Beyond Predictive Resampling: Learning Input-Agnostic Downsampling for Aligned Vision Recognition

Kai Zhao[1], Liting Ruan[1], Haoran Jiang[1], Xiaoqiang Zhu[1], Xianchao Zhang[2] , Dan Zeng[1]*

[1] School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China
[2] Provincial Key Laboratory of Multimodal Perceiving and Intelligent Systems, Jiaxing University, Jiaxing 314001, China
kz@kaizhao.net, dzeng@shu.edu.cn

Abstract

Images are typically sampled on a uniform grid, despite their non-uniform information distribution—some regions are rich in content while others are not. The mismatch leads to inefficient computation allocation in deep learning models. To address this, recent studies have proposed predictive downsampling methods that adaptively downsample images based on predicted per-pixel importance, allocating more pixels to informative areas. However, these methods require high-resolution processing to accurately estimate importance, which undermines their efficiency: the prediction itself must process the full-resolution image, consuming most of the computational budget. This high-resolution importance prediction is necessary because each input may differ significantly in structure and content. In this paper, we take a different approach and introduce a learn-to-downsample paradigm tailored for aligned vision recognition tasks, such as face recognition and palmprint recognition, where input alignment ensures consistent spatial structure across images. This alignment ensures structural consistency across images, allowing a shared, input-agnostic downsampling template applicable to all inputs. Furthermore, instead of relying on implicit importance maps, we introduce a flow-based representation that explicitly models the spatial warping from the original image to the downsampled version. The flow representation is not only more efficient but also more controllable: we regularize the flow using its Jacobian determinant to precisely control the sampling density and coverage, enabling interpretable and tunable sampling patterns. Extensive experiments on two aligned recognition tasks, face and palmprint recognition, demonstrate that our method substantially reduces computational cost with minimal accuracy degradation, achieving a significantly better performance-efficiency trade-off than existing predictive downsampling methods.

Introduction

Images are typically uniformly sampled in the spatial domain, leading to a rectangular grid of pixels. Modern Deep Learning (DL) architecutres, *i.e*. Convolutional Neural Networks (CNNs) (Krizhevsky, Sutskever, and Hinton 2012; He et al. 2016) and Vision Transformers (ViTs) (Dosovitskiy et al. 2021), are designed to process such uniformly sampled pixels, and the computation is evenly allocated across the image. However, visual information is often not uniformly distributed across the image, with some regions containing more important features than others (Itti, Koch, and Niebur 2002; Bruce and Tsotsos 2009). The mismatch between uniformly sampled pixels and the non-uniform nature of visual information leads to inefficient computation (Fayyaz et al. 2022), as the model allocates equal resources to both textureless regions and those rich in visual detail.

Recently, adaptive image sampling has attracted growing interest due to its ability to adjust sampling density dynamically based on image content (Recasens et al. 2018; Jin et al. 2022). By assigning denser sampling to informative regions and sparser sampling to less relevant areas, it produces downsampled images of lower spatial resolution while preserving critical visual details. This enables faster inference and lower memory consumption without significantly compromising accuracy.

However, most existing methods follow the predictive sampling paradigm, where the sampling density is determined by a sub-network conditioned on the input image. Density prediction is necessary because the input images present diverse structural content and therefore require different sampling patterns. To make accurate predictions, the sub-network must process the full-resolution image, which introduces significant computational overhead and consumes a large portion of the total budget.

In most vision tasks, input images exhibit diverse and unpredictable structures. A notable exception is aligned vision recognition, where inputs are geometrically aligned and share similar structural content across samples. *Face recognition* and *palmprint recognition* are two representative examples of aligned vision recognition. In these tasks, the input images are typically aligned to a canonical pose (Zhang et al. 2016; Zhao et al. 2022), to ensure the recognition model focuses on the discriminative features rather than the geometric variations. The consistent input structure in aligned vision recognition presents a unique opportunity: *can we learn an input-independent sampling pattern that generalizes across all inputs?* Such a pattern would enable downsampling without accessing the full-resolution input, thereby eliminating the computational overhead associated with image-dependent processing.

In this paper, we move beyond the conventional predictive sampling paradigm and introduce a *learned sampling*

*Corresponding Author.

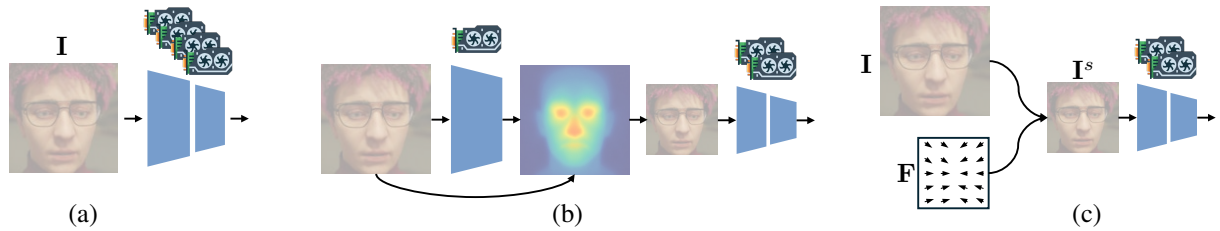

Figure 1: Diagramatic comparison of (a) full image processing, (b) predictive sampling, and (c) learned sampling. Full image processing requires the model to process the entire input image **I**, which is computationally expensive. Predictive sampling predicts a sampling pattern based on the input image, but still requires processing the full-resolution image to estimate the sampling density. Learned sampling uses a flow field **F** as a model parameter, which guides the downsampling process without requiring access to the input image.

paradigm tailored for aligned vision recognition. Our key insight is to exploit the consistent structural patterns present in aligned images to learn a universal, input-independent sampling strategy. Specifically, we learn a *flow field* to guide downsampling without requiring access to the input at all. This flow field is treated as a model parameter, independent of the input, allowing efficient downsampling at the very beginning of the network and greatly reducing computation for subsequent layers. Fig 1 illustrates the difference between predictive sampling and our proposed learned sampling paradigm.

Our method models downsampling as flow-guided image warping, which explicitly maps the high-resolution input to a compact representation. The sampling pattern is controlled via the Jacobian determinant of the flow, enabling flexible and spatially aware resampling. We introduce Jacobian-based regularization to ensure valid, topology-preserving transformations and to promote compression in less informative regions. In summary, our contributions are:

- We propose a learned sampling paradigm for aligned vision recognition, which eliminates the need for input-dependent processing and enables efficient downsampling.
- We introduce a flow-based representation to model the downsampling transformation, allowing more efficient sampling and flexible control over the sampling pattern.
- We design a set of regularization terms based on the Jacobian determinant to ensure valid transformations and to encourage compression of sampled regions.

We evaluate our method on two aligned vision recognition tasks: *face recognition* and *palmprint recognition*. Extensive experiments demonstrate that our approach greatly reduces computational cost while maintaining competitive recognition accuracy. Compared to predictive sampling methods, our method achieves a better trade-off between performance and efficiency. These results suggest a new paradigm for efficient image processing in aligned vision tasks, offering a principled and effective alternative to dynamic, prediction-based downsampling approaches.

Related Work

Learning to spatially resample or zoom into salient regions of visual data has long been studied as a means to improve efficiency and accuracy. Early attention models used dynamic glimpses to focus on informative parts of an image (Mnih et al. 2014), demonstrating that not all input pixels need equal processing. Spatial Transformer Networks (Jaderberg et al. 2015) introduced a learnable module that can apply global parametric transformations (e.g. cropping, scaling, rotations) to input features, allowing a network to automatically align or magnify important content.

Subsequent approaches explored differentiable resampling strategies: Recasens et al. (Recasens et al. 2018) proposed a saliency-based "Learning to Zoom" layer that learns non-uniform downsampling, producing a distorted yet task-enhancing image where informative regions are magnified. Jin et al. (Jin et al. 2022) developed a learnable downsampling module for ultra-high resolution segmentation, which adaptively allocates higher sampling density to more complex or uncertain image regions (e.g. around object boundaries) based on a low-res preview, yielding better segmentation than uniform resizing. These methods perform input-dependent resampling, often guided by saliency or learned attention maps, to preserve critical details while reducing resolution.

Similarly, adaptive focus mechanisms within network architectures have been proposed. For example, Zhao et al. (Zhao et al. 2025) present a "Boltzmann attention" scheme in a transformer that dynamically narrows the attention field: initially broad attention allows exploration, then an annealing process focuses it on likely object locations, greatly improving small-object detection efficiency. Other works in this vein include deformable convolutions (Dai et al. 2017), which learn input-dependent sampling offsets for each convolution kernel, effectively attending to salient local features.

The idea of allocating computation dynamically has also been explored extensively in model architectures. In image recognition, adaptive computation frameworks like MSDNet (Huang et al. 2018) use multi-scale features with early-exit classifiers to allow "easier" examples to be processed at lower cost, while harder ones receive deeper analysis. Skip-

Net (Wang et al. 2018) and BlockDrop (Wu et al. 2018) learn to skip unnecessary ResNet blocks on a per-input basis via learned gating networks (sometimes using reinforcement learning), reducing inference cost without sacrificing accuracy. Similarly, Veit et al. (Veit and Belongie 2018) introduce ConvNet-AIG, which inserts lightweight gates into a CNN to conditionally bypass certain layers, effectively learning a dynamic inference graph for each input. In natural language processing, where model depths are even greater, recent studies apply analogous ideas: Jiang et al. (Jiang et al. 2024) (D-LLM) add a decision module at each Transformer layer to determine if that layer can be skipped for a given token or sequence, significantly accelerating large language model inference.

Rather than predicting a different zooming per image, our method learns a single sampling pattern that consistently captures the most informative regions across all inputs—providing a fixed, low-cost attention mechanism tailored to align vision recognition.

Methodology

In this section, we present our proposed method for adaptive image sampling for aligned vision recognition.

Overview

Our method learns a spatially adaptive flow field to resample the input image via image warping. This process enables downsampling with unevenly spaced sampling points and preserves critical visual content. Instead of relying on fixed uniform sampling grids, we use a learnable flow field to guide where pixels are sampled.

Let $\mathbf{I} \in \mathbb{R}^{h \times w \times 3}$ denote the input image, where h and w are the height and width. Let $0 < s < 1$ be a user-defined downsampling factor. We aim to generate a smaller image $\mathbf{I}^s \in \mathbb{R}^{h' \times w' \times 3}$, where $h' = \lceil h \cdot s \rceil$ and $w' = \lceil w \cdot s \rceil$. To resample the image, we introduce a learnable flow field $\mathbf{F} \in \mathbb{R}^{h' \times w' \times 2}$, which defines a 2D offset (or displacement) vector at each location in the downsampled image $\mathbf{I}^s$. The flow field is input-independent and is a parameter of the neural network.

Image Downsampling via Flow-guided Warping

Each pixel in the downsampled image $\mathbf{I}^s(u, v)$ is computed by warping the original image $\mathbf{I}$ using a learned flow vector $\vec{\mathbf{f}}_{u,v} = (f^x_{u,v}, f^y_{u,v})$:

$$\mathbf{I}^s(u, v) = \mathbf{I}\big(u/s + f^x_{u,v},\ v/s + f^y_{u,v}\big), \tag{1}$$

where the sampling location is generally non-integer and evaluated via bilinear interpolation. When $\vec{\mathbf{f}}_{u,v} = (0, 0)$ for all (u, v), Eq (1) reduces to standard uniform downsampling by a factor of s.

This formulation allows the model to learn non-uniform sampling patterns that are adaptive to the task. Fig 2 illustrates how sampling locations are derived from the flow field $\vec{\mathbf{f}}$, using a 3×3 input image downsampled to a 2×2 output with a downsampling factor of $s = 2/3$.

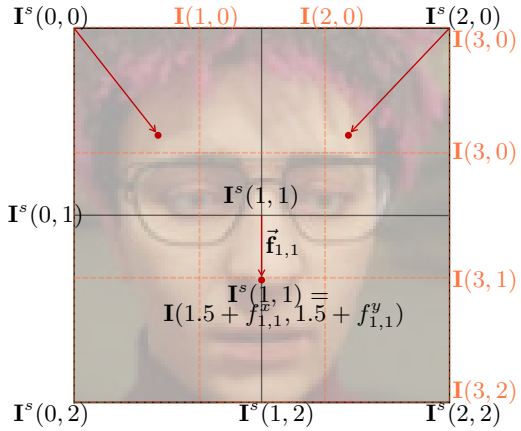

Figure 2: Illustration of location sampling when warping a 3×3 image (yellow dotted grid) to a 2×2 image (black solid grid) with a downsampling factor of $s = 2/3$. The flow vector $\vec{\mathbf{f}}$ (red arrows $\rightarrow$) guides the sampling locations in the original image $\mathbf{I}$.

Regularization through Jacobian Determinant

To prevent the learned flow field $\mathbf{F}$ from producing erratic or unstable sampling patterns, we introduce a set of regularization terms that promote spatial coherence and geometric validity. Specifically, we encourage $\mathbf{F}$ to be smooth, globally contractive, and diffeomorphic, ensuring that the resulting warping is both stable and topology-preserving.

Diffeomorphism A diffeomorphism is a smooth and invertible mapping whose inverse is also smooth; that is, it preserves the topology of the space without any folding, tearing, or inversion. In our context, this implies that the flow field should maintain the spatial ordering of the sampling grid and prevent any overlaps or pixel misalignments.

To ensure that the flow field $\mathbf{F}$ is diffeomorphic and does not induce local folding or inversion, we incorporate the Jacobian determinant[1] as a regularization term. The Jacobian determinant $J \in \mathbb{R}^{h' \times w'}$ of the flow field $\mathbf{F}$ is defined as:

$$J = \det\left(\mathbf{I} + \nabla \mathbf{F}\right), \tag{2}$$

where $\mathbf{I}$ is the identity matrix and $\nabla \mathbf{F}$ is the Jacobian matrix of the flow field. Each element of J measures the local area distortion caused by the flow field:

- $J[i, j] > 1$ indicates expansion of the local area,
- $0 < J[i, j] < 1$ indicates shrinkage,
- $J[i, j] = 0$ indicates a fold or tear,
- $J[i, j] < 0$ indicates an inversion.

For example, if the flow $\mathbf{F} = 0$, then $J_{i,j} = 1$ for all (i, j), indicating no local deformation — that is, the transformation preserves both area and spatial structure everywhere.

[1] https://en.wikipedia.org/wiki/Jacobian_matrix_and_determinant

To prevent local folds or tears, we penalize all negative values in J:

$$\mathcal{L}_{\text{diffeo}} = \sum_{i,j} \max(0, -J_{i,j}), \tag{3}$$

which encourages the flow field to be diffeomorphic.

Shrinkage In addition to enforcing local regularity via individual Jacobian determinants, we also penalize the global average of J:

$$\mathcal{L}_{\text{shrink}} = \sum_{i,j} J_{i,j}. \tag{4}$$

The average Jacobian in Eq (4) quantifies the net area change induced by the flow field, measuring whether the overall transformation is expanding or contracting the sampling grid.

Since the input images often contain large textureless regions— particularly in the background—this regularization encourages the flow to globally shrink the sampling area, effectively skipping uninformative regions.

By minimizing Eq (4), we promote flow fields that focus sampling capacity on informative content while suppressing the allocation of samples to unimportant areas.

Smoothness To ensure that the learned sampling grid is spatially coherent, we introduce a smoothness regularization term on the flow field. Specifically, we penalize both the first-order gradients and second-order derivatives (Laplacians) of the flow:

$$\mathcal{L}_{\text{smooth}} = \sum_{i,j} \|\nabla \vec{\mathbf{f}}_{i,j}\|_2^2 + \sum_{i,j} \|\Delta \vec{\mathbf{f}}_{i,j}\|_2^2, \tag{5}$$

where $\nabla \vec{\mathbf{f}}_{i,j}$ denotes the spatial gradient at location (i, j) and $\Delta \vec{\mathbf{f}}_{i,j}$ denotes the Laplacian.

End-to-End Training

The sampling operation in Eq (1) and the regularization in Eqs (3) to (5) are fully differentiable with respect to the flow field $\mathbf{F}$. Thus, we can jointly train the flow module and the recognition network via backpropagation.

The overall training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{smooth}} + \mathcal{L}_{\text{diffeo}} + \mathcal{L}_{\text{shrink}}, \tag{6}$$

where $\mathcal{L}_{\text{rec}}$ is the task-specific recognition loss, e.g., cross-entropy for classification. This allows the flow field to adaptively learn task-relevant sampling patterns from data.

After training, the learned flow field $\mathbf{F}$ can be used to downsample any aligned input image.

Experiments

Experimental Setup

We evaluate our method on two aligned vision recognition tasks: *face recognition* and *palmprint recognition*. In both tasks, the models receive images that are geometrically aligned to a canonical pose based on detected landmarks.

We compared our method against state-of-the-art adaptive sampling methods, including: (Recasens et al. 2018), (Jin et al. 2022), and (Zhao et al. 2025). We also compared against the baseline of uniform downsampling with bilinear interpolation. For all downsampling methods, we apply the same downsampling procedure to both training and test images, ensuring they are processed to the same resolution.

Running Efficiency

Before quantitative evaluation, we first compare the running efficiency of different adaptive sampling methods. In particular, we compare our method with the baseline of uniform downsampling and two predictive downsampling methods: (Recasens et al. 2018) and (Jin et al. 2022).

We use a baseline input image size of 112×112, and downsample the input to 56×56 and 84×84, 96×96 respectively. The efficiency comparison does not rely on the specific recognition task.

	112	96	84	56
Uniform	5,525	6,675	7,322	11,969
(Recasens et al. 2018)		5,817	5,973	8,539
(Jin et al. 2022)		5,817	5,973	8,539
Ours		6,623	7,3,09	11,849

Table 1: FPS of different downsampling methods on a single NVIDIA RTX 3090 GPU.

As demonstrated in Tab 1, our method achieves FPS on par with the baseline of uniform downsampling, while being notably faster than the other two adaptive sampling methods, thanks to the elimination of importance prediction and the input-agnostic nature of our approach. (Recasens et al. 2018) and (Jin et al. 2022) require additional computation to predict the importance of each pixel, which slows down the process. Since the two methods share the same computation, and the only difference is the additional loss term in (Jin et al. 2022), the FPS of the two methods are similar.

Face Recognition

Experimental Setup For face recognition, we use the MS1MV2 (Deng et al. 2022) dataset as the training set. We train various face recognition models and evaluate their performance with different downsampling methods.

Following state-of-the-art methods (Li et al. 2021a; Meng et al. 2021; Li et al. 2021b; Kim, Jain, and Liu 2022), we report the face verification accuracy on seven benchmarks: LFW (Huang et al. 2008), CFP-FP (Sengupta et al. 2016), CPLFW (Zheng and Deng 2018), AgeDB (Moschoglou et al. 2017), CALFW (Zheng, Deng, and Hu 2017), IJB-B and IJB-C (Whitelam et al. 2017). LFW contains 6,000 pairs of in-the-wild face images. CFP-FP and CPLFW have larger pose variation (CFP-FP specifically compares frontal views to profile views). AgeDB and CALFW have larger age variation.

The IJB-B dataset contains 1,845 subjects with 21.8K still images and 55K frames from 7,011 videos. As the extension of IJB-B, the IJB-C dataset cov- ers about 3,500 identities with a total of 31,334 images and 117,542 unconstrained video frames. In the 1:1 verification, the number of positive/negative matches are 10k/8M in IJB- B and 19k/15M in IJB-C. We present the True Accept Rates (TARs) at False

Dataset	DS	SphereFace			ArcFace			MagFace		
		112	84	56	112	84	56	112	84	56
LFW	Uniform	99.67	92.19	84.56	99.81	92.44	85.16	99.83	92.51	85.20
	(Recasens et al. 2018)		93.55	89.17		93.57	89.79		94.01	89.97
	(Jin et al. 2022)		93.52	89.18		93.53	89.80		94.02	89.98
	Ours		94.57	90.11		94.84	90.59		94.70	91.01
CFP-FP	Uniform	96.84	88.50	75.32	98.40	88.79	76.77	98.46	88.82	79.80
	(Recasens et al. 2018)		90.09	79.76		89.12	78.45		89.12	78.45
	(Jin et al. 2022)		89.31	78.84		89.21	78.98		89.12	78.45
	Ours		89.71	80.45		90.12	80.45		90.95	80.45
CPLFW	Uniform	91.27	82.17	71.29	92.72	83.15	73.44	92.87	83.22	75.45
	(Recasens et al. 2018)		84.10	73.55		85.14	76.38		85.92	78.45
	(Jin et al. 2022)		84.29	73.41		86.00	73.74		86.17	78.17
	Ours		85.25	77.54		86.15	78.44		87.22	78.85
AgeDB	Uniform	97.05	88.91	76.01	98.05	89.12	77.29	98.17	89.01	80.15
	(Recasens et al. 2018)		90.11	81.29		91.32	80.46		91.27	82.95
	(Jin et al. 2022)		90.25	80.91		90.87	80.25		91.52	83.21
	Ours		90.72	81.51		91.44	81.29		92.12	84.15

Table 2: Face verification accuracy (%) on LFW, CFP-FP, CPLFW, and AgeDB datasets using different methods.

Dataset	DS	TAR@FAR 1e-6			TAR@FAR 1e-5			TAR@FAR 1e-4		
		112	84	56	112	84	56	112	84	56
CASIA	Uniform	91.86	85.14	79.17	94.07	87.11	78.20	95.86	89.54	83.39
	(Recasens et al. 2018)		87.21	82.54		87.21	78.94		90.12	85.37
	(Jin et al. 2022)		89.14	83.84		89.18	80.04		91.25	87.17
	Ours		89.21	83.91		89.21	80.94		91.21	87.37
PolyU	Uniform	91.09	86.37	80.24	94.11	87.32	79.11	96.10	90.10	83.91
	(Recasens et al. 2018)		89.21	82.54		87.92	79.31		90.12	85.37
	(Jin et al. 2022)		90.14	83.17		87.85	79.55		91.25	86.17
	Ours		90.37	84.19		89.21	80.94		91.12	87.37
TongJi	Uniform	92.19	87.46	81.18	95.81	86.11	73.20	97.24	87.62	81.39
	(Recasens et al. 2018)		89.04	84.35		88.51	76.14		88.15	84.50
	(Jin et al. 2022)		89.54	83.17		87.85	77.12		90.37	85.71
	Ours		89.87	84.17		88.25	77.55		89.25	86.17
MPD	Uniform	39.40	34.19	31.28	52.97	47.11	41.20	59.60	56.31	51.20
	(Recasens et al. 2018)		36.51	33.76		50.01	43.49		59.44	52.65
	(Jin et al. 2022)		36.71	34.01		49.71	43.50		59.25	52.65
	Ours		37.01	34.21		49.65	43.57		59.54	52.90
XJTU-UP	Uniform	63.57	57.82	51.25	71.24	68.81	63.34	82.12	78.59	66.39
	(Recasens et al. 2018)		60.64	55.70		70.94	65.91		79.01	67.50
	(Jin et al. 2022)		61.14	56.20		71.44	66.41		79.51	68.00
	Ours		61.37	56.41		71.67	66.64		79.75	68.25

Table 3: Face verification accuracy (%) on LFW, CFP-FP, CPLFW, and AgeDB datasets using different methods. The best results for each setting are highlighted in red.

Accept Rates (FAR) of 1e-6, 1e-5, and 1e-4, as detailed in Tab 4.

Quantitative Results Tab 2 shows the face verification accuracy on LFW, CFP-FP, CPLFW, and AgeDB datasets. We observe that our method consistently outperforms the baselines and other adaptive sampling methods across all datasets. For example, on the LFW dataset, our method achieves 94.57% accuracy with SphereFace, 94.84% with ArcFace, and 94.70% with MagFace, surpassing the best results of 93.55%, 93.57%, and 94.01% from other methods.

Tab 4 presents the results on the IJB-B and IJB-C datasets. On the IJB-B and IJB-C datasets, our method also achieves the highest TARs at all FARs, demonstrating its effectiveness in adaptive sampling for face recognition. These results indicate that our method can effectively learn to sample more discriminative regions in face images, leading to improved recognition performance.

Palmprint Recognition

Experimental Setup For palmprint recognition, we follow the setup in (Jin et al. 2025) and evaluate our method on CASIA (Sun et al. 2005), PolyU (Zhang et al. 2003), TongJi (Zhang et al. 2017), MPD (Zhang et al. 2020), and XJTU-UP (Shao, Zhong, and Du 2020) datasets. We follow a 1:1 training and testing protocol, where the model is trained on half of the identities, and tested on the other half. We augment the identities by flipping the palmprint images horizontally, and resize them to 112×112 pixels. We refer the readers to respect papers for details of these datasets.

Method	Res	DS	FPS	IJB-B (TAR@FAR)			IJB-C (TAR@FAR)		
				1e-6	1e-5	1e-4	1e-6	1e-5	1e-4
SphereFace	Baseline (112×112)		5,525	39.40	73.58	89.19	68.86	83.33	91.77
	96	Uniform	6,675	35.51	69.89	86.11	63.28	79.13	88.57
		(Recasens et al. 2018)	5,817	38.51	71.28	88.52	67.78	82.51	90.73
		(Jin et al. 2022)		38.54	71.85	88.91	68.12	82.33	91.65
		Ours	6,675	39.37	73.51	89.14	68.82	83.31	91.76
	84	Uniform	7,322	33.15	71.31	87.33	64.86	76.87	85.76
		(Recasens et al. 2018)	5,973	37.51	70.74	86.89	65.61	77.13	86.25
		(Jin et al. 2022)		36.88	71.05	87.04	65.37	77.14	86.19
		Ours	7,322	38.65	72.85	88.64	66.89	77.43	87.97
	56	Uniform	11,969	29.37	62.59	79.80	56.32	67.36	72.97
		(Recasens et al. 2018)	8,539	34.45	67.17	80.25	58.16	75.33	74.01
		(Jin et al. 2022)		34.58	67.25	80.15	58.21	75.23	74.16
		Ours	11,969	35.69	67.83	80.64	58.89	75.70	74.93
ArcFace	Baseline (112×112)		5,525	38.68	88.50	94.09	85.65	92.69	95.74
	96	Uniform	6,675	38.68	84.50	90.09	82.65	90.69	92.83
		(Recasens et al. 2018)	5,817	38.51	87.28	92.52	84.78	91.11	94.04
		(Jin et al. 2022)		38.54	86.85	92.91	85.12	91.13	94.15
		Ours	6,675	39.37	87.51	93.14	85.49	92.11	95.71
	84	Uniform	7,322	32.15	71.31	87.33	68.86	80.87	88.76
		(Recasens et al. 2018)	5,973	37.40	85.48	91.54	82.87	84.89	90.69
		(Jin et al. 2022)		37.51	85.58	91.19	82.86	85.33	90.77
		Ours	7,322	37.72	86.58	92.57	83.92	85.34	90.85
	56	Ours	11,969	29.86	64.17	81.25	57.14	73.36	74.50
		(Recasens et al. 2018)	8,539	34.50	70.17	85.25	65.16	77.33	79.01
		(Jin et al. 2022)		34.23	70.25	85.15	65.21	77.23	79.42
		Ours	11,969	35.12	71.63	86.72	67.50	78.45	80.01
MagFace	Baseline (112×112)		5,525	42.32	90.36	94.51	90.24	94.08	95.97
	96	Uniform	6,675	33.51	69.89	86.11	63.28	79.13	88.57
		(Recasens et al. 2018)	5,817	38.51	71.28	91.52	67.78	82.51	93.73
		(Jin et al. 2022)		38.54	71.85	91.91	68.12	82.33	94.65
		Ours	6,675	39.37	73.51	94.14	88.82	93.31	95.16
	84	Uniform	7,322	32.15	71.31	87.33	68.86	80.87	88.76
		(Recasens et al. 2018)	5,973	39.40	73.58	89.19	86.17	86.33	89.72
		(Jin et al. 2022)		39.40	73.58	88.19	86.28	85.79	90.10
		Ours	7,322	38.65	72.85	92.64	86.89	86.43	90.97
	56	Uniform	11,969	31.37	64.29	81.21	57.55	74.07	75.69
		(Recasens et al. 2018)	8,539	35.45	67.17	89.25	69.16	83.83	87.01
		(Jin et al. 2022)		35.58	67.25	89.15	68.77	83.93	86.76
		Ours	11,969	36.81	67.75	89.79	69.71	83.70	87.93

Table 4: Face verification accuracy (%) on the IJB-B and IJB-C datasets, comparing different recognition models and downsampling (DS) methods. The best results for each resolution are highlighted in red.

We use the combo of ArcFace (Deng et al. 2022) loss with a modified ResNet-50 (He et al. 2016) backbone that is widely used in recent palmprint recognition works (Jin et al. 2025, 2024; Zhao et al. 2022; Shen et al. 2023).

The performance is evaluated on each individual dataset in terms of True Accept Rate (TAR) at False Accept Rate (FAR) of 1e-6, 1e-5, and 1e-4.

Interpretability Study

In this section, we analyze the learned downsampling grids and their effectiveness in focusing on discriminative regions of the input images.

Learned Downsampling Grids Fig 3 visualizes the learned downsampling grids (top) and warpped images (bottom) for face (left) and palmprint (right) recognition tasks.

The learned grids adaptively focus on the discriminative regions of the input images. For face recognition, the grids are denser around the eyes and nose, which are crucial for identity discrimination. For palmprint recognition, the grids are denser around the core of the palmprint, where the unique features are located.

Jacobian Regularization To further analyze the learned grids, we visualize the Jacobian determinants of the downsampling grids in Fig 4. The Jacobian determinant indicates the local expansion or compression of the sampling grid. Higher values indicate local expansion, leading to sparser sampling in those regions. The model learns to compress the sampling grid in the facial areas and sparse it in the background. This is particularly beneficial for face recognition, as it allows the model to focus on the most discriminative features while ignoring less relevant background informa-

Figure 3: Learned downsampling grids (top) and their corresponding warpped images (bottom) for face recognition (left) and palmprint recognition (right). The face image is from the synthetic DiGiFace dataset, and the palmprint image is blurred for privacy.

tion.

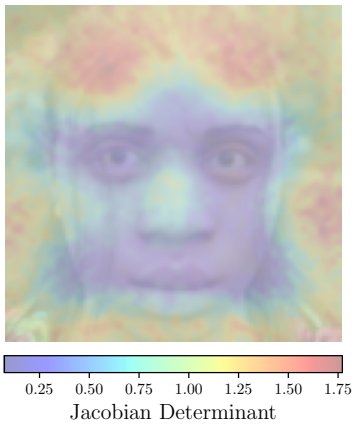

Figure 4: Jacobian determinants of the learned downsampling grids. Higher values indicate local expansion and thus sparser sampling in those regions. The model learns to compress the sampling grid in the facial areas and sparse it in the background.

Input-Dependent vs. Input-Independent

The key difference between our method and prior approaches (Recasens et al. 2018; Jin et al. 2022) lies in the design of the downsampling grids. While previous methods generate input-dependent grids predicted from each image, our method learns a fixed, input-independent grid.

In recognition tasks, geometric alignment is essential for learning transformation-invariant and identity-specific features. Input-dependent sampling can disrupt this alignment, potentially degrading performance. This ablation study evaluates both approaches—predictive downsampling (input-dependent grids) and learn-to-downsample (input-independent grids)— using our method and those from (Recasens et al. 2018) and (Jin et al. 2022). The former predicts a unique grid per sample, while the latter learns a single grid shared across the dataset.

Results in Tab 5 show that learn-to-downsample consistently outperforms predictive downsampling across all methods. This highlights the importance of maintaining consistent image structure, which enables more robust, invariant feature learning.

The results in Tab 5 also reveal that our flow-based representation outperforms the importance maps used in (Recasens et al. 2018; Jin et al. 2022), as evidenced by the higher TARs using both learn and predict downsampling.

Method	DS	TAR@1e-6	TAR@1e-5	TAR@1e-4
Face Recognition (IJB-B dataset, 56×56)				
Recasens	Learn	34.56	70.75	85.42
	Pred	34.50	70.17	85.25
Jin	Learn	34.87	70.95	86.01
	Pred	34.23	70.25	85.15
Ours	Learn	35.12	71.63	86.72
	Pred	34.70	71.18	86.17
Palmprint Recognition (CASIA dataset, 56×56)				
Recasens	Learn	82.37	78.59	85.13
	Pred	82.54	78.94	85.37
Jin	Learn	83.79	79.67	86.98
	Pred	83.84	80.04	87.17
Ours	Learn	89.31	80.94	87.37
	Pred	88.12	79.67	86.98

Table 5: Comparison of predictive and learn-to-downsample methods. All images are downsampled to 56×56. Face recognition is evaluated on the IJB-B dataset, and palmprint recognition on the CASIA dataset. Predictive methods consistently outperforms learn-to-downsample methods, and our method achieves the best performance in both cases.

Conclusion

In this paper, we introduced a novel learned sampling paradigm for aligned vision recognition. Our method learns a flow field to guide downsampling, which is input-independent and can be treated as a model parameter. This approach enables efficient downsampling without requiring access to the input image, thereby significantly reducing computational overhead. We demonstrated the effectiveness of our method on two aligned vision recognition tasks: face recognition and palmprint recognition. We believe that our method opens up new possibilities for efficient and effective image processing in aligned vision recognition tasks.

Aknowledgments

This work was partially supported by the National Natural Science Foundation of China (62372284, 62476143), the National Key Laboratory of Science and Technology on Space-Born Intelligent Information Processing (TJ-02-22-01), and the Pioneer R&D Program of Zhejiang Province (2024C01024).

References

Bruce, N. D.; and Tsotsos, J. K. 2009. Saliency, attention, and visual search: An information theoretic approach. *Journal of vision*, 9(3): 5–5.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773.

Deng, J.; Guo, J.; Yang, J.; Xue, N.; Kotsia, I.; and Zafeiriou, S. 2022. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 5962–5979.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*.

Fayyaz, M.; Abbasi Kouhpayegani, S.; Rezaei Jafari, F.; Sommerlade, E.; Vaezi Joze, H. R.; Pirsiavash, H.; and Gall, J. 2022. Adaptive token sampling for efficient vision transformers. *European Conference on Computer Vision (ECCV)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

Huang, G.; Chen, D.; Li, T.; Wu, F.; van der Maaten, L.; and Weinberger, K. 2018. Multi-Scale Dense Networks for Resource Efficient Image Classification. In *International Conference on Learning Representations*.

Huang, G. B.; Mattar, M.; Berg, T.; and Learned-Miller, E. 2008. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition*.

Itti, L.; Koch, C.; and Niebur, E. 2002. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11): 1254–1259.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial transformer networks. *Advances in neural information processing systems*, 28.

Jiang, Y.; Wang, H.; Xie, L.; Zhao, H.; Qian, H.; Lui, J.; et al. 2024. D-llm: A token adaptive computing resource allocation strategy for large language models. *Advances in Neural Information Processing Systems*, 37: 1725–1749.

Jin, C.; Tanno, R.; Mertzanidou, T.; Panagiotaki, E.; and Alexander, D. C. 2022. Learning to Downsample for Segmentation of Ultra-High Resolution Images. In *International Conference on Learning Representations*.

Jin, J.; Shen, L.; Zhang, R.; Zhao, C.; Jin, G.; Zhang, J.; Ding, S.; Zhao, Y.; and Jia, W. 2024. PCE-Palm: Palm Crease Energy Based Two-Stage Realistic Pseudo-Palmprint Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2616–2624.

Jin, J.; Zhao, C.; Zhang, R.; Shang, S.; Xu, J.; Zhang, J.; Wang, S.; Zhao, Y.; Ding, S.; Jia, W.; et al. 2025. Diff-Palm: Realistic Palmprint Generation with Polynomial Creases and Intra-Class Variation Controllable Diffusion Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 26367–26376.

Kim, M.; Jain, A. K.; and Liu, X. 2022. Adaface: Quality adaptive margin for face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 25, 1097–1105.

Li, B.; Xi, T.; Zhang, G.; Feng, H.; Han, J.; Liu, J.; Ding, E.; and Liu, W. 2021a. Dynamic class queue for large scale face recognition in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Li, S.; Xu, J.; Xu, X.; Shen, P.; Li, S.; and Hooi, B. 2021b. Spherical confidence learning for face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Meng, Q.; Zhao, S.; Huang, Z.; and Zhou, F. 2021. Magface: A universal representation for face recognition and quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Mnih, V.; Heess, N.; Graves, A.; and Kavukcuoglu, K. 2014. Recurrent models of visual attention. *Advances in neural information processing systems*, 27.

Moschoglou, S.; Papaioannou, A.; Sagonas, C.; Deng, J.; Kotsia, I.; and Zafeiriou, S. 2017. Agedb: the first manually collected, in-the-wild age database. In *CVPRW*.

Recasens, A.; Kellnhofer, P.; Stent, S.; Matusik, W.; and Torralba, A. 2018. Learning to zoom: a saliency-based sampling layer for neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, 51–66.

Sengupta, S.; Chen, J.-C.; Castillo, C.; Patel, V. M.; Chellappa, R.; and Jacobs, D. W. 2016. Frontal to profile face verification in the wild.

Shao, H.; Zhong, D.; and Du, X. 2020. Effective deep ensemble hashing for open-set palmprint recognition. *Journal of Electronic Imaging*, 29(1): 013018.

Shen, L.; Jin, J.; Zhang, R.; Li, H.; Zhao, K.; Zhang, Y.; Zhang, J.; Ding, S.; Zhao, Y.; and Jia, W. 2023. RPG-Palm: Realistic Pseudo-data Generation for Palmprint Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19605–19616.

Sun, Z.; Tan, T.; Wang, Y.; and Li, S. Z. 2005. Ordinal palmprint represention for personal identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 1, 279–284. IEEE.

Veit, A.; and Belongie, S. 2018. Convolutional networks with adaptive inference graphs. In *Proceedings of the European conference on computer vision (ECCV)*, 3–18.

Wang, X.; Yu, F.; Dou, Z.-Y.; Darrell, T.; and Gonzalez, J. E. 2018. Skipnet: Learning dynamic routing in convolutional networks. In *Proceedings of the European conference on computer vision (ECCV)*, 409–424.

Whitelam, C.; Taborsky, E.; Blanton, A.; Maze, B.; Adams, J.; Miller, T.; Kalka, N.; Jain, A. K.; Duncan, J. A.; Allen, K.; et al. 2017. Iarpa janus benchmark-b face dataset. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 90–98.

Wu, Z.; Nagarajan, T.; Kumar, A.; Rennie, S.; Davis, L. S.; Grauman, K.; and Feris, R. 2018. Blockdrop: Dynamic inference paths in residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8817–8826.

Zhang, D.; Kong, W.-K.; You, J.; and Wong, M. 2003. Online palmprint identification. *IEEE Transactions on pattern analysis and machine intelligence*, 25(9): 1041–1050.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 23(10): 1499–1503.

Zhang, L.; Li, L.; Yang, A.; Shen, Y.; and Yang, M. 2017. Towards contactless palmprint recognition: A novel device, a new benchmark, and a collaborative representation based identification approach. *Pattern Recognition*, 69: 199–212.

Zhang, Y.; Zhang, L.; Zhang, R.; Li, S.; Li, J.; and Huang, F. 2020. Towards Palmprint Verification On Smartphones. *arXiv preprint arXiv:2003.13266*.

Zhao, K.; Shen, L.; Zhang, Y.; Chuhan, Z.; Wang, T.; Zhang, R.; Shouhong, D.; Jia, W.; and Shen, W. 2022. BézierPalm: A Free lunch for Palmprint Recognition. In *European Conference on Computer Vision (ECCV)*.

Zhao, T.; Kiblawi, S.; Usuyama, N.; Lee, H. H.; Preston, S.; Poon, H.; and Wei, M. 2025. Boltzmann Attention Sampling for Image Analysis with Small Objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 25950–25959.

Zheng, T.; and Deng, W. 2018. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5: 7.

Zheng, T.; Deng, W.; and Hu, J. 2017. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*.