binom{x-1}{2}$. The probability that x is the highest among the 3 chosen numbers in the megapixel MNIST is therefore $\binom{x-1}{2}/120$. The probabilities for each digit from 0 to 9, are as follows{ x = 0: 0.00, x = 1: 0.00, x = 2: 0.01, x = 3: 0.03, x = 4: 0.05, x = 5: 0.08, x = 6: 0.13, x = 7: 0.18, x = 8: 0.23, x = 9: 0.30 }. By consistently predicting digit 9, the model maximizes its random performance which sits at 30%.

9.3 Effect of the number of noise digits on optimization

We experimented with changing the number of noise digits on the canvas when there thickness is set to 1.925 (thickness used in all experiments). Results in Table 11 show that increasing the number of noise digits does not affect performance. For task "Maj" for instance when 100 noise digits are placed on the 3000×3000canvas, the validation sits at 99% and remains at 99% with 800 noise digits on the canvas. This indicates that the main contributor of our proposed noise generation component to the performance is the thickness of the noise digit.

	Validation					Training				
Noise	Maj (%)	Max (%)	Top (%)	Multi (%)	loss	Maj (%)	Maj (%)	Top (%)	Multi (%)	loss
100	99.0	94.5	92.7	83.6	0.195	100.0	100.0	99.9	98.1	0.003
200	98.8	93.9	93.6	84.0	0.193	100.0	100.0	99.9	97.8	0.004
300	99.0	94.2	94.4	83.8	0.194	100.0	99.9	99.9	98.2	0.003
400	98.9	94.6	92.5	82.7	0.199	100.0	100.0	100.0	97.9	0.004
600	99.0	94.2	94.2	82.6	0.194	100.0	100.0	100.0	97.5	0.004
800	99.1	92.7	93.5	83.4	0.211	100.0	100.0	99.9	97.9	0.003

Table 11: Full training and validation results for experiments varying the number of noise digits on the canvas (refer to Section 4.8). The size of the digits and the noise component are fixed at 28×28 on a 3000×3000 canvas. The noise is measured in the number of noise components added to the canvas.

9.4 Diversity regularisation

In this section, we visualize the resulting training dynamics on the validation sets with different Diversity regularisation weights corresponding to our experiment in Section 4.5. Results in Figure 12 showcase the dynamic validation accuracy of tasks "Max" (left) and "Maj" (right) across 100 epochs on the megapixel MNIST dataset with a 3000×3000 canvas and an O2I ratio of 0.01%. It is evident that the Diversity loss affects the validation accuracy throughout training. For task "Max" (left), when the Diversity is set to 0, validation accuracy plateaus at 70% after epoch 60 whereas, for weightings $1 \times D$, $2 \times D$, and $3 \times D$ it plateaus at 55%, 59%, and 40% respectively after epoch 60. Noticeably, higher Diversity doesn't always result in lower performance. For instance, for the task Majority, (right of Figure 12) the weight $2 \times D$ converges higher (58%) than the weight $1 \times D$ (40%). This is in line with the increased standard deviation brought by the Diversity loss observed in Section 4.5.

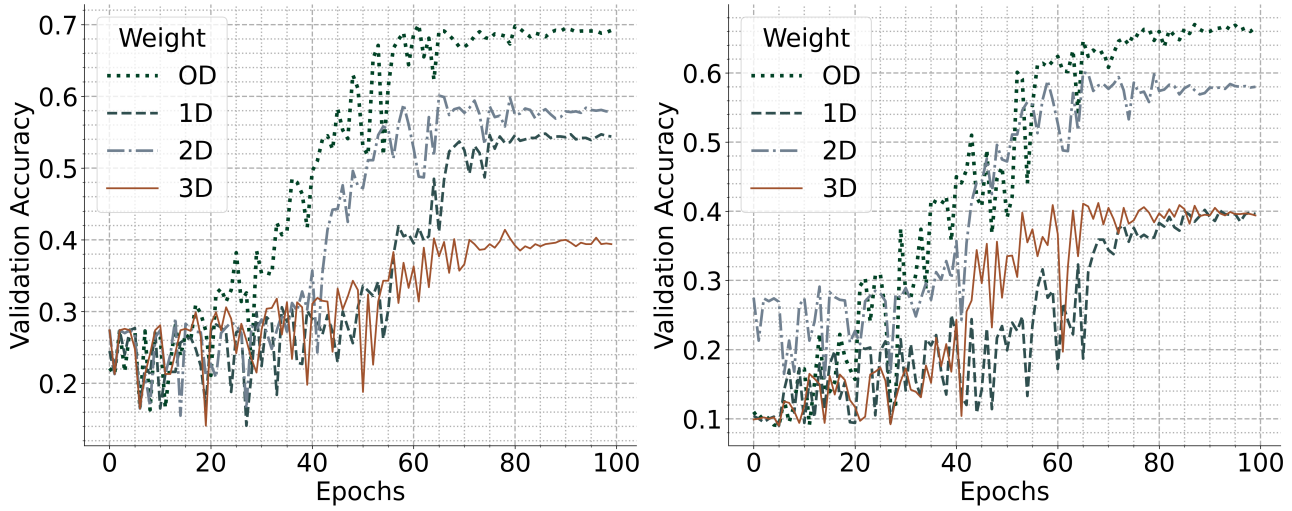


Figure 12: Validation accuracy for tasks "Max" (Left) and "Maj" (Right) for different weights of Diversity loss, namely $\{0 \times D, 1 \times D, 2 \times D, 3 \times D\}$. For $\{0 \times D\}$ the Semantic loss is also set to 0. The object-to-image ratio is fixed at 0.01%.

9.5 Stochastic Top-K Instance Masking

In this section, we further explore the effects of Stochastic Top-K Instance Masking on the resulting attention distributions. For each configuration of P and K conducted in the experiments in Section 4.6 on megapixel MNIST, we plot the resulting attention distribution at the 100th epoch on a single image from the megapixel MNIST test set. K controls the number of instances to mask and P the proportion of instances to mask in the subset K. The resulting distributions, shown in Figure 13, highlight the shortcomings of Stochastic Top-K Instance Masking when applied to the megapixel MNIST dataset. Notably, as the hyperparameters change towards masking a larger proportion P of the K instances (right of Figure 13), we would expect the attention distribution for the patches to have lower variance where the attention distribution is less dominated by a subset of informative patches.

The attention distributions however showcase that as P increases, the variance of the distribution neither increases nor decreases. Rather it sporadically goes up and down with no apparent trend. For instance, the attention distribution on the right of Figure 13 shows the lowest variance is achieved when $P = 60\%$ and $K = 10$. The highest variance on the other hand is obtained when when $P = 80\%$ and $K = 10$. When masking is deactivated ($P = 0\%$ and $K = 0$), the variance falls between these two extremes. The same inconsistencies are evident as the hyperparameters change towards masking more instances K while P remains fixed (left of Figure 13). As K increases, the variance increases and decreases around the baseline where $P = 0\%$ and $K = 0$. For instance, when K is set to 30 and P to 20%, it results in the highest variance. However, when K is set to 40, the resulting distribution features the lowest variance. We would expect the variance of the attention distribution to gradually decrease with K as a higher number of informative patches are being masked.

We find that this is due to the number of informative patches. As the attention Maps in Section 9.6.1 illustrate, the attention distribution is dominated by 2 to three instances for both the megapixel MNIST and Swedish traffic signs dataset. With the STKIM implementation, masking the top K patches results in either masking no informative patches which slightly decreases variance, or masking one of the 3 to 5 very informative patches which drastically decreases variance as attention is dominated by one less instance. This is due to the absence of ambiguous patches in the megapixel MNIST dataset, which are omnipresent in the CAMELYON16 dataset (the dataset where STKIM was shown to be effective). We conclude that STKIM is not suitable for datasets where the number of informative patches is 1) consistently very low and 2) where no ambiguous patches are present.

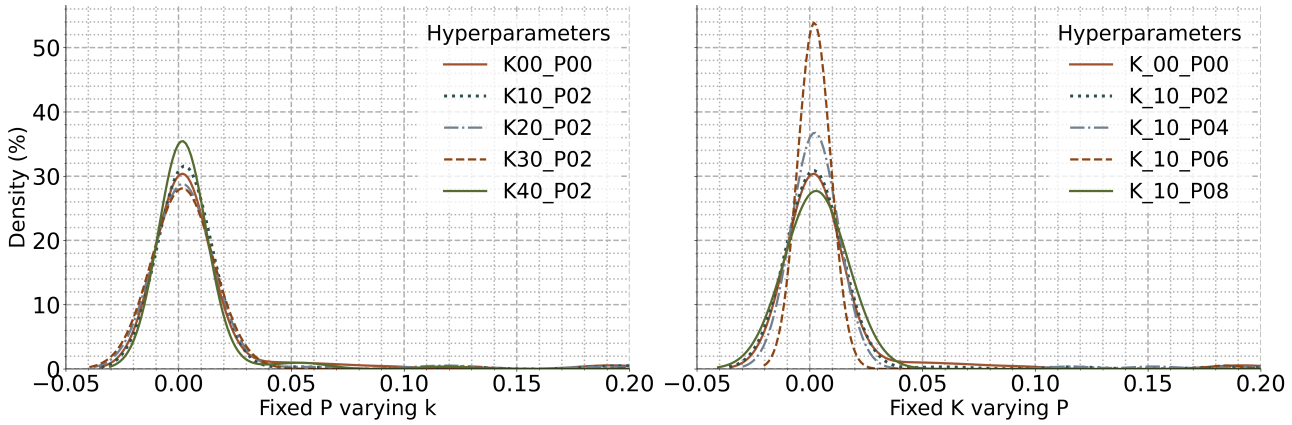


Figure 13: Attention density corresponding to the different hyperparameter values tested for Stochastic Top-K Instance Masking (refer to Section 4.6). The plot on the left corresponds to attention distributions with a fixed P and different values of K, while the plot on the right showcases the attention distributions with different values of P for a fixed K. The densities were obtained from the same image in the megapixel MNIST dataset.

9.6 Attention Maps

9.6.1 Megapixel MNIST

In this section, we showcase attention maps for the first instance in the validation set of the megapixel MNIST dataset. There are four maps, each with a different Object-to-Image (O2I) ratio, showcased in Figures 14 and 15. Each map was produced by running identical models for 60 epochs, with 1000 training samples and the following object-to-image ratios: 0.01% (left of Figure 14), 0.034% (right of Figure 14), 0.078% (left of Figure 15), and 0.13% (right of Figure 15) on a 1500×1500 canvas. The attention maps demonstrate that the lower the object-to-image ratio, the more uncertain the attention maps are in distinguishing between the noise and MNIST digits. Although all of them solve the task, the lowest object-to-image ratio (0.01%) (left of Figure 14) includes many informative patches (in purple) where the patch overlays noise. On the other hand, at the highest object-to-image ratio (0.13%) in Figure 15 on the right, there are no informative patches that cover noise digits. This pattern is consistent with the O2I ratios in the center of both extremes, where at 0.034% (right of Figure 14), more patches cover noise digits than at 0.078%, where fewer informative patches cover noise digits. The implications are that at higher object-to-image ratios, IPS can better distinguish noise digits from MNIST digits.

Additionally, the attention maps showcase how IPS is flexible in using the same patch size for detecting regions of interest that are superior and inferior to the dimensions of the patch. In Figure 14 on the left, where the object size is 28×28, one or two patches (50×50) are sufficient to cover each digit. In contrast, in Figure 15 on the right, where object size is 112×112 (O2I: 0.13%), up to eight patches are used by IPS to detect a single MNIST digit. This further supports that changing the object-to-image ratio using the same patch size is valid, as multiple patches can attend to a single digit.

9.6.2 Swedish traffic signs

In the megapixel MNIST dataset, the object-to-image ratio stays fixed across the training and validation sets allowing us to tailor the patch size to best fit the region of interest. In the Swedish traffic signs benchmark, however, the object-to-image ratio changes (from 0.05% to 0.4%) as some signs are further away from the cameras than others. Here, we visualize how IPS detects regions of interest of varying sizes using the same patch size or field of view. We trained two models: the first with a patch size and stride of 25×25, and the second with a patch size and stride of 100×100. The model weights were taken after 140 epochs of training, and the maps were plotted on images 82 and 86 of the Swedish traffic signs dataset. Both images contain the same scene but with different object-to-image ratios, 0.05% for image 82 (Figures 16 and 18) and 0.4% for image 86 (Figures 17 and 19) . The attention values are normalized before plotting.

Figures 16, 17, 18, and 19 illustrate that when the patch size is lower than or equal to the smallest region of interest, inference is more stable across object-to-image ratios. When the patch size is fixed at 25×25, in Figures 16 and 17 we can see that the model is able to find clusters of patches where a single one is more informative than the others. In Figure 16 (O2I ratio: 0.05%), there is a cluster of 2 patches overlaying the sign where one (in yellow) is substantially more informative than all patches in the image. In Figure 17 (O2I ratio: 0.4%), there are 12 patches covering the region of interest where a single patch (yellow) is more informative than all other patches.

In contrast, in Figures 18 and 19, the patch size of 100×100 is substantially bigger than the smallest object. In Figure 18 (O2I ratio: 0.05%), the object of interest (27×27) occupies less than one-third of the patch (100×100). The image shows that all of the informative patches in the image are in the same range of attention (yellow) meaning IPS fails to find the single patch that solves the problem. This is undesirable as being able to delineate salient regions and non-informative patches is vital in certain domains. For Digital histopathology, for instance, attention is used to highlight regions that likely contain cancer cells for further diagnosis. When the object-to-image ratio increases to 0.4% in Figure 19, a single patch is highlighted as informative. These observations imply that a smaller patch size is advantageous in an inference setting as IPS can more easily delineate between informative and non-informative regions at low O2I ratios. A good rule of thumb would be to ensure that the patch size matches the size of the smallest object in the dataset.

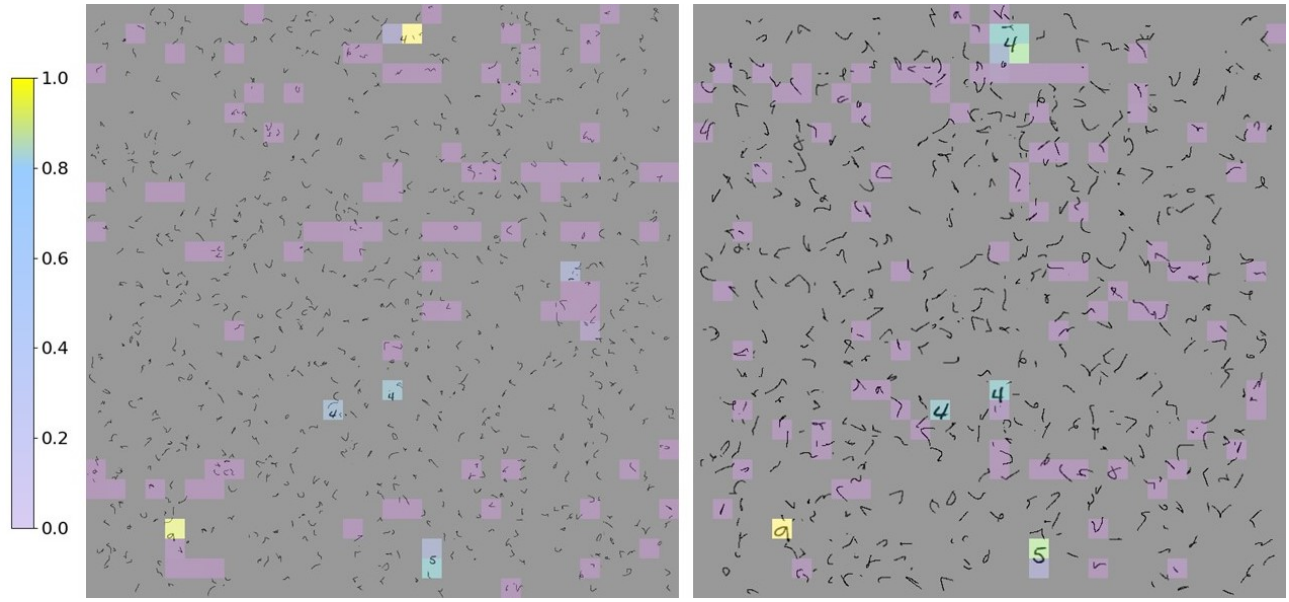


Figure 14: Attention maps for different object-to-image ratios: 0.01% (left) and 0.034% (right) on a 1500×1500 canvas, with 800 noise digits on the left and 600 on the right. The maps display the top M (100) most informative patches at the end of a full forward pass with IPS. The digit and noise size on the left is 28×28 and on the right, 56×56.

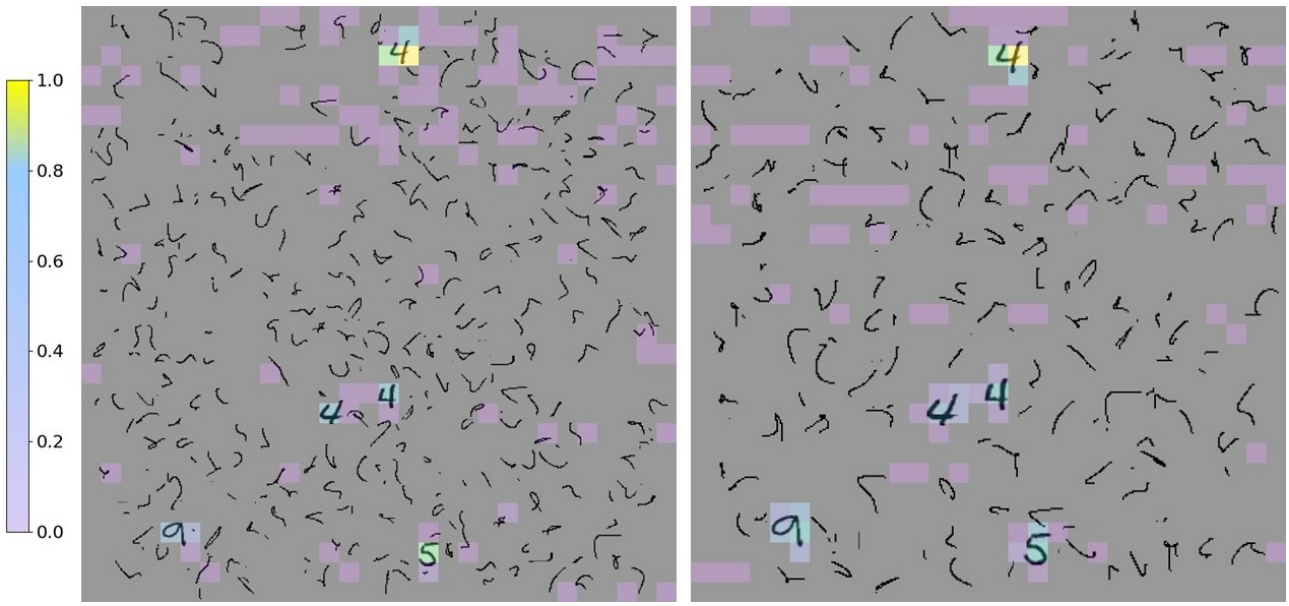


Figure 15: Attention maps for different object-to-image ratios: 0.078% (left) and 0.13% (right) on a 1500×1500 canvas, with 400 noise digits on the left and 200 on the right. The maps display the top M (100) most informative patches at the end of a full forward pass with IPS. The digit and noise size on the left is 84×84 and on the right, 112×112.

Figure 16: Attention maps for image 82 (validation set) of the Swedish traffic signs dataset. IPS was run for 140 epochs with a patch size and stride of 25. The memory buffer M is set to 20 patches. All 20 patches M selected by IPS are outlined in red and the color corresponds to the normalized attention value.

Figure 17: Attention maps for image 86 (validation set) of the Swedish traffic signs dataset. IPS was run for 140 epochs with a patch size and stride of 25. The memory buffer M is set to 20 patches.

Figure 18: Attention maps for image 82 (validation set) of the Swedish traffic signs dataset. IPS was run for 140 epochs with a patch size and stride of 100. The memory buffer M is set to 10 patches. All 20 patches M selected by IPS are outlined in red and the color corresponds to the normalized attention value.

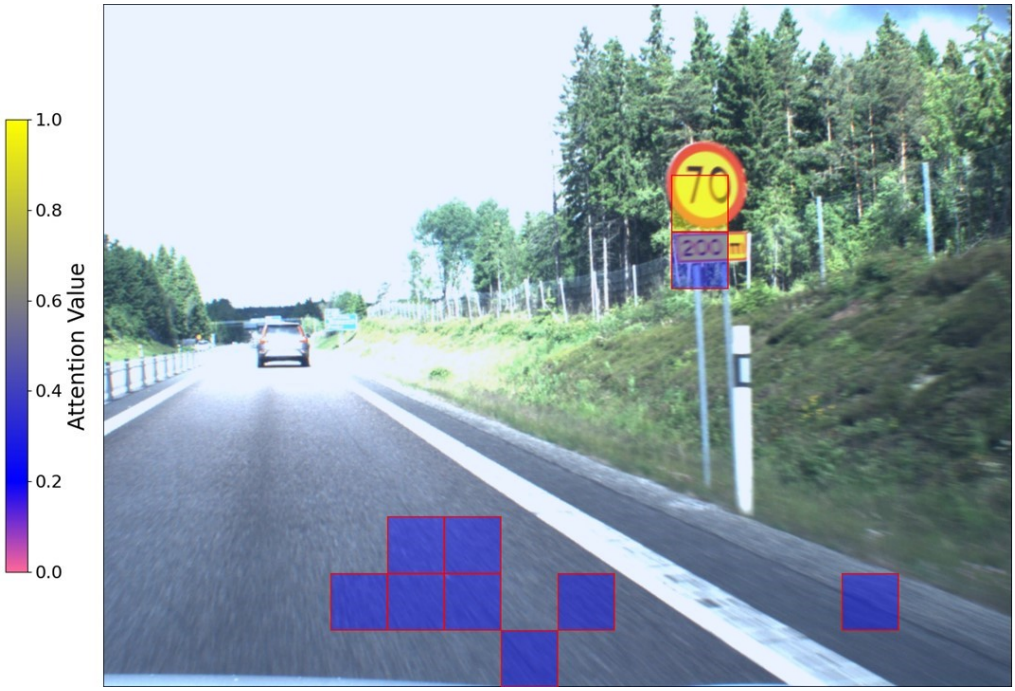


Figure 19: Attention maps for image 86 (validation set) of the Swedish traffic signs dataset. IPS was run for 140 epochs with a patch size and stride of 100. The memory buffer M is set to 10 patches.