Sub-token ViT Embedding via Stochastic Resonance Transformers

Dong Lao [1] **Yangchao Wu** [1] **Tian Yu Liu** [1] **Alex Wong** [2] **Stefano Soatto** [1]

Abstract

Vision Transformer (ViT) architectures represent images as collections of high-dimensional vectorized tokens, each corresponding to a rectangular non-overlapping patch. This representation trades spatial granularity for embedding dimensionality, and results in semantically rich but spatially coarsely quantized feature maps. In order to retrieve spatial details beneficial to fine-grained inference tasks we propose a training-free method inspired by "stochastic resonance." Specifically, we perform *sub-token spatial transformations* to the input data, and aggregate the resulting ViT features after applying the *inverse transformation*. The resulting "Stochastic Resonance Transformer" (SRT) retains the rich semantic information of the original representation, but grounds it on a finer-scale spatial domain, partly mitigating the coarse effect of spatial tokenization. SRT is applicable across any layer of any ViT architecture, consistently boosting performance on several tasks including segmentation, classification, depth estimation, and others by up to 14.9% without the need for any fine-tuning. Code: https://github.com/donglao/srt.

1. Introduction

The Transformer architecture (Vaswani et al., 2017), originally designed for modeling language which is naturally quantized into discrete objects (sub-word "tokens"), is a poor fit for vision tasks due to the lack of a natural scale for spatial discretization: The same object can disappear within a pixel or fill the entire image plane depending on its distance from the camera. In theory, one could create tokens for patches of all sizes and positions, but at significant computational expense due to the complexity of transformers, which is quadratic in the number of tokens. Despite the counter-intuitive nature of spatial quantization, Vision Transformers (ViTs) (Dosovitskiy et al., 2020) achieve state-of-the-art performance in many vision tasks. So we focus on developing methods to harness pre-trained ViTs and overcome their limitations in representing spatial details at fine granularity due to the fixed spatial quantization of tokens.

The standard remedy for quantization artifacts is anti-aliasing. In one-dimensional digitized signals such as audio, anti-aliasing refers to weighted averaging of nearby samples in the discrete topology, or equivalently averaging versions of the signal translated by *integer multiples of the original sampling interval*. For images, in addition to sampling the translation group, one also has to sample the scale group, so as to capture the varying size of the projection of objects onto the image plane. Various network architectures comprise spatial average pooling, which is translational anti-aliasing, whereas the notion of domain-size pooling and domain anti-aliasing has been championed by (Dong & Soatto, 2015). Anti-aliasing is typically performed by convolving the discrete signal with a generic (not data-dependent) kernel. The optimal kernel is unbounded, so any finite implementation is necessarily lossy and cannot "recreate information" lost in the sampling process. Similarly, super-resolution algorithms hallucinate missing details either using generic priors or data other than the signal in question (Buades et al., 2005).

"Stochastic Resonance" (Benzi et al., 1981) is a qualitatively different process whereby the limitations imposed by a fixed quantization threshold can be overcome simply by shifting the signal by sub-threshold additive perturbations. This results in sampling beyond the Nyquist limit otherwise imposed by the quantizer. We extend this process, originally employed in cochlear implants, to translation *not* of the value of the signal (additive perturbations) but its domain (translation). The same process can also be applied to domain size (scale). Stochastic resonance can be thought of as a form of data augmentation or adaptive quantization. We further simplify it by choosing deterministic, rather than randomly sampled, perturbations. The perturbed token embeddings are aggregated statistically to first-order (mean or median) to yield a sub-token embedding. First-order statistics can be used for visual tasks, for instance, unsupervised object segmentation, and second-order statistics as a weight for adaptive regularization.

[1]UCLA Vision Lab [2]Yale Vision Lab. Correspondence to: Dong Lao <lao@cs.ucla.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

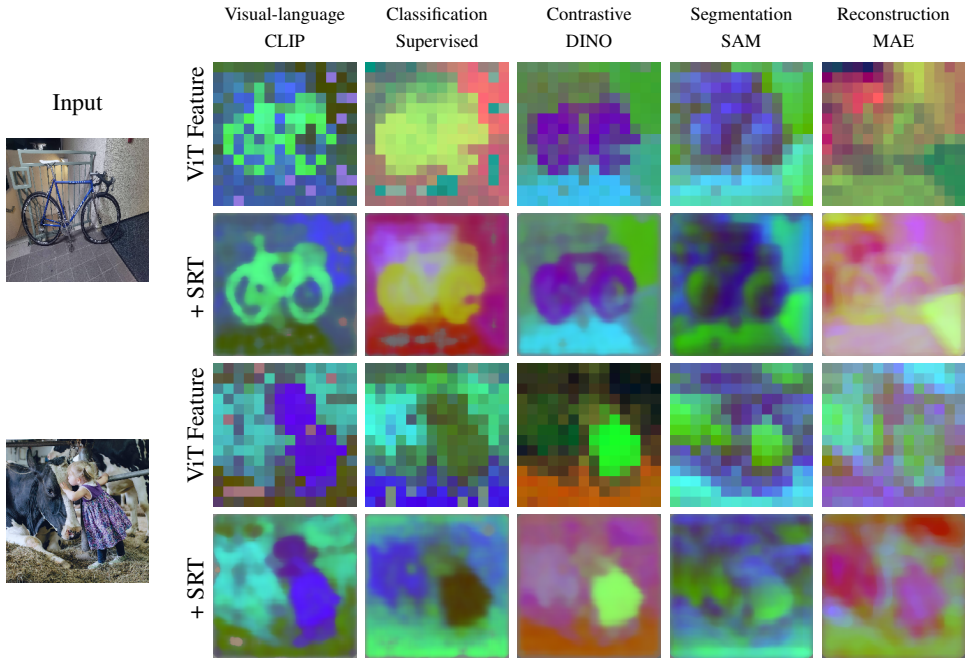

Figure 1. **High-resolution ViT features computed by stochastic resonance.** *Stochastic Resonance enables enhancing tokenized ViT features during inference without the need for additional training or modifying ViT forward pass. Here we present enhanced features from different pre-trained ViT models, visualized via Principal Component Analysis: CLIP (Radford et al., 2021) captures major image components. Interestingly, although Supervised (Dosovitskiy et al., 2020) and DINO (Caron et al., 2021) are trained by different pipelines and training loss, they prioritize similar regions. This may be due to they are trained on the same dataset and thus capture similar inductive bias. In contrast, SAM (Kirillov et al., 2023) and MAE (He et al., 2022) capture local features over high-level semantics.*

This simple approach is well suited to pre-trained transformers since it only requires acting on inputs and outputs without modifying (or even knowing) the weights or the forward pass of the model. We call the resulting method "Stochastic Resonance Transformer" although we do not modify the transformer nor do we use artificial noise, to reflect closer proximity of our method to Stochastic Resonance than to traditional super-resolution or anti-aliasing methods. The simplicity of the method allows us to leverage ViTs, pre-trained on large datasets, such as CLIP (Radford et al., 2021) and DINO (Caron et al., 2021), to improve their handling of spatial quantization. This may help attenuate some of the biases of certain training procedures, for instance, the object-centric nature of contrastive learning, e,g, DINO, which biases the representation towards centered objects that occupy a large portion of the visual field.

Stochastic Resonance can be used as a form of sub-token ensembling, to enhance feature maps in ViTs and reveal some of the local fine-grained underlying structure. SRT can be applied to any ViT layer, on any task, without altering network architecture or pre-trained network weights. We can use SRT to visualize fine-grained features, or optionally map them back to the original ViT feature scale by pooling to be used for inference, where we notice performance increases on a wide range of vision tasks. Additionally, fine-tuning pre-trained ViTs by distillation from ensembled features maintains their original inference time and cost. To the best of our knowledge, SRT is the first approach to recoup spatial granularity from embedding dimensions in ViT feature maps. Unlike conventional ensemble methods that augment model inputs and combine outputs, SRT operates at the feature level, allowing seamless integration into any ViT pipelines, including for tasks that demand intermediate features or attention maps. Our contributions are summarized as follows:

- We introduce a novel technique, namely the Stochastic Resonance Transformer (SRT), that computes fine-grained ViT embeddings at test time without additional training or modifications to the ViT's forward pass.
- SRT can be seamlessly integrated into any task that utilizes ViT as a feature extractor, serving as a test-time augmentation and ensemble method.
- We provide an efficient implementation of SRT, including parallelization and recursive aggregation, which reduces computational and memory requirements.
- We showcase the effectiveness of SRT by consistent improvement on a range of diverse vision tasks. Notably, it demonstrates significant enhancements on dense predic-

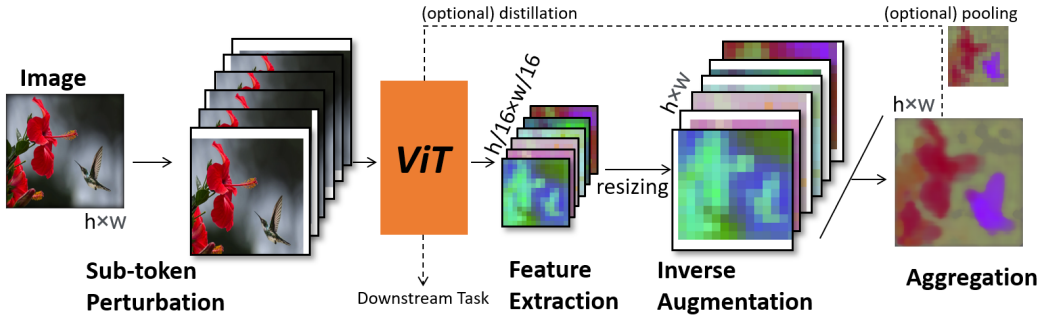

Figure 2. **Schematic for SRT.** *SRT applies controlled perturbations to input images, extracting features through Vision Transformers (ViTs). These features are then upsampled to higher resolution and aligned using the inverse of the applied perturbations. Statistical aggregation, including mean and median, along the perturbation dimension, produces fine-grained feature representations. These features find utility in visualization and can also be seamlessly integrated back into the network for enhanced performance in downstream tasks.*

tion tasks, of up to 14.9% on depth prediction.
- SRT also yields a versatile visualization tool that can be applied to any layer of any pre-trained ViT model, offering valuable insights into ViT model characteristics.

2. Stochastic Resonance Transformer

2.1. Method

Given an image x with $N \times M$ resolution, a Vision Transformer (ViT) divides it into tokens, where each token represents a $n \times m$ rectangular patch. While tokens can technically overlap, practical ViT models often use non-overlapping tokens for efficiency due to the quadratic complexity of transformers with respect to the number of tokens. Consequently, in a certain layer of ViT, this approach yields a feature map with dimensions $\frac{N}{n} \times \frac{M}{m} \times C$, where C is the size of the feature vector determined by architecture, downsampled from the original image and subsequently losing sub-token spatial information.

Given a trained ViT model, we aim to obtain features in a higher resolution that preserve the spatial information on a pixel level, ideally matching with the original image input. Fig. 2 illustrates our proposed pipeline of SRT. To enhance the features, we introduce sub-token perturbations to the input, i.e. transforming the coordinates of the input and resampling onto a new image plane, and extract embeddings from the resulting perturbed image. Note that any group transformation (translation, rotation, flipping, etc.) and a combination of them (*e.g.* (Wu et al., 2023)) can be chosen as the perturbation, provided that its inverse transformation is available. However, our specific interest lies in introducing perturbation through translation. This preference arises from several factors: 1) translation preserves the object's scale, unlike zooming; 2) translation can be applied at the pixel level grid, eliminating interpolation artifacts, as opposed to rotation; 3) translation allows an efficient implementation of SRT (Section 2.3).

We then upsample the resulting low-resolution embeddings back to the original image resolution $N \times M$ and apply an inverse of the perturbation to the spatial coordinates of the embeddings, and through an inverse warp, align it with the original input image. By repeating this process on different sub-token perturbations for t times, we generate a collection of embeddings, denoted by $N \times M \times C \times t$, that are spatially aligned to the input frame of reference. We can then compute statistics, *e.g.* mean or median, along the t dimension. Consequently, we obtain a feature field $N \times M \times C$, with the same spatial resolution as the original image. As showcased in Fig. 1, the embeddings are enhanced to sub-token resolution. This process is similar to Stochastic Resonance, where introducing white noise to the input signal enhances a signal beyond the native resolution. These embeddings offer promising downstream applications, as in Section 3.

For any task that utilizes ViT as a feature extractor, we can take an additional step by applying average pooling to again tokenize this high-resolution feature, to map it to $\frac{N}{n} \times \frac{M}{m} \times C$. It's important to note that this feature differs from the one obtained from one single forward pass of ViT, as it is an aggregate of multiple perturbed inputs. This process can be viewed as test-time augmentation and ensemble. Since this feature is compatible with the original ViT architecture, it can be seamlessly integrated back into the layer from which we perturbed the features, and applies to any model at any layer, regardless of pre-training, without requiring additional learned modules or altering the forward pass. Such a pipeline improves performance on diverse computer vision tasks, as validated by Section 3. Next, we formalize the aforementioned pipeline.

2.2. Formalization

$x \in \mathbb{R}^{N \times M \times K}$ is a K-channel signal (*e.g.*, $K = 3$ for a color image.) Let $\pi : \mathbb{R}^{N \times M} \to \mathbb{R}^{n \times m}; x \mapsto x$ a projection (subsampling, $n \ll N, m \ll M$), with the corresponding

inverse (interpolation) map $\pi^{-1} : \mathbb{R}^{n\times m} \to \mathbb{R}^{N\times M}; x \mapsto x$ be piecewise constant. This is a trivial form of subsampling and interpolation with a constant kernel.

Now, let $\phi : \mathbb{R}^{NMK} \to \mathbb{R}^{nmC}$ be a trained model with C channels of feature maps, typically $C \gg K$. Finally, let $T : \mathbb{R}^{N\times M} \to \mathbb{R}^{N\times M}; x \mapsto Tx$ be a compact and invertible transformation, for instance, edge-padded shift by a number of pixels smaller than $(N-n)/n \times (M-m)/m$. We consider uniform random padded shifts (translation) and consider the following measurement process:

$$y_t = \phi(T_t x) \tag{1}$$

for all random transformations T_t. We wish to enhance the output of ϕ from $n \times m$ to $N \times M$. We call this process *immersion* since each point x maps to $z = \phi(x)$ but $z \neq T^{-1}\phi(Tx)$. In other words, x is mapped injectively but not bijectively, since there are as many (vector)-valued embeddings as sampled value of T. We do so iteratively by averaging (or by a linear transformation K_t) with respect to the innovation process:

$$\epsilon_t = \underbrace{\pi\left(T_t^{-1}\pi^{-1}y_t\right)}_{\hat{y}_t} - K_t\phi(x) \tag{2}$$

now the fine-grained features which we call $\hat{x}_t$ are obtained by an observer architecture, which implements a closed-loop dynamical system of the form:

$$\begin{cases} \hat{x}_{t+1} = \hat{x}_t + T_t^{-1}\pi^{-1}y_t \quad \hat{x}_0 = 0; \\ y_t = \phi(T_t x) \end{cases} \tag{3}$$

This is just a moving average in higher resolution, whereby the variance of $\hat{x}$ will decrease to a steady state (by Central Limit Theorem), following the practice of stochastic resonance. It is a mixture of upsampling/interpolation and inference-time data augmentation, or ensembling.

2.3. Efficient Implementation

In theory, there is no limitation on the types of sub-token transformations that can be employed. We opt for a straightforward approach by applying translations (with padding) and this practice demonstrates effective results. We sample translations at the pixel level, avoiding the need for sub-pixel interpolation, which could introduce unwanted artifacts.

For a ViT utilizing token sizes of $m \times n$, we impose a constraint on the maximum magnitude of translation, limiting it to $\frac{m}{2} \times \frac{n}{2}$. This constraint allows the model to explore all possible token selections within the image. It is worth noting that excessive translation can be counterproductive when applied to downstream vision tasks, as it can result in information loss at the image boundaries. A detailed discussion can be found in Section 3.2, where we study the relation between perturbation level and model performance.

While naive implementation can lead to significant computational drawbacks, running inference on each augmented image can be trivially parallelized. Greater implementation speed-ups can also be achieved by bypassing the upsampling step (which is computationally expensive), since the aggregated result can be deterministically computed from the original feature maps of each augmented image when average pooling is used. With ViT-16/S architecture, on DAVIS-2017 (Pont-Tuset et al., 2017) our implementation of SRT runs at 1.0 seconds per image on a Nvidia 3090 GPU using a perturbation level of 3 pixels. To further speed up, one may optionally fine-tune the ViT model by distilling utilizing SRT, so that the inference time and cost remain, as demonstrated in Section 3.2.

3. Experiments

3.1. Visualization of SRT Features

SRT demonstrates significant promise in visualizing features of ViT models. It achieves this without necessitating modifications to the ViT's forward pass. In Fig. 1, we present visualizations of the final layer features from five popular ViT models, all employing the ViT-B/16 architecture. Notably, all visualizations are computed by a standard consumer laptop. We employ SRT with a turbulence level of 7 pixels to traverse non-overlapping augmented tokens extensively. The resultant high-dimensional features then go through Principal Component Analysis (PCA), with the top three components mapped to RGB channels to facilitate effective visualization. Despite sharing the same architecture, the five models exhibit distinct characteristics owing to variations in their pre-training supervision. For instance, CLIP (Radford et al., 2021) is trained through contrastive visual-language pre-training and captures major image components in the displayed examples. The Supervised model (Dosovitskiy et al., 2020) is trained for ImageNet classification, while DINO (Caron et al., 2021) undergoes contrastive learning. Interestingly, despite their diverse training regimes, both models prioritize similar image regions, potentially due to their shared dataset and resulting common inductive bias. In contrast, SAM (Kirillov et al., 2023) is trained on massive segmentation masks without semantic labels or object-centric priors, and MAE (He et al., 2022) is trained through inpainting of randomly masked image regions. Both methods emphasize local image features over high-level semantics. Our versatile visualization tool provides valuable insights into the characteristics of ViT models, offering substantial potential for practical applications.

In Figure 3 we offer additional visualization of ensembled SRT features across various network layers. The visualization indicates a noticeable trend: deeper layers reveal clearer high-level semantic boundaries, while shallower layers highlight more local features than deeper ones.

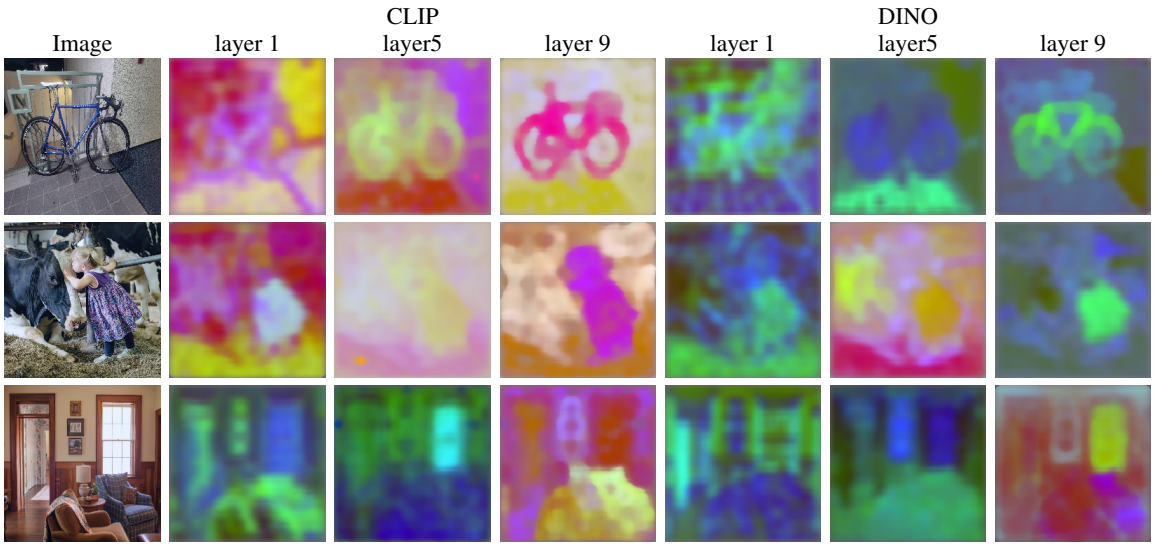

Figure 3. **Visualization of ensembled SRT features in different ViT layers.** *Architecture: ViT-S/16. Deeper layers reveal clearer high-level semantic boundaries, while shallower layers highlight more local features compared to high-level ones*

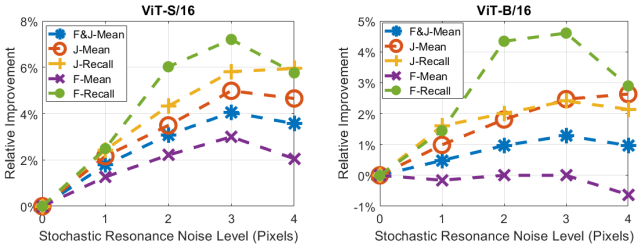

Figure 4. **Relative improvement on DAVIS-2017 dataset vs different noise levels.** *There exists an inherent trade-off between perturbation level and performance gain. Smaller perturbation ranges result in weaker improvements from the baseline model due to lower input diversity, while larger perturbations are susceptible to greater information loss. 3 pixels is found to be the optimal perturbation level on both ViT-S/16 and Vit-B/16.*

3.2. Semi-supervised Video Object Segmentation

We apply SRT to evaluate its performance using the DAVIS-2017 video instance segmentation benchmark (Pont-Tuset et al., 2017). We adhere to the experimental methodology established in (Jabri et al., 2020), which employs a "semi-supervised" video object segmentation approach on the original 480p resolution. Provided with the initial annotation of the objects of interest in the first frame, this method subsequently propagates the segmentation between consecutive frames. Notably, the method utilizes the last layer feature of the Vision Transformer (ViT) to guide this segmentation propagation process. Consequently, the quality of the ViT features directly impacts the final segmentation results. For optimal outcomes, these features must possess discriminative and semantically meaningful characteristics to effectively support this segmentation task.

In our study, we evaluate various Vision Transformer (ViT) models pre-trained using the DINO (Caron et al., 2021) contrastive scheme. We adopt three different architectures, specifically ViT-S/16, ViT-B/16, and ViT-S/8, each varying in their spatial patch size (16x16 pixels and 8x8 pixels). Our results in Tab. 1 indicate that, on average, SRT enhances the original baseline models by a relative 2.4% in terms of the F&J score. The most significant improvement is observed with ViT-S/16, where we achieve 4.1%. Importantly, these enhancements are achieved without any modifications to the model or pre-trained weights. However, we address a potential criticism of our approach, which could be seen as trivial test-time augmentation combined with feature-level ensemble. To counter this concern, we perform a heuristic by naively augmenting images by color jitter and performing feature-level ensemble (1, Naive ensemble), and we find that this method is, in fact, detrimental to performance. We also reproduce the approach proposed by (Amir et al., 2021) that uses overlapping tokens at inference time, which negatively impacts the results. We investigate whether inference costs induced by SRT can potentially be mitigated via distillation. To this end, we attempt to learn the ensembled SRT representations using the following self-distillation objective:

$$\min_{w} \sum_{x \in \mathcal{D}} ||\phi_w(x) - SRT(x, w_0)||, \tag{4}$$

where ϕ and (w_0) w are the ViT and its (original) parameters, and x the image in the target dataset. Our preliminary results on DINO-ViT/16 improve from the baseline by 1.3% after the self-distillation step. Note that Eq. (4) is task agnostic and requires no label, thus effectively adapts pre-

Method	F&J	J-mean	J-recall	F-mean	F-recall
DINO-ViT-S/16	0.617	0.602	0.740	0.634	0.764
+ SRT	**0.642**	**0.632**	**0.783**	**0.653**	**0.819**
Distill by SRT	0.625	0.609	0.745	0.642	0.780
+ Overlapping tokens	0.591	0.577	0.706	0.605	0.741
+ Naive ensemble	0.477	0.455	0.468	0.500	0.542
DINO-ViT-B/16	0.622	0.608	0.748	0.637	0.760
+ SRT	**0.630**	**0.623**	**0.766**	**0.637**	**0.795**
DINO-ViT-S/8	0.706	0.675	0.815	0.737	0.846
+ SRT	**0.720**	**0.688**	**0.827**	**0.752**	**0.868**

Table 1. **Results on DAVIS-2017 video object segmentation.** *Applying SRT improves over the baseline models uniformly over all metrics, as measured across 3 variants of ViTs trained using the DINO (Caron et al., 2021) contrastive learning objective. SRT yields significant improvements even for ViT-S/8 trained with finer patch sizes (8x8). One may optionally fine-tune the original ViT model by distilling by SRT, which increases performance while inference time and cost remain one single forward pass.*

trained ViT features to any given target dataset. We leave the investigation of this to future work.

Fig. 3.2 illustrates the relative improvement across different perturbation levels of SRT applied to ViT-S/16 and ViT-B/16. While higher perturbation levels offer greater input diversity, they are also susceptible to information loss. We anticipate a trade-off between perturbation level and performance gain and empirically identify a perturbation level of 3 pixels as the optimal point for both.

3.3. Depth Prediction and Semantic Segmentation

We extend the application of SRT to monocular depth estimation, a task that leverages ViT features from multiple ViT layers, in contrast to video object segmentation which primarily utilizes the last layer features. This choice of task highlights the versatility of SRT, showcasing its seamless compatibility with various ViT layers and architectures. Specifically, we evaluate three ViT architectures: ViT-S/14, ViT-B/14, and ViT-L/14, each equipped with two prediction heads (linear and DPT (Ranftl et al., 2021)). We adopt the experimental settings provided by DINOV2, which offers pre-trained backbones and corresponding prediction heads. Our assessment utilizes the NYU-V2 dataset (Nathan Silberman & Fergus, 2012) under its original 640×480 resolution.

Tab. 2 presents the results, demonstrating consistent improvements over baseline methods. The most significant enhancements are observed in the RMSE and RMSE_log metrics, where we achieve relative improvements of 4.7% and 14.9% with linear heads, and 3.6% and 11.0% with DPT heads, respectively. Notably, these metrics are sensitive to outliers, highlighting the effectiveness of SRT in mitigating instability in ViT features and enhancing robustness. On the other hand, SRT can reduce some of the bias introduced by training procedures that puts more focus on global representation (e.g. DINO). SRT re-introduces localized features through aggregating local perturbations of each region, leading to improvements in RMSE (often caused by lack of details, e.g., over-smoothing across object boundaries) while maintaining global coherence as measured by AbsRel.

For ablation, we compare our method with output-space ensemble (marked as "OE"), which employs the same perturbations as SRT, but aggregates the model output instead of intermediate features. We find no significant improvements, and in some cases, this method is even detrimental. This underscores the robustness of SRT's ensemble scheme that operates on the feature level instead of the output.

Similar to depth prediction, we show results on semantic segmentation by employing the protocol from DINOV2 and ADE20K dataset (Zhou et al., 2017). The results are presented in Tab. 3. SRT consistently improves mIoU on all three pre-trained ViTs, by as much as 1.7% in relative improvement, but is less significant than depth estimation. Depth comes in the form of positive real values; whereas, segmentations are represented as integer categories (logits quantized by choosing the max). Metrics measuring fidelity of segmentation operate on the categorical level, where spurious changes in logits do not affect the category they are mapped to so long as it is the max. Hence, while SRT can suppress spurious outputs in logits, the gain is less significant than depth where any difference in value is penalized.

3.4. Unsupervised Salient Region Segmentation

We employ SRT in conjunction with TokenCut (Wang et al., 2022) for unsupervised salient region segmentation tasks. TokenCut is a graph-based approach that applies the Normalized Cut algorithm to partition ViT tokens into two distinct clusters, representing the salient foreground and the background respectively. The key challenge is to ensure that the features are not only discriminative across clusters but also consistent within clusters. We adopt three datasets: ECSSD (Shi et al., 2015), DUTS (Wang et al., 2017), and DUT-OMRON (Yang et al., 2013), following the TokenCut.

In Tab. 4, we report results both before and after post-processing (bilateral solver) to assess both the raw quality of ViT embeddings and final segmentation accuracy. Under both settings, SRT improves the original ViTs pre-trained by DINO, with an average increase in the maxF metric of 1.8%. Notably, this improvement is constrained by the architecture of TokenCut, as it operates at the coarse segmentation level of ViT tokens. Directly applying TokenCut to the enhanced feature map is computationally impractical due to its $O(n^2)$ complexity in constructing a fully connected graph for graphcut. Given SRT's capability to provide fine-grained features, we anticipate future research on the effective lever-

Backbone	Head	Method	RMSE	RMSE_log	AbsRel	SqRel	a1	a2	a3
DINOV2-ViT-B/14	Linear	Baseline	0.396	0.135	0.100	0.061	0.903	0.983	0.996
		+OE	0.376	0.121	0.093	0.059	0.918	0.984	0.997
		+SRT	**0.349**	**0.108**	**0.087**	**0.052**	**0.930**	**0.990**	**0.998**
	DPT	Baseline	0.323	0.109	0.074	0.044	0.941	0.987	0.996
		+OE	0.314	0.101	**0.073**	**0.043**	0.944	0.988	**0.997**
		+SRT	**0.305**	**0.096**	**0.073**	**0.043**	**0.945**	**0.989**	**0.997**
DINOV2-ViT-S/14	Linear	Baseline	0.471	0.162	0.125	**0.084**	0.853	0.972	0.994
		+OE	0.486	0.153	0.126	0.095	0.858	0.974	0.994
		+SRT	**0.457**	**0.140**	**0.118**	0.085	**0.876**	**0.980**	**0.996**
	DPT	Baseline	0.336	0.114	0.080	**0.048**	0.933	0.986	0.996
		+OE	0.347	0.114	0.080	0.053	0.932	0.985	0.996
		+SRT	**0.334**	**0.104**	0.080	0.051	**0.935**	**0.988**	0.996
DINOV2-ViT-L/14	Linear	Baseline	0.373	0.127	0.093	0.054	0.916	0.985	0.996
		+OE	0.401	0.131	0.097	0.062	0.908	0.982	0.996
		+SRT	**0.365**	**0.113**	**0.090**	**0.053**	**0.924**	**0.989**	**0.998**
	DPT	Baseline	0.311	0.105	**0.070**	0.042	0.946	0.988	**0.997**
		+OE	0.317	0.103	0.072	0.044	0.942	0.987	0.996
		+SRT	**0.297**	**0.092**	**0.070**	**0.041**	**0.947**	**0.991**	**0.997**

Table 2. **Results on NYU-V2 depth prediction.** *Our method can be extended without modification to improve intermediate features to yield improved performance on the downstream depth prediction tasks. While ensembling of outputs (OE) can often be detrimental to performance, applying SRT on the features from pre-trained backbones (inputs to prediction heads) can improve performance over baselines by* 4.7% *and* 14.9% *on RMSE and RMSE_log, using the linear prediction head and by* 3.6% *and* 11.0% *using the DPT head.*

Method	head	baseline	d=1	d=2	d=3
ViT-S/14	linear	44.24	44.44	44.57	**44.64**
ViT-B/14	linear	47.28	47.63	47.85	**47.98**
ViT-L/14	linear	47.79	48.18	48.44	**48.62**

Table 3. **Results on Semantic Segmentation on ADE20K in mIOU** *Experiments run with evaluation pipeline from InternImage (Wang et al., 2023) and DINOV2 (Oquab et al., 2023). d denotes the translation in pixels, ranging from -d to d with respect to a coordinate location across horizontal and vertical directions, when ensembling with SRT. As the size of the ensemble grows, the segmentation mIOU increases. d: perturbation level.*

age of SRT's high-resolution embeddings.

3.5. Image Retrieval and Unsupervised Object Detection

Incorporating SRT into vision tasks involves updating ViT features based on fine-tuned high-resolution features. However, questions remain regarding whether the observed enhancements in dense prediction tasks are solely due to increased awareness of semantic boundaries in images and whether this method extends to non-dense prediction tasks. To address these concerns, we conducted a sanity check using image retrieval and unsupervised object detection tasks.

For image retrieval, we applied a nearest-neighbor protocol following DINO, using the Oxford image retrieval datasets (Radenović et al., 2018) and ViT-S/16 trained on ImageNet. Notably, our base model's pre-training poses a substantial domain gap to the target datasets. Note that, we do not naively average the class tokens from augmented images, but ensemble the features by SRT prior to the attention mechanism in the last layer. In this way, the final class token is computed from the ensemble SRT feature. Although image retrieval primarily requires distinctive image-level features (rather than pixel-level), aiming to match images to queries at a higher level, SRT exhibited effective adaptation, resulting in a notable 2.6% relative improvement.

Regarding unsupervised object detection, we utilized TokenCut and the VOC07 dataset (Everingham et al., 2010). Unsupervised object detection focuses on region-level discriminative features, utilizing bounding boxes instead of segmentation masks for object shapes. Despite this, we observed a 1.0% relative improvement in the detection rate, reaffirming that SRT does not compromise the information within the original ViT embeddings. These results serve as a critical validation of SRT's capacity to obtain fine-grained ViT features without distorting their original information.

4. Related Work

Stochastic Resonance was proposed by (Benzi et al., 1981) and first applied in climate dynamics (Benzi et al., 1982) and later in signal processing (Wellens et al., 2003; Kosko & Mitaim, 2001; Chen et al., 2007) and acoustics (Shu-Yao et al., 2016; Wang et al., 2014). It is used to enhance a

Datasets	ECSSD			DUTS			DUTS-OMRON		
Feature Extractor	maxF	IoU	Acc.	maxF	IoU	Acc.	maxF	IoU	Acc.
DINO ViT-S/16	80.3	71.2	91.8	67.2	57.6	90.3	60.0	53.3	88.0
+SRT	**82.4**	**71.7**	**92.1**	**68.8**	**58.5**	**90.7**	**61.0**	**54.0**	**88.2**
DINO ViT-S/16 w/ bilateral solver	87.4	**77.2**	93.4	75.5	**62.4**	91.4	69.7	61.8	89.7
+SRT	**88.4**	77.0	**93.6**	**76.5**	**62.4**	**91.7**	**70.6**	**62.4**	**89.9**
DINO ViT-B/16	80.3	71.0	91.5	66.4	56.7	89.5	56.7	50.5	85.4
+ SRT	**81.8**	**72.6**	**92.2**	**68.8**	**58.3**	**90.6**	**58.0**	**51.6**	**86.1**
DINO ViT-B/16 w/ bilateral solver	86.8	76.6	93.0	74.1	60.9	90.6	65.6	58.4	87.1
+ SRT	**88.2**	**78.0**	**93.7**	**68.8**	**58.3**	**90.6**	**67.2**	**59.7**	**87.8**

Table 4. **Results on unsupervised salient region segmentation.** *Despite architectural constraints, our method yields consistent improvement on all three datasets, with an average increase of 1.8% in the maxF metric.*

Task	Metric	Baseline	d=1	d=2	d=3	d=4	d=5	d=6
Image Retrieval	mAP (Medium)	34.6	34.8	35.1	35.2	35.3	35.3	**35.5**
	mAP (Hard)	13.0	13.1	13.2	13.1	13.2	**13.2**	13.1
Object Discovery	Detection Rate	68.7	68.9	68.9	69.2	**69.4**	69.3	69.2

Table 5. **Results on Image Retrieval and Object Discovery.** *SRT generalizes to non-dense prediction tasks operating on higher-level region/image features to yield equal or better performance compared to the standard inference baseline. On the Oxford image retrieval task, SRT on the DINO-ViT-S/16 model yields up to* 2.6% *relative improvement from the baseline model. On the unsupervised object detection task, SRT improves the detection rate by up to* 1.0%*. d: translation in pixels when ensembling with SRT. d: perturbation level.*

signal beyond the native resolution of the sensor by adding white noise. We use the same principle to adapt generic ViT image features for dense prediction downstream tasks. By randomly translating the images, (i.e. introducing noise in the spatial dimension), we can enhance ViT features to be smoother and better suited for dense prediction tasks. We leave extensions to other groups or semi-groups of transformations (*e.g.*, scale or domain size) to future work.

Test-time data augmentation involves aggregating model predictions from augmented test input to a final prediction. Applying such a technique increases the robustness of predictions (Prakash et al., 2018; Song et al., 2017; Cohen et al., 2019) and prediction accuracy (Krizhevsky et al., 2012; Szegedy et al., 2015; Simonyan & Zisserman, 2014; Jin et al., 2018; Matsunaga et al., 2017) in a variety of tasks. It can be used during adaptation of model weights at test-time (Park et al., 2024). It can also used to estimate the uncertainty of the model (Matsunaga et al., 2017; Smith & Gal, 2018; Ayhan & Berens, 2022; Wang et al., 2019). Different transformations are used to target different potential tasks: (Pang et al., 2019) linearly combines the testing input and a randomly sampled clean image to generate classification prediction. (Isensee et al., 2018) performs flipping and rotation to the test input image to generate 64 different inputs and finally aggregates the outputs to perform medical image segmentation. (Krizhevsky et al., 2012) crops the images into smaller patches and ensemble the results for classification. Self-ensembling (Bousselham et al., 2021) is also closely related to our work. (Bousselham et al., 2021) leverages multi-scale features fed into multiple independent decoders to create an ensemble within a single model. (Liu et al., 2018) ensembles outputs from networks augmented with random noise layers to improve model robustness. SRT aggregates information via adding spatial translations as noise and can be considered a general case of test-time augmentation, where ensembling is performed at the feature level at intermediate layers of a ViT, instead of the output level, which is novel.

Knowledge distillation aims to transfer the knowledge from stronger teacher models to weaker student models to improve their performance. (Hinton et al., 2015) trains a student model to mimic the soft output distribution of the teacher model. (Romero et al., 2014) extends this idea to distill the intermediate features learned by the teacher models. We consider a form of self-distillation (Zhang et al., 2019), in which the student itself is used as the teacher to improve learned representations.

Dense ViT feature extractor. Our work is closely related to (Amir et al., 2021),which employs ViT for generating dense visual descriptors. To extract these fine-grained features, (Amir et al., 2021) reduce the stride allowing for overlapping tokens and performing a single forward pass with ViT. In SRT, instead of a single pass, we conduct multiple passes using perturbed inputs. This modification reduces the computational complexity from quadratic to linear.

Properties discovered by SRT. Additionally, our findings underscore the segmentation capabilities of ViT embed-

	Method	RMSE	R_log	AbsRel	SqRel	a1	a2	a3
ViT-S	Baseline	0.336	0.114	0.080	0.048	0.933	0.986	0.996
	Bilinear	0.573	0.178	0.146	0.125	0.8	0.964	0.995
	Bicubic	0.572	0.124	0.146	0.178	0.801	0.964	0.995
ViT-B	Baseline	0.323	0.109	0.074	0.044	0.941	0.987	0.996
	Bilinear	0.568	0.185	0.146	0.120	0.792	0.96	0.992
	Bicubic	0.579	0.188	0.149	0.124	0.787	0.959	0.991
ViT-L	Baseline	0.311	0.105	0.070	0.042	0.946	0.988	0.997
	Bilinear	0.732	0.246	0.183	0.195	0.695	0.91	0.973
	Bicubic	0.720	0.241	0.181	0.188	0.701	0.916	0.975

Table 6. **Results on NYU-V2 depth prediction using interpolated features** *For the interpolation method, we bilinearly or bicubically interpolate the DINOV2 feature up to the image dimension and perform an average pooling to return the feature to the original dimension. The results are much worse than the baseline method, which uses the original DINOV2 features without ensembling.*

Architecture	ResNet20	ResNet32	ResNet56
Accuracy	91.95	92.68	93.50
Accuracy w/ SRT	**92.41**	**93.14**	**93.87**
Relative error reduced	5.6%	6.3%	5.7%

Table 7. **Results on Cifar-10 classification with ResNet.** *Stochastic resonance consistently improves classification accuracy by an average of 5.87% and as much as 6.3% on ResNet32, without additional training.*

dings, aligning with recent claims in the field (Caron et al., 2021; Yu et al., 2023). Enhanced features exhibit sharp, fine-grained semantically relevant boundaries. Furthermore, SRT method leverages the convexity properties (Park & Kim, 2022) of ViT embeddings, enabling convex combinations (average pooling as a special case) during inference. SRT is also related to (Darcet et al., 2024), which shows local ViT tokens may be repurposed, possibly incurring decreased performance on dense prediction tasks. SRT resolves this issue by obtaining "localized" features by ensembling.

5. Discussion

Ensemble vs super-resolution. Although both increase the spatial resolution, SRT achieves it by ensemble, which differs from super-resolution: Given a signal x that is subsampled to $\tilde{x}$, super-resolution aims to retrieve an approximation $\hat{x}$ of x given $\tilde{x}$. Since information is lost in the sampling, the reconstruction depends crucially on the choice of prior. In this sense, super-resolution is a form of hallucination: Attribute details to $\hat{x}$ that are not in $\tilde{x}$, in the hope that they will somehow match those in x. This requires strong faith in prior knowledge about x, $P(x)$.

Given the same signal x, one could instead generate multiple samples $\tilde{x}_i$, each with a different kernel, and then reconstruct a single estimate from the samples $\hat{x} = F(\tilde{x}_1, \dots, \tilde{x}_N)$. Now, the estimator F has more information available about x than in super-resolution: The sigma-algebra spanned by the random variables $\tilde{x}_i$ is a superset of the (trivial) sigma algebra spanned by the single sample $\tilde{x}$ in super-resolution. In other words, SRT which aggregates different samples from a process contains more information than any single sample about the process. To validate this claim, we provide further ablation studies on comparing with enhancing the features by spatial interpolation and smoothing in Table 6 on NYU-V2 dataset with DINOV2 backbone and DPT head, showing that simple feature interpolation is detrimental to the task.

Limitations. SRT has several limitations. The basic embodiment increases inference cost and latency, as each perturbed image necessitates a ViT forward pass. To address this, one viable approach is knowledge distillation, which involves fine-tuning the network to mimic the feature-level output of SRT. We illustrate this process using the DAVIS-2017 training dataset with DINO-ViT-S/16, achieving improved results (F&J-score $0.617 \Rightarrow 0.625$) without the use of labels or operations on the validation set. This establishes a label-free, task-free transductive fine-tuning scheme that adapts pre-trained ViT features to new target datasets. Future directions may involve refining the distillation process on different layers and exploring the integration of Stochastic Resonance directly into ViT architectures.

Conclusions. SRT offers a versatile feature-level ensemble method that applies to any layer within any network architecture that utilizes ViT as a feature extractor, eliminating the need for modifications to the model's forward pass, in contrast to most test-time augmentation and ensemble methods that operate at the output level that require task-specific designs. Compared with increasing token numbers, SRT avoids the quadratic complexity related to the number of ViT tokens, and is amenable to parallelization through batching, ensuring computational efficiency. Furthermore, the method allows ensembling without memory-intensive resizing of all embeddings to full resolution, which can be executed recursively, as described in Section 2.3. Practical implementations demonstrate smooth execution on even laptop GPUs, confirming the efficiency of our approach.

It is worth noting that stochastic resonance is not limited to ViT architectures, as demonstrated in Table 7, where we apply the same stochastic resonance mechanism to ResNet (He et al., 2016) on the image classification task, and on average reduces error by a relative 5.87%. Stochastic resonance also applies to other forms of quantization, such as sale or domain size. However, our emphasis in this paper is on ViTs that mostly use non-overlapping tokens, making them particularly suited to our approach.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgements

This work was supported by ONR N00014-22-1-2252 and ARO W911NF-17-1-0304.

References

Amir, S., Gandelsman, Y., Bagon, S., and Dekel, T. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021.

Ayhan, M. S. and Berens, P. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *Medical Imaging with Deep Learning*, 2022.

Benzi, R., Sutera, A., and Vulpiani, A. The mechanism of stochastic resonance. *Journal of Physics A: mathematical and general*, 14(11):L453, 1981.

Benzi, R., Parisi, G., Sutera, A., and Vulpiani, A. Stochastic resonance in climatic change. *Tellus*, 34(1):10–16, 1982.

Bousselham, W., Thibault, G., Pagano, L., Machireddy, A., Gray, J., Chang, Y. H., and Song, X. Efficient self-ensemble for semantic segmentation. *arXiv preprint arXiv:2111.13280*, 2021.

Buades, A., Coll, B., and Morel, J.-M. A non-local algorithm for image denoising. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 2, pp. 60–65. Ieee, 2005.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Chen, H., Varshney, P. K., Kay, S. M., and Michels, J. H. Theory of the stochastic resonance effect in signal detection: Part i—fixed detectors. *IEEE transactions on Signal Processing*, 55(7):3172–3184, 2007.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pp. 1310–1320. PMLR, 2019.

Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024.

Dong, J. and Soatto, S. Domain-size pooling in local descriptors: Dsp-sift. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5097–5106, 2015.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338, 2010.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P. F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.

Jabri, A., Owens, A., and Efros, A. Space-time correspondence as a contrastive random walk. *Advances in neural information processing systems*, 33:19545–19560, 2020.

Jin, H., Li, Z., Tong, R., and Lin, L. A deep 3d residual cnn for false-positive reduction in pulmonary nodule detection. *Medical physics*, 45(5):2097–2107, 2018.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Kosko, B. and Mitaim, S. Robust stochastic resonance: Signal detection and adaptation in impulsive noise. *Physical review E*, 64(5):051110, 2001.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Liu, X., Cheng, M., Zhang, H., and Hsieh, C.-J. Towards robust neural networks via random self-ensemble. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 369–385, 2018.

Matsunaga, K., Hamada, A., Minagawa, A., and Koga, H. Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. *arXiv preprint arXiv:1703.03108*, 2017.

Nathan Silberman, Derek Hoiem, P. K. and Fergus, R. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Pang, T., Xu, K., and Zhu, J. Mixup inference: Better exploiting mixup to defend adversarial attacks. *arXiv preprint arXiv:1909.11515*, 2019.

Park, H., Gupta, A., and Wong, A. Test-time adaptation for depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Park, N. and Kim, S. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022.

Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., and Van Gool, L. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017.

Prakash, A., Moran, N., Garber, S., DiLillo, A., and Storer, J. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8571–8580, 2018.

Radenović, F., Iscen, A., Tolias, G., Avrithis, Y., and Chum, O. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5706–5715, 2018.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Ranftl, R., Bochkovskiy, A., and Koltun, V. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12179–12188, 2021.

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Shi, J., Yan, Q., Xu, L., and Jia, J. Hierarchical image saliency detection on extended cssd. *IEEE transactions on pattern analysis and machine intelligence*, 38(4):717–729, 2015.

Shu-Yao, J., Fei, Y., Ke-Yu, C., and En, C. Application of stochastic resonance technology in underwater acoustic weak signal detection. In *OCEANS 2016-Shanghai*, pp. 1–5. IEEE, 2016.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Smith, L. and Gal, Y. Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*, 2018.

Song, Y., Kim, T., Nowozin, S., Ermon, S., and Kushman, N. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *arXiv preprint arXiv:1710.10766*, 2017.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., and Vercauteren, T. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

Wang, J., He, Q., and Kong, F. Adaptive multiscale noise tuning stochastic resonance for health diagnosis of rolling element bearings. *IEEE Transactions on instrumentation and measurement*, 64(2):564–577, 2014.

Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., and Ruan, X. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 136–145, 2017.

Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., Wang, X., and Qiao, Y. Internimage: Exploring large-scale vision foundation models

with deformable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14408–14419, June 2023.

Wang, Y., Shen, X., Yuan, Y., Du, Y., Li, M., Hu, S. X., Crowley, J. L., and Vaufreydaz, D. Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. *arXiv preprint arXiv:2209.00383*, 2022.

Wellens, T., Shatokhin, V., and Buchleitner, A. Stochastic resonance. *Reports on progress in physics*, 67(1):45, 2003.

Wu, Y., Liu, T. Y., Park, H., Soatto, S., Lao, D., and Wong, A. Augundo: Scaling up augmentations for unsupervised depth completion. *arXiv preprint arXiv:2310.09739*, 2023.

Yang, C., Zhang, L., Lu, H., Ruan, X., and Yang, M.-H. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3166–3173, 2013.

Yu, Y., Chu, T., Tong, S., Wu, Z., Pai, D., Buchanan, S., and Ma, Y. Emergence of segmentation with minimalistic white-box transformers. *arXiv preprint arXiv:2308.16271*, 2023.

Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., and Ma, K. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3713–3722, 2019.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.