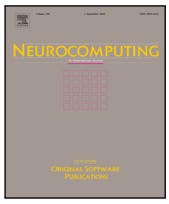
Neurocomputing 614 (2025) 128846

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



An analysis of pre-trained stable diffusion models through a semantic lens

Simone Bonechi [a,b,*], Paolo Andreini [b], Barbara Toniella Corradini [b,c], Franco Scarselli [b]

[a] Department of Social, Political and Cognitive Science, University of Siena, Siena, Italy
[b] Department of Information Engineering and Mathematics, University of Siena, Siena, Italy
[c] Department of Computer Engineering, University of Florence, Florence, Italy

ARTICLE INFO

Keywords:
Generative models
Diffusion models
Stable diffusion
Pre-trained models
Semantic preservation
Layer analysis
Multimodal systems

ABSTRACT

Recently, generative models for images have garnered remarkable attention, due to their effective generalization ability and their capability to generate highly detailed and realistic content. Indeed, the success of generative networks (*e.g.*, BigGAN, StyleGAN, Diffusion Models) has driven researchers to develop increasingly powerful models. As a result, we have observed an unprecedented improvement in terms of both image resolution and realism, making generated images indistinguishable from real ones. In this work, we focus on a family of generative models known as Stable Diffusion Models (SDMs), which have recently emerged due to their ability to generate images in a multimodal setup (*i.e.*, from a textual prompt) and have outperformed adversarial networks by learning to reverse a diffusion process. Given the complexity of these models that makes it hard to retrain them, researchers started to exploit pre-trained SDMs to perform downstream tasks (*e.g.*, classification and segmentation), where semantics plays a fundamental role. In this context, *understanding how well the model preserves semantic information may be crucial to improve its performance.*

This paper presents an approach aimed at providing insights into the properties of a pre-trained SDM through the semantic lens. In particular, we analyze the features extracted by the U-Net within a SDM to explore whether and how the semantic information of an image is preserved in its internal representation. For this purpose, different distance measures are compared, and an ablation study is performed to select the layer (or combination of layers) of the U-Net that best preserves the semantic information. We also seek to understand whether semantics are preserved when the image undergoes simple transformations (*e.g.*, rotation, flip, scale, padding, crop, and shift) and for a different number of diffusion denoising steps. To evaluate these properties, we consider popular benchmarks for semantic segmentation tasks (*e.g.*, COCO, and Pascal-VOC). Our experiments suggest that the first encoder layer at 16×16 resolution effectively preserves semantic information. However, increasing inference steps (even for a minimal amount of noise) and applying various image transformations can affect the diffusion U-Net's internal feature representation. Additionally, we propose some examples taken from a video benchmark (DAVIS dataset), where we investigate if an object instance within a video preserves its internal representation even after several frames. Our findings suggest that the internal object representation remains consistent across multiple frames in a video, as long as the configuration changes are not excessive.

1. Introduction

In recent years, generative models have gained significant traction for their remarkable ability to create increasingly realistic images, revolutionizing fields such as digital art [1], virtual reality [2], and medical imaging [3–5]. These advancements have been largely driven by sophisticated architectures like Generative Adversarial Networks (GANs) [6] and Variational Autoencoders (VAEs) [7], which enable the synthesis of high-fidelity images that often blur the line between computer-generated and real-world pictures.

Generative models aim to learn the underlying distribution of a dataset to generate unseen, realistic samples that resemble the real ones. Their ability to effectively capture the nuances and complexities of the original data enables the creation of novel content that preserves the realism and diversity of the input data. Following groundbreaking generative models like GAN and VAE, which marked significant advancements in Deep Learning, subsequent models (*e.g.*, DCGAN [8], WGAN [9]) have pushed the state-of-the-art by improving the quality of generation in terms of both image resolution and level of detail. However, while these models struggled to generate complex content

* Corresponding author at: Department of Social, Political and Cognitive Science, University of Siena, Siena, Italy.
E-mail address: simone.bonechi@unisi.it (S. Bonechi).

https://doi.org/10.1016/j.neucom.2024.128846
Received 4 July 2024; Received in revised form 28 August 2024; Accepted 31 October 2024
Available online 8 November 2024

at high resolutions, the ambition to generate increasingly more realistic and diverse images has led to the design of progressively more complex and resource-intensive architectures. Models like StyleGAN [10] and BigGAN [11], while capable of producing high-resolution and remarkably realistic images, require training times that can stretch over weeks, making their training difficult in terms of computational resources and time.

The emergence of multimodal models, particularly Visual-Language Models (VLMs), with their expanding parameter count, has further exacerbated the challenges of training a generative model from scratch. A VLM aims to create images from diverse sources of content (*e.g.*, text), thereby learning a multimodal latent space where different types of information can coexist. Moving from the early multimodal models (*e.g.*, VQ-VAE-2 [12]), the advent of new engines based on vision and language, such as Transformers [13], Diffusion Denoising Probabilistic Models (DDPMs) [14] and Large Language Models (LLMs) [13] has indeed led to a remarkable improvement but it has also led to a rapid escalation in the number of parameters. Additionally, the vast datasets required for training (*e.g.*, LAION-400M [15], LAION-5B [16]), involving hundreds of millions of image–text data, make these models extremely powerful but also nearly impossible to train in absence of vast computational resources.

The introduction of DDPMs marked a significant advancement in image generation. DDPMs are likelihood-based generative models that utilize a U-Net backbone and variational inference to learn a denoising Markov chain. They have set a new benchmark in image generation, surpassing BigGAN and VQ-VAE-2 in terms of Fréchet Inception Distance [17] metrics on ImageNet [18]. Recent advancements in this field have also demonstrated remarkable results in text-to-image generation with models such as DALL-E [19], Imagen [20], and Stable Diffusion Models (SDMs) [21].

Given the impressive performance of VLMs but the impracticality of retraining generative multimodal models from scratch, *new research avenues have emerged that leverage pre-trained models*. Indeed, as VLMs are trained on large-scale datasets, they can perform tasks using a zero-(few-)shot approach, eliminating the need for retraining.

However, using a pre-trained model requires a deep understanding of how the information is encoded within its internal representation and how to retrieve this information. Indeed, overlooking these aspects could potentially compromise the model's performance, particularly in a downstream task. For instance, it would be beneficial to determine whether the latent vector of a VLM identifies specific objects in the image, captures specific details of objects such as textures and colors, or both simultaneously. Indeed, models like CLIP [22], which typically learn multimodal embedding spaces by correlating text and images, suggest that cosine similarity between embeddings can reflect semantic similarity between objects. This indicates that the multimodal space is highly structured, which enhances its interpretability [23]. However, it remains uncertain whether this principle extends to the multimodal space of Diffusion Models. If we could apply insights to complex generative models such as SDMs, we might infer that the internal representation of a *zebra* is closer to that of a *horse* than to that of a *truck*. Therefore, a current research challenge is understanding how to effectively retrieve information from the internal representation of pre-trained diffusion models to fully leverage their potential for tasks such as image classification [24–26] and segmentation [27–29] without any training. In this work, we aim to tackle this challenge by analyzing the internal representation of a pre-trained diffusion model through the semantics lens. Thus, we wonder:

Does the U-Net in SDM reflect the semantic similarity between objects in its internal representation?

To the best of our knowledge, while many studies offer insights into the internal dynamics of Stable Diffusion, there is a lack of systematic contributions in this area. This work aims to provide new insights into whether and how features of an SDM preserve semantics, particularly by examining the features extracted from the diffusion U-Net at different depths. We propose a distance-based approach to provide an overview of each layer's capability in preserving object semantics within an image, as well as the relationship among the internal representation of various objects in the U-Net encoder and decoder. In a nutshell, as shown in Fig. 1, we feed an image to the SDM and we exploit the target segmentation mask to extract the internal features corresponding to different objects in the scene. Given the corresponding prototypes, we then infer distances between the objects in the multimodal feature space. If semantics are preserved, we expect the distances to exhibit specific patterns, *i.e.*, instances of the same class should be closer to each other, according to certain metrics, than instances of different classes, both within and across images.

To study how well the diffusion U-Net preserves semantic information within its feature representation during the diffusion process, we employ several distance measures and an ablation study to identify the specific U-Net layer(s) that best preserve semantic meaning. Extending our seminal work [30], we explore the impact of simple image transformations (*e.g.*, rotation, flip, scale, padding, crop, and shift) and a different number of denoising steps on semantic preservation. To evaluate our approach, we used two semantic segmentation datasets (COCO [31] and Pascal-VOC [32]). We also delved into video data using the DAVIS [33] dataset, to evaluate whether the feature representation of the same object instance is preserved across multiple frames.

The main contribution of the paper can be summarized as follows:

- We present an approach based on calculating distances in the feature space of a diffusion U-Net.
- We provide insights for studying the contribution of various U-Net layers, revealing that the first 16×16 layer of the U-Net encoder appears to be the most effective at preserving the semantic information of objects within the image.
- We investigate how the number of inference steps affects the semantics, confirming that adding noise for an increasing number of time steps progressively degrades the semantic information.
- We analyze whether semantics are preserved when the image undergoes simple transformations (rotation, flip, scale, padding, crop, and shift). Interestingly, cropping an object with padding seems to preserve semantic content, whereas scaling introduces a slight loss in representation.
- Finally, we demonstrate that semantic information can be preserved within a video sequence even after several frames, which could provide meaningful insights for downstream tasks such as object tracking and object re-identification.

The paper is organized as follows: Section 2 revises the literature related to the main aspects of diffusion models and their interpretability, while Section 3 describes the datasets and metrics used in this study. Then, the proposed method is detailed in Section 4. Sections 5 and 6 present the experimental setup and the obtained results respectively. Finally, Section 7 concludes and discusses possible future developments.

2. Related work

2.1. Diffusion (probabilistic) models

Diffusion Models (DMs), also known as Diffusion Probabilistic Models, are a class of generative models inspired by thermodynamics that has garnered attention for their capability to produce high-quality images by learning to reverse a diffusion process [34]. A DM is based on a forward and a reverse process.

In the *Forward diffusion process*, from image to noise, the input image ($\mathbf{x}_0$) sampled from the real data distribution $q(\mathbf{x}_0)$ ($\mathbf{x}_0 \sim q(\mathbf{x}_0)$) undergoes incremental degradation by incorporating noise sampled

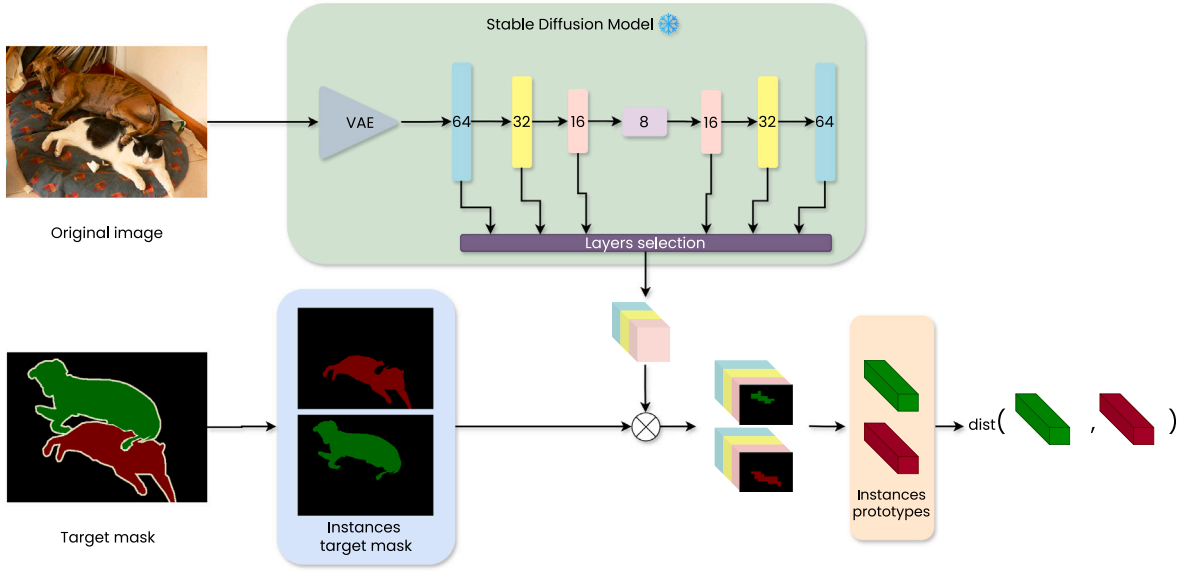


Fig. 1. Overview of proposed approach. Given an image, the corresponding SDM internal features are extracted at various depths and resolutions, and filtered based on the target masks. The filtered features are averaged and the prototypes of the objects thereby obtained are compared in terms of distance within the feature space. *Our aim is to demonstrate the semantic proximity of classes within the feature space of an SDM*.

from a Gaussian distribution across a predetermined number of time steps T. After the last forward step, the image has lost its original appearance and $\mathbf{x}_T$ is equivalent to an isotropic Gaussian distribution (*i.e.*, pure noise). The Markov chain that describes the degradation of the distribution $q(\mathbf{x})$ of $\mathbf{x}_0$ after each time step of the forward process can be formalized as in Eq. (1):

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \prod_{t=1}^{T} \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}), \tag{1}$$

where $\mathcal{N}$ is a Gaussian distribution, $\mathbf{I}$ is the identity matrix, and β coefficients are either given by a scheduler or fixed [35,36] to ensure that $\mathbf{x}_T$ is nearly an isotropic Gaussian for sufficiently large T.

The *Reverse diffusion process* aims to invert the diffusion process and learn the distribution $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$, meaning to recognize the specific noise patterns introduced at each step and generate new samples by removing noise. The problem of calculating the conditional probabilities is intractable as it depends on the entire dataset, but a model p_θ (*e.g.*, a neural network) can be used to approximate the reverse process

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = p(\mathbf{x}_T) \prod_{t=1}^{T} \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)). \tag{2}$$

In particular, the reverse process is driven by a U-shaped neural network trained to predict the amount of noise added to the image at time step t following the distribution in Eq. (2) by minimizing the loss function (in Eq. (3))

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0, t, \bar{\epsilon}} \parallel \bar{\epsilon} - \epsilon_\theta(\mathbf{x}_t, t) \parallel_2^2, \tag{3}$$

where $\epsilon_\theta(\mathbf{x}_t, t)$ is the predicted amount of noise at time step t, denoted as, and $\bar{\epsilon}$ is the true amount of added noise.

The first implementation of diffusion models, DDPMs, could generate images by introducing noise using a Markov chain. Given the need to calculate the distribution at each stage, performing diffusion as a Markovian process incurred an almost prohibitive computational cost. To overcome this issue, DDIMs (Denoising Diffusion Implicit Models) [37] generalized to a non-Markovian process, demonstrating the ability to produce higher-resolution images with lower computational cost. However, both DDPMs and DDIMs generate images directly from the pixel space making these approaches significantly more computationally expensive compared to other generative methods (*e.g.*, GANs, VAEs). Latent Diffusion Models (LDMs) [21] use the encoder $\mathcal{E}$ of a VAE to create a latent encoding of the image in a lower-dimensional space, then apply diffusion to the latent vector and finally decode the latent vector via the VAE decoder $\mathcal{D}$. Applying the diffusion operations to the latent space not only enhanced speed and reduced costs but also allowed to extend the model with new capabilities, such as generating images based on specific additional inputs, using a conditioning mechanism [38]. Possible types of conditioning include textual prompts (text-to-image diffusion models) or another image (image-to-image diffusion models). To enable the conditioned image generation, the denoising U-Net makes use of a cross-attention mechanism to incorporate conditioning information during the reverse process.

In this work, we aim to explore the semantic properties of a pre-trained diffusion model known as SDM, a type of text-to-image LDM that has emerged as the state-of-the-art in image generation due to its capability to produce photorealistic images based on any given textual prompt.

2.2. Use of diffusion models internal features

DMs are usually trained on huge datasets containing hundreds of millions of examples (*e.g.*, LAION-400M and LAION-5B contain 400 million and 5.85 billion of image–text pairs, respectively), making re-training almost unaffordable. New research directions have therefore emerged aiming to use pre-trained DMs to improve their generative power, for instance enhancing the embeddings of images or engineering textual prompts to generate better samples [39,40]. In addition, a certain number of works use DMs to perform other tasks than mere image generation, such as classification and segmentation. In [41], the authors explore a DDPM to identify which layers are most effective for semantic segmentation. They demonstrate that it is possible to aggregate the internal features of a DDPM using K-Means and obtain a spatially coherent representation of the image. The analysis is conducted on various blocks of the decoder of the U-Net at different diffusion timesteps t. Finally, an MLP is trained to predict the semantic label of a pixel based on its U-Net features. ODISE (Open-vocabulary DIffusion-based panoptic SEgmentation) [42] is a model for panoptic image segmentation. ODISE relies on learning fine masks obtained from raw masks extracted from the layers of a U-Net within an SDM. They demonstrate superior performance compared to clustering-based approaches of the internal representation. LD-ZNet [43] shows that the

internal features of LDMs contain rich semantic information to perform text-based segmentation of synthetic images. In ASYRP paper [44], a special latent space h is introduced, which exhibits significant algebraic properties. To this aim they introduce a 1×1 convolutional layer, which processes the concatenation sequence of bottleneck representations from the U-Net for each time step. Despite the excellent properties of h enabling straightforward modification of image characteristics, it is unrelated to the feature space or latent space of the SDM. Adopting the Riemannian geometry serves in [45] to understand the latent space of a diffusion model. The authors focus on finding a vector basis for the latent space by leveraging the pullback metric associated with their encoding feature maps. Their discoveries allow to move through the latent space and perform image editing via parallel transport of the vector basis. Moreover, other works have shown alternative uses of the internal representation of a pre-trained SD. In [46], the internal representation is exploited to quantify the social bias in text-to-image generative models. Instead, in [47], a neural network is trained to modify the features of a pre-trained SD to follow specific multimodal conditioning and improve generated image quality.

3. Preliminaries

Section 3.1 presents the three datasets used in this study, while Sections 3.2 and 3.3 describe the similarity measures and the metrics employed to analyze the extracted features, respectively.

3.1. Datasets

3.1.1. Pascal-VOC 2012

The Pascal-VOC dataset [32] is a popular benchmark for image segmentation. It is characterized by images with a limited number of object categories. The primary object is typically positioned in the center of the scene. Each image in the training and validation sets comes with pixel-level annotation. The dataset consists of 20 object categories, a background class, and a "do not care" class for uncertain regions. In this study, we randomly selected 1000 Pascal-VOC images,[1] ensuring that each of them contains at least one object with a size greater than 1% of the total image area. This criterion helps to focus on images with well-represented objects.

3.1.2. COCO 2017

The COCO dataset [31] is a large-scale image dataset containing more than 100,000 images designed for object detection, segmentation, and captioning tasks. The images often capture a complex scene with various types of objects. Each image in the training and validation set comes with instance-level annotations for 80 object categories, along with background labels. Although COCO-2017 offers a wide range of object categories, in this study we extract a subset of 1000 images[1] tailored to our needs following these criteria:

- Object Size: Each image must contain at least one object with a size greater than 1% of the total image area. This ensures to focus on images with "well-represented" objects, avoiding images where the object size is too small to provide a meaningful representation in the SDM.
- Object Category: The object category should belong to one of the 20 Pascal-VOC classes. This ensures to focus on a common set of classes between the two datasets.

[1] The list of the images/videos employed in this work can be downloaded at: https://github.com/bcorrad/Diff-Props.

3.1.3. DAVIS

The Densely-Annotated VIdeo Segmentation (DAVIS) dataset [48] provides high-definition video sequences with pixel-level object mask annotations for each frame. In this work, we use the DAVIS 2017 [33] dataset, which is a larger and more challenging version released in 2017. The complexity of the dataset was increased not only by annotating multiple objects per scene but also by re-annotating a number of original samples, so as to include more distractors, smaller objects, finer structures, occlusions, and faster movements. The expanded dataset is composed of 150 video sequences containing 10,459 annotated frames. In this paper, we consider a subset of DAVIS 2017 composed of 544 frames from 8 video samples[1], showing multiple instances of people in the scene. Such a selection aims to investigate whether features referred to a specific person instance in a video frame can retain semantic information in the subsequent frames.

3.2. Distance measures

This section describes the metrics used to compare feature vectors within the latent space of the U-Net in the SDM. These metrics allow us to quantify the relationships between features in the latent space. Given two n-dimensional vectors $\mathbf{X} = (x_1, x_2, \dots, x_n)$ and $\mathbf{Y} = (y_1, y_2, \dots, y_n)$ we can define the following measures.

Euclidean distance. The Euclidean distance between two vectors, $\mathbf{X}$ and $\mathbf{Y}$ is defined as

$$D_{\text{euc}}(\mathbf{X}, \mathbf{Y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}\,. \tag{4}$$

Distance correlation. The Distance Correlation [49] is a statistical measure that captures both linear and nonlinear relationships between variables, offering a comprehensive view of their dependence beyond traditional correlation metrics. It is defined as

$$D_{\text{corr}}(\mathbf{X}, \mathbf{Y}) = \sqrt{\frac{dCov^2(\mathbf{X}, \mathbf{Y})}{\sqrt{dVar^2(\mathbf{X}) dVar^2(\mathbf{Y})}}}\,, \tag{5}$$

where the distance covariance $dCov^2(\mathbf{X}, \mathbf{Y})$ and the distance variance $dVar^2(\mathbf{X})$ between $\mathbf{X}$ and $\mathbf{Y}$ are

$$dVar^2(\mathbf{X}) = dCov^2(\mathbf{X}, \mathbf{X}) \tag{6}$$

$$dCov^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (a_{ij} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a})(b_{ij} - \bar{b}_{i\cdot} - \bar{b}_{\cdot j} + \bar{b})\,. \tag{7}$$

If x_i, x_j are two samples of $\mathbf{X}$ and y_k, y_l are two samples of $\mathbf{Y}$, it is possible to define:

$a_{ij} = D_{euc}(x_i, x_j)$, $b_{kl} = D_{euc}(y_k, y_l)$ and $\bar{a}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} a_{ij}$, $\bar{a}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} a_{ij}$, $\bar{b}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} b_{ij}$, $\bar{b}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} b_{ij}$

Manhattan distance. The Manhattan distance [50] in an arbitrary space is defined as the sum of the absolute differences between the corresponding coordinates along each dimension

$$D_{\text{man}}(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{n} |x_i - y_i|\,. \tag{8}$$

Cosine distance. The cosine distance between two non-zero vectors, $\mathbf{X}$ and $\mathbf{Y}$, is defined as

$$D_{\cos}(\mathbf{X}, \mathbf{Y}) = 1 - \frac{\mathbf{X} \cdot \mathbf{Y}}{\|\mathbf{X}\| \|\mathbf{Y}\|}\,, \tag{9}$$

where $\cdot$ denotes the dot product of the vectors, and $\|\mathbf{X}\|$ and $\|\mathbf{Y}\|$ are the Euclidean norms of the vectors. The $\frac{\mathbf{X}\cdot\mathbf{Y}}{\|\mathbf{X}\|\|\mathbf{Y}\|}$ is the cosine similarity between $\mathbf{X}$ and $\mathbf{Y}$, which quantifies the similarity between vectors based on the angle between them. The cosine distance is simply the complement of the cosine similarity.

3.3. Metrics and reliability measures

Dunn index. The Dunn Index [51] measures the compactness (intra-cluster similarity) and the separation between clusters (inter-cluster dissimilarity). It can be defined as the ratio of the smallest inter-cluster distance to the largest intra-cluster distance:

$$\text{Dunn Index} = \frac{\min_{i \neq j} d_{min}(C_i, C_j)}{\max_k d_{max}(x \in C_k)} \tag{10}$$

where $d_{min}(C_i, C_j)$ represents the minimum pairwise distance between any two clusters C_i, C_j and $d_{max}(x \in C_k)$ calculates the maximum distance between any two points within a single cluster. The higher the Dunn Index, the better.

Kruskal–wallis. The Kruskal–Wallis test [52] is a non-parametric method for assessing the equality of medians across multiple groups and serves as an alternative to one-way ANOVA [53] when its assumptions, such as normality and homogeneity of variances, are not satisfied. Indeed, the Kruskal–Wallis test does not require the data to follow a specific distribution. In this study, we apply the Kruskal–Wallis test to assess whether there are statistically significant differences between the groups of interest. We chose this test after assessing the non-normal distribution of our samples using the Shapiro–Wilk test [54].

4. Proposed method

The goal of this work is to provide new insights into whether and how object representation is preserved in the features extracted from the diffusion U-Net at different depths. In the following, we provide an in-depth explanation of the proposed method. Specifically, we focus on the process of extracting prototypes from each object in an image (see Section 4.1) and we detail how to obtain the sets of distances for the comparison of these sets (see Section 4.2).

4.1. Prototype extraction

In order to measure the distance between features of different objects, we need to retrieve a feature-based prototype of the objects in the image. Prototypes can be obtained from the internal representations of pre-trained neural networks by masking their internal features with segmentation maps provided by the target labels [55,56] or through unsupervised methods [42]. Alternatively, prototypes can be created by averaging the embedding of multiple images containing only a single instance of the desired [28] class. Inspired by these works, we propose the following processing pipeline to extract the object prototype exploiting the features of the SDM (see Fig. 1).

1. *Image Encoding:* The images are first encoded through the VAE of the SDM.
2. *U-Net feature extraction:* The images are passed through the diffusion U-Net backbone, and their internal representation is extracted.
3. *Feature Concatenation (Optional):* Once the features are extracted, they are bilinearly resized to match the highest spatial dimension (height and width) across all selected layers. Features are then concatenated along the channel dimension, resulting in a single feature tensor that contains the information from all the chosen layers.
4. *Object feature masking:* To isolate object-specific features, we leverage the masks of the target segmentation instance masks as in [55,56]. The target instance mask is scaled at the resolution of the feature maps, and then we use the target instance mask to select the object feature maps. Therefore, we can isolate specific object instances within the feature space and create new filtered feature maps with information only about the masked objects.
5. *Prototype extraction:* For each object instance in the feature space, inspired by [56], we then compute the average of all its feature components, yielding a prototype vector for each object.

4.2. Collection of distance sets

Once the procedure in Section 4.1 is performed on the datasets, we obtain a prototype representation of each object within the images. To evaluate how much object semantics are preserved in the latent space, inspired by [27], the prototype vectors are compared using the described distance metrics in Section 3.2. More precisely, we consider different case studies (sets of distances), related to different pairs of object instances.

Formally, a set is defined as $S = \{D(I_{i,c}, J_{j,k})\}$, where I an J indicate two different object instances, i and j are the images containing the instances, c and k are the object classes, and D is the distance used to compare the two prototypes. The following sets of distances have been defined.

Same class - same image (S_{scsi}). The set contains the comparison between instances belonging to the same class within the same image

$$S_{\text{scsi}} = \{D(I_{i,c}, J_{j,k}) | i = j, c = k\}. \tag{11}$$

Same class - different image (S_{scdi}). In this set, each instance from an image is compared with all other instances of the same class located in different images

$$S_{\text{scdi}} = \{D(I_{i,c}, J_{j,k}) | i \neq j, c = k\}. \tag{12}$$

Same instance - different image (S_{sidi}). In this set, an instance from an image is compared with the same instance in different images

$$S_{\text{sidi}} = \{D(I_{i,c}, J_{j,k}) | I = J, i \neq j, c = k\}. \tag{13}$$

Different class - same image (S_{dcsi}). The set contains the comparison between instances of different classes within the same image

$$S_{\text{dcsi}} = \{D(I_{i,c}, J_{j,k}) | i = j, c \neq k\}. \tag{14}$$

Different class - different image (S_{dcdi}). In this set, each instance from an image is compared with all instances of different classes in different images.

$$S_{\text{dcdi}} = \{D(I_{i,c}, J_{j,k}) | i \neq j, c \neq k\} \tag{15}$$

Fig. 2 presents an example of the prototype extraction procedure from three images.

Using prototypes a, b, c, d, e, and f (in Fig. 2), the four distance sets are defined as follows:

$$S_{\text{scsi}} = \{D(c,d), D(e,f)\},$$

$$S_{\text{scdi}} = \{D(b,c), D(b,d), D(b,e), D(b,f), D(c,e), D(c,f)\}$$

$$S_{\text{dcsi}} = \{D(a,b)\}$$

$$S_{\text{dcdi}} = \{D(a,c), D(a,d), D(a,e), D(a,f)\}$$

$$S_{\text{sidi}} = \{D(c,e), D(d,f)\}$$

5. Experimental setup

This section describes our experimental setup, in which we explore the layers and conditions that better preserve the semantics of objects in the latent space of a diffusion U-Net. If the multimodal space of SDM reflected the semantic proximity of the objects, then objects belonging to the same class should be closer than those from different classes. In Section 5.1, we evaluate different distance measures to determine the most effective one. In Section 5.2, we analyze features extracted at

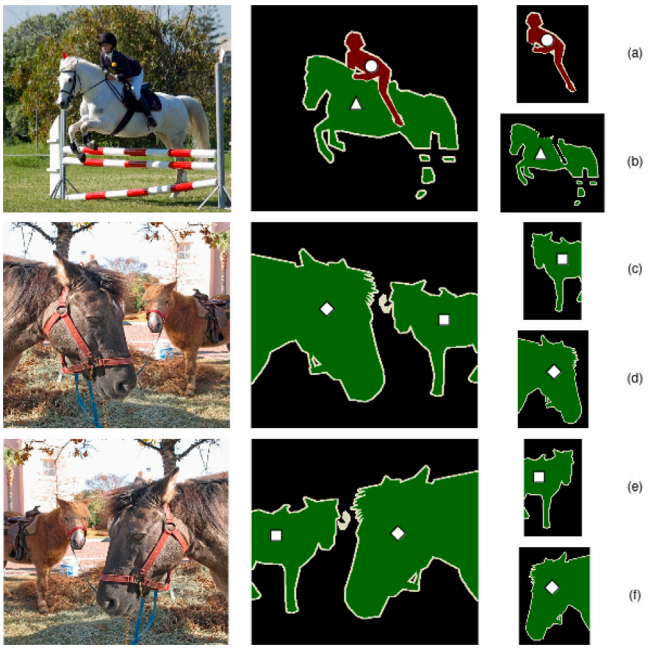


Fig. 2. Instance maps drive the feature extraction process. The objects in the image are depicted by representing their class with colors (*e.g. person* in red and *horse* in green) and their individual instances with symbols (*e.g.* ○, △, □, ◇). This information, provided by the instance segmentation label map, is used to extract a mask for each object in the image. After resizing, this mask is used to filter the feature maps and extract the object's prototype ((a), (b), (c), (d), (e), (f)). Top: An image with two objects of different classes. Middle: An image with two objects of the same class. Bottom: Flipped version of the Middle image.

different depths of the diffusion U-Net to identify which layer (or combination of layers) best retains semantic information. In Section 5.3, we examine the impact of the number of diffusion time (inference) steps on the extracted features. Additionally, in Section 5.4, we assess the effect of various image transformations. Finally, Section 5.5 investigates whether and how semantic information is maintained between consecutive frames in a video. For all the experiments, we leverage a publicly available implementation of SDM-v1.5.[2] All the images are resized to 512×512 before being input to the SDM.

5.1. Distance measure evaluation

In this experiment, we compare the distance measures introduced in Section 3.2 to find the one that best reflects the semantic distance of the objects. We generate the prototypes of all the objects in the Pascal-VOC and COCO subsets. In this phase, we consider the layers in the encoder and the decoder of the diffusion U-Net at the spatial resolutions of 64×64, 32×32, and 16×16. The features are picked from the SDM at the last inference step ($t = 0$), *i.e.*, the last denoising step. Thus, we obtain four distance sets: *Same Class - Same Image* (S_{scsi}), *Same Class - Different Image* (S_{scdi}), *Different Class - Same Image* (S_{dcsi}), and *Different Class - Different Image* (S_{dcdi}). The distribution of the four sets is analyzed using box plots and the p-value is used to evaluate the statistical significance of the pairwise comparison between sets. Additionally, Dunn Index is employed to analyze to measure the degree of separation between the sets.

[2] https://github.com/hkproj/pytorch-stable-diffusion.

5.2. U-Net features analysis

To evaluate the role of different layers we exploit the 1000 images from both Pascal-VOC and COCO and we extract the four distance sets (S_{scsi}, S_{scdi}, S_{dcsi} and S_{dcdi}) using the features extracted from different combinations of layers of the diffusion U-Net. As in the previous phase, the features are picked from the SDM at the last inference step ($t = 0$). Each set is characterized by the mean and standard deviation of all the per-object comparisons: for instance, we calculate the per-object distance in the *Same Class - Same Image* (S_{scsi}) setup for all the images; once we collected all the distances, we calculate their mean and standard deviation; finally, we perform pairwise comparisons between sets to verify that the distances between different objects are higher than the ones between the same object in different configurations. We aim to find the layer that produces well-separated sets of features and that reflects the semantic distance between objects. This translates into having different instances of the same class (*e.g.*, two different dogs) closer than instances of different classes (*e.g.*, a dog and a truck) in the same image or in distinct images. When the semantic is preserved we would observe the following relation between the sets:

$$\bar{S}_{scsi} < \bar{S}_{scdi} < \bar{S}_{dcsi} < \bar{S}_{dcdi} \tag{16}$$

where $\bar{S}$ denotes the mean of all the distances in a given set. To verify this desired trend, we analyze the distribution of the four sets using box plots and we employ the Kruskal–Wallis test to assess whether the results are statistically significant. Additionally, we also calculate the Dunn Index, which quantifies the degree of separation between the sets.

5.3. Inference steps evaluation

Diffusion models differ from traditional generative models because they generate images by learning to progressively remove noise. The denoising process is performed within a U-Net for a number of time steps T (from 1 to 1000): thus, it may be crucial to evaluate the impact of a different number of time steps on the internal representation of a diffusion U-Net. We perform feature extraction on the Pascal-VOC and COCO subsets at different denoising steps. The features are used to compute the four sets of pairwise distances: S_{scsi}, S_{scdi}, S_{dcsi} and S_{dcdi}. For each set of pairwise distances, we analyze the distribution using box plots. As before, we want to confirm that the relation in Eq. (16) is satisfied to evaluate the impact of noise on the features extracted from the diffusion U-Net. Additionally, the p-value is also calculated to evaluate the statistical significance of the distance sets.

5.4. Impact of image transformations

Diffusion models are trained on huge text–image datasets to create a multimodal space where text and images coexist. The vast amount of data enables them to deal with a large number of objects, while data complexity allows them to generate images with objects in a wide range of configurations. Hence, we wonder whether the learned internal object representation of the U-Net within a diffusion model is influenced by the following elementary transformations:

- Crop Padding – We crop the image around a specific object and padded it to its original size, aiming to understand the influence of context on the extracted features.
- Crop Resizing – Similar to the prior case, we trim around an entity, bilinearly resizing to the initial dimensions. This experiment examines the impact of changing the object scales on features.
- Crop Shift – We crop around an object and we randomly shift the bounding box by 20%, 40%, and 60% of its size. The image is cropped and padding is added to match the original image size. We evaluate how progressive occlusion affects the extracted features.
- Flip – The image is horizontally and vertically flipped.

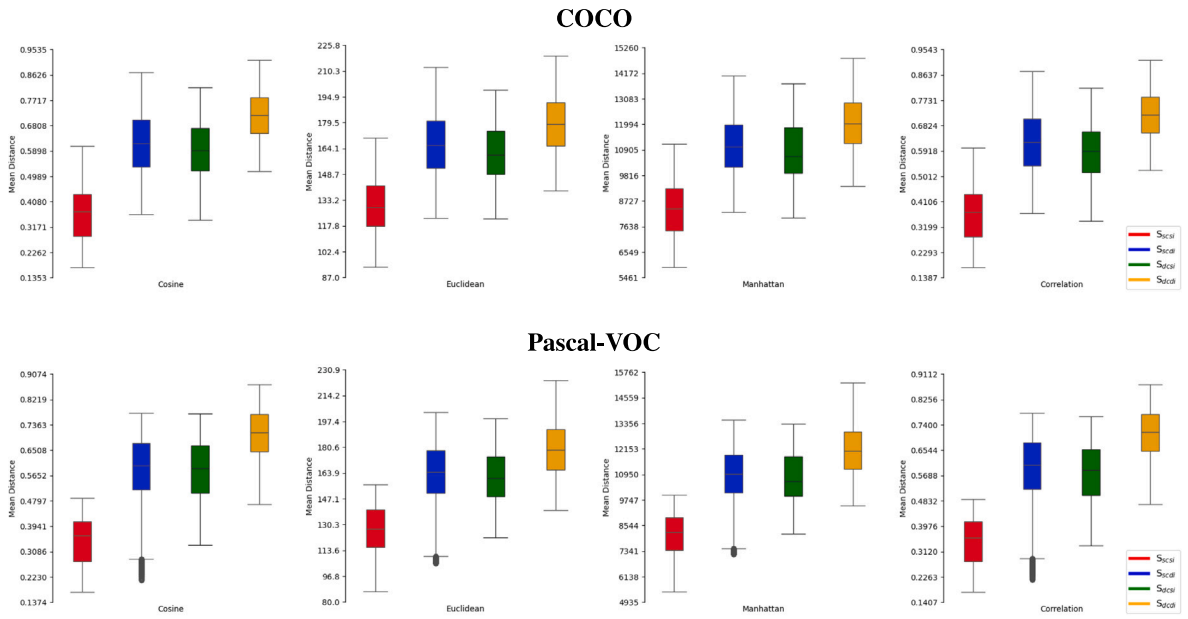


Fig. 3. Distance selection. Each box plot reports the distribution of the distances calculated on the following sets: *Same Class - Same Image* (S_{scsi}, red), *Same Class - Different Image* (S_{scdi}, blue), *Different Class - Same Image* (S_{dcsi}, green and *Different Class - Different Image* (S_{dcdi}, yellow). The evaluation is separately performed on COCO (top) and Pascal-VOC (bottom).

- Rotation – The image is rotated by 90°, 180°, and 270°.

We apply these transformations to images from the Pascal-VOC and COCO datasets. To investigate the influence on object representation, we compare the features extracted from an image before and after applying a transformation. This produces three sets of distances: S_{scsi}, S_{dcsi} and S_{sidi}. If the transformations did not affect the internal representation of the features, we would obtain a very low value for $\bar{S}_{\text{sidi}}$, which represents the mean distances between instances of the same object before and after the transformation. In this scenario, the desired outcome is for $\bar{S}_{\text{sidi}}$ to be smaller than the others, as in the following relation:

$$\bar{S}_{sidi} < \bar{S}_{scsi} < \bar{S}_{dcsi} \tag{17}$$

We analyze the distribution of each set using box plots to evaluate the impact of the transformations on semantic similarity. The statistical significance of the distance sets is determined by computing the relevant p-values.

5.5. Object's features in subsequent video frames

In this study, we explore if semantic information about an object remains consistent across subsequent frames in a video. We use videos from the DAVIS dataset containing instances of *person* class. The features are extracted from a reference frame and from subsequent frames at offsets 1, 2, 3, 4, 5, 10, 20, 30, and 40. For each offset, we calculated two sets of distances: S_{scdi}, S_{sidi} capturing the differences between the features of the reference frame and the features of subsequent frames. Additionally, we compute the S_{scsi} set, which compares features within a single frame. For each set, we analyze the distribution of the four sets using the box plots, and the *p*-value is used to evaluate semantic consistency across distant frames.

6. Results

The results of the comparison between different distance measures are presented in Section 6.1. In Section 6.2, we conduct an ablation study to investigate whether the U-Net layers within the SDM effectively preserve semantic information. Section 6.3 evaluates the influence of using different inference steps on the SDM latent space. Finally, Section 6.4 presents the results of applying different image transformations, while Section 6.5 details the analysis of features extracted from different video frames.

6.1. Distance measure evaluation

Following the experimental setup described in Section 5.1, we calculate the four sets of distances (S_{scsi}, S_{scdi}, S_{dcsi} and S_{dcdi}). Our goal is to identify the distance measure that best separates features between classes (S_{scsi} vs. S_{dcsi}, S_{scsi} vs. S_{dcdi}, S_{scdi} vs. S_{dcsi}, S_{scdi} vs. S_{dcdi} and S_{dcsi} vs. S_{dcdi}) while minimizing the distances related to the same class (S_{scsi} vs. S_{scdi}). In Fig. 3 the box plots of the distribution of the four sets using different distance measures are presented. As we can observe, all the measures produce a similar separation between the sets. However, the correlation and cosine distances appear to better differentiate the S_{scsi} and the S_{dcdi} sets. This is further confirmed by the Dunn index reported in Table 1 where the correlation distance consistently leads to a higher value in most cases. While other metrics may occasionally outperform it by a small margin, the correlation distance generally achieves comparable performance (the p-values confirm the statistical significance of these results). We, therefore, chose the correlation distance as the evaluation metric from now on. In Table S.1 of the supplementary material we also report the numerical values of the mean and the standard deviation of four sets of distances.

6.2. U-Net features analysis

Features extracted from different layers are compared employing the experimental setup described in Section 5.2: we compare different spatial resolutions from the encoder and the decoder, and from specific layers. We discuss the results for each case study in the following.

Selection of the spatial resolution. To find the most convenient diffusion U-Net layer(s) for feature extraction in terms of semantics preservation, we compare features at different spatial resolutions (16 × 16, 32 × 32, 64 × 64), within both the encoder and decoder. In particular, we perform feature extraction from the following (combination of) layers: 64 + 32 + 16, 64, 32, and 16. Notice that, if more than a resolution is employed (*e.g.*, in 64 + 32 + 16), we concatenate the features at different resolutions along the channel dimension. The box plots in

Table 1
Distance comparison. Dunn Index and p-value comparison between distance sets using the two datasets subsets.

Sets	Metric	Pascal-VOC				COCO 2017			
		D_{euc}	D_{cor}	D_{man}	D_{cos}	D_{euc}	D_{cor}	D_{man}	D_{cos}
S_{scsi} vs. S_{scdi}	p-value	3.1×10^{-32}	1.3×10^{-33}	$\mathbf{1.2 \times 10^{-38}}$	3.7×10^{-33}	1.4×10^{-45}	$\mathbf{2.3 \times 10^{-53}}$	9.6×10^{-51}	2.3×10^{-52}
	Dunn Index	0.30312	0.33343	**0.41411**	0.33027	0.35941	**0.50012**	0.39991	0.49589
S_{scsi} vs. S_{dcsi}	p-value	1.3×10^{-20}	9.2×10^{-22}	$\mathbf{2.2 \times 10^{-23}}$	9.5×10^{-22}	5.6×10^{-17}	4.0×10^{-21}	7.0×10^{-20}	$\mathbf{4.5 \times 10^{-21}}$
	Dunn Index	0.65856	0.80092	**0.80684**	0.78081	0.37309	0.59327	0.42288	**0.59863**
S_{scsi} vs. S_{dcdi}	p-value	1.9×10^{-52}	$\mathbf{2.0 \times 10^{-54}}$	4.4×10^{-54}	3.3×10^{-54}	2.3×10^{-60}	$\mathbf{2.8 \times 10^{-70}}$	1.6×10^{-64}	7.4×10^{-70}
	Dunn Index	0.60352	0.79657	0.66042	**0.79998**	0.48857	**0.67989**	0.56369	0.67637
S_{scdi} vs. S_{dcsi}	p-value	0.02742	0.00028	0.01164	**0.00012**	0.00516	**0.00195**	0.01620	0.00816
	Dunn Index	0.04510	0.08312	0.06120	**0.08946**	0.07081	**0.08545**	0.05874	0.07451
S_{scdi} vs. S_{dcdi}	p-value	0	0	0	0	0	0	0	0
	Dunn Index	0.23696	**0.29406**	0.27972	0.29364	0.11407	0.19403	0.13810	**0.19816**
S_{dcsi} vs. S_{dcdi}	p-value	9.1×10^{-18}	1.5×10^{-18}	1.0×10^{-19}	$\mathbf{2.0 \times 10^{-17}}$	3.8×10^{-15}	$\mathbf{5.4 \times 10^{-27}}$	5.6×10^{-16}	1.4×10^{-25}
	Dunn Index	0.21439	**0.26777**	0.22345	0.26179	0.19077	**0.27705**	0.20624	0.27002

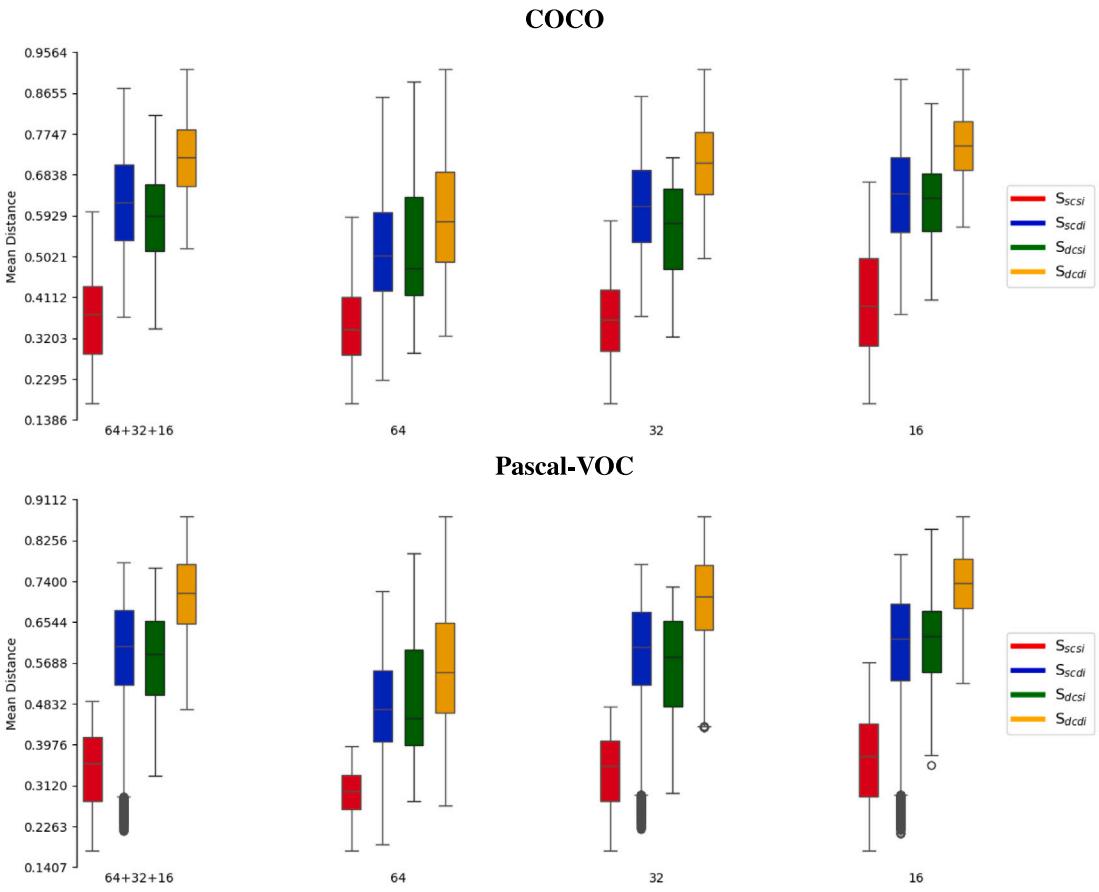


Fig. 4. Semantics preservation at different spatial resolutions of the diffusion U-Net. Each box plot reports the distribution of the distances calculated on the following sets: *Same Class - Same Image* (S_{scsi}, red), *Same Class - Different Image* (S_{scdi}, blue), *Different Class - Same Image* (S_{dcsi}, green and *Different Class - Different Image* (S_{dcdi}, yellow). The evaluation is separately performed on COCO (top) and Pascal-VOC (bottom).

Fig. 4 summarize the distribution of the four sets of distances (S_{scsi}, S_{scdi}, S_{dcsi} and S_{dcdi}) on Pascal-VOC and COCO. In the supplementary material, the corresponding numerical values are reported in Table S.2 and the statistical significance of the distributions is given by the p-values in Table S.3.

In the Pascal-VOC dataset, we observe that the desired trend (described in Eq. (16)) is respected by the features extracted from all the layers. However, the features extracted from COCO do not allow to distinguish between objects of the same class in different images and objects of different classes in the same image: for instance, we may not be able to distinguish a dog from a truck in the same image (green boxes in Fig. 4), or two dogs in different images (blue boxes in Fig. 4). This could suggest that the features extracted from these layers are influenced also by the context of an image bringing the representation of objects of different classes closer. A possible explanation for this behavior may lay in the fact that (a) COCO is a more complex dataset (the scene usually contains many objects with smaller dimensions) than Pascal-VOC and (b) internal representation may be influenced by the context.

However, 16×16 layers exhibit the best separation ability on Pascal-VOC and the lowest difference between $\bar{S}_{scdi}$ and $\bar{S}_{dcsi}$ in the COCO

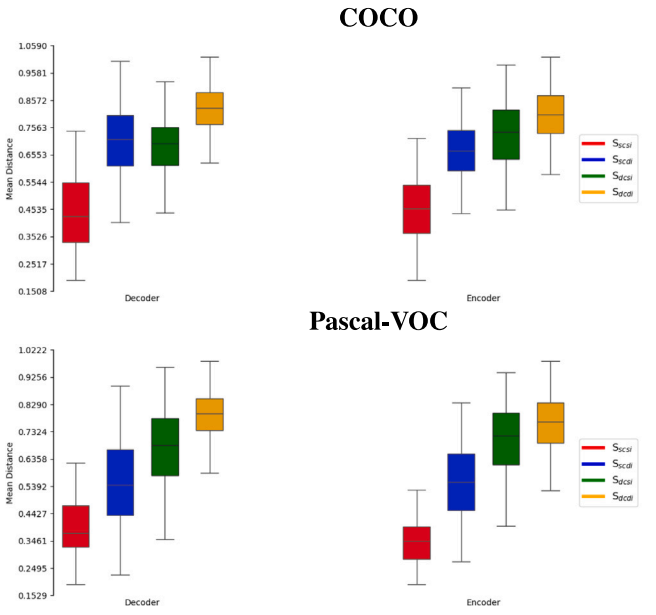


Fig. 5. Semantics preservation in the encoder and the decoder of the diffusion U-Net. Each box plot reports the distribution of the distances calculated on the following sets: *Same Class - Same Image* (S_{scsi}, red), *Same Class - Different Image* (S_{scdi}, blue), *Different Class - Same Image* (S_{dcsi}, green and *Different Class - Different Image* (S_{dcdi}, yellow). The evaluation is separately performed on COCO (top) and Pascal-VOC (bottom).

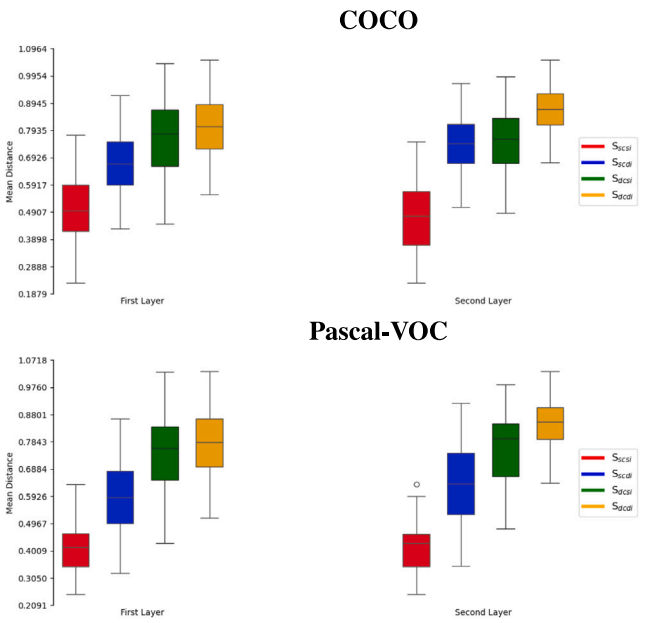


Fig. 6. Semantics preservation in the 1st and 2nd layer of the diffusion U-Net at resolution **16×16**. Each box plot reports the distribution of the distances calculated on the following sets: *Same Class - Same Image* (S_{scsi}, red), *Same Class - Different Image* (S_{scdi}, blue), *Different Class - Same Image* (S_{dcsi}, green and *Different Class - Different Image* (S_{dcdi}, yellow).

dataset. This finding encourages us to delve deeper into the analysis of these specific layers.

Encoder features vs. Decoder features. The diffusion U-Net is composed of an Encoder and a Decoder. The Encoder reduces the image resolution while increasing the number of features; the Decoder then restores the features to the original image resolution. The layers at resolution 16×16 are analyzed by collecting features from the Encoder and the Decoder separately. Fig. 5 shows the distribution of the distances within each set (S_{scsi}, S_{scdi}, S_{dcsi} and S_{dcdi}). Interestingly, the layers of the Encoder at resolution 16×16 seem to better express object semantics: in both Pascal-VOC and COCO datasets we observe the desired trend of Eq. (16). In Table S.4 of the supplementary material, we report the corresponding mean and standard deviation of the distances in the four sets (S_{scsi}, S_{scdi}, S_{dcsi} and S_{dcdi}). To assess the statistical significance of the distributions we also compute the pairwise p-values (Table S.5 in the supplementary material). Since diffusion U-Net has two encoder layers at resolution 16×16, we continue our analysis further to determine the extent to which each layer contributes to semantic preservation.

Single encoder layer evaluation. We focus on the features collected from the 1st Layer and the 2nd Layer of the encoder having a spatial resolution of 16×16, separately. In Fig. 6 we plot the distribution of the distances within each set (S_{scsi}, S_{scdi}, S_{dcsi} and S_{dcdi}). Our analysis of the 16×16 encoder layers reveals a significant difference between the features extracted from the first and second layers. The first layer reliably captures the desired trend in the data (Eq. (16)). Notably, the gap between the distances of objects belonging to different classes in the same image ($\bar{S}_{dcsi}$) and objects of different classes in different images ($\bar{S}_{dcdi}$) is smaller than the distances of objects of the same class in different images ($\bar{S}_{scdi}$). For instance, the distribution of *cat and dog in the same image* is closer to the distribution of *cat and dog in different images* than to the distribution of *cats in different images*. Additionally, S_{scsi} and S_{scdi} are closer than using all the Encoder features, *i.e.* intra-class similarity is better preserved. In the supplementary material, we report the mean and the standard deviation of each set in each setup (Table S.4) and also the p-values and the Dunn Indexes (Table S.7).

Using the first layer we not only obtain statistically significant p-values ($p < 0.05$) for all the compared subsets, but also the Dunn Index shows a clear separation between sets of features corresponding to different classes.

The second layer, on the other hand, barely respects the desired trend (Eq. (16)) on the COCO dataset. Indeed, using the second layer we obtain a non-significant p-value (0.36489) for the comparison between objects of the same class in different images and objects of different classes in the same image (S_{scdi} vs. S_{dcsi}) also in terms of the corresponding Dunn Index (S_{scdi} vs. S_{dcsi} returns 0.01823 for the second layer and 0.14350 for the first layer). This indicates that the features extracted from the second layer do not allow to distinguish objects of the same class from objects of different classes.

6.3. Impact of inference steps

Following the procedure in Section 5.3, we investigate if and how the number of inference steps of the diffusion U-Net affects its features representation. To this aim, we extract the 16×16 features from the first layer of the encoder given a different number of inference steps (1, 50, 100, 200, 300, 400, and 500). The expected result is that after a certain number of inference steps the features cannot retain the semantics and Eq. (16) does not hold anymore.

Fig. 7 plots the distribution of the four sets of distances: S_{scsi}, S_{scdi}, S_{dcsi}, S_{dcdi}. As expected, adding even a small amount of noise ($t = 50$) to the image results in corrupted features. In the supplementary material, we report the mean and the standard deviation of each set at different inference steps (Table S.8) and the corresponding p-values (Table S.9).

6.4. Impact of image transformations

To investigate the impact of image transformation on the diffusion features, we employed the experimental setup detailed in Section 5.4, extracting the feature from the first encoder layer at resolution 16×16. For each transformation, we compute the distances between objects of

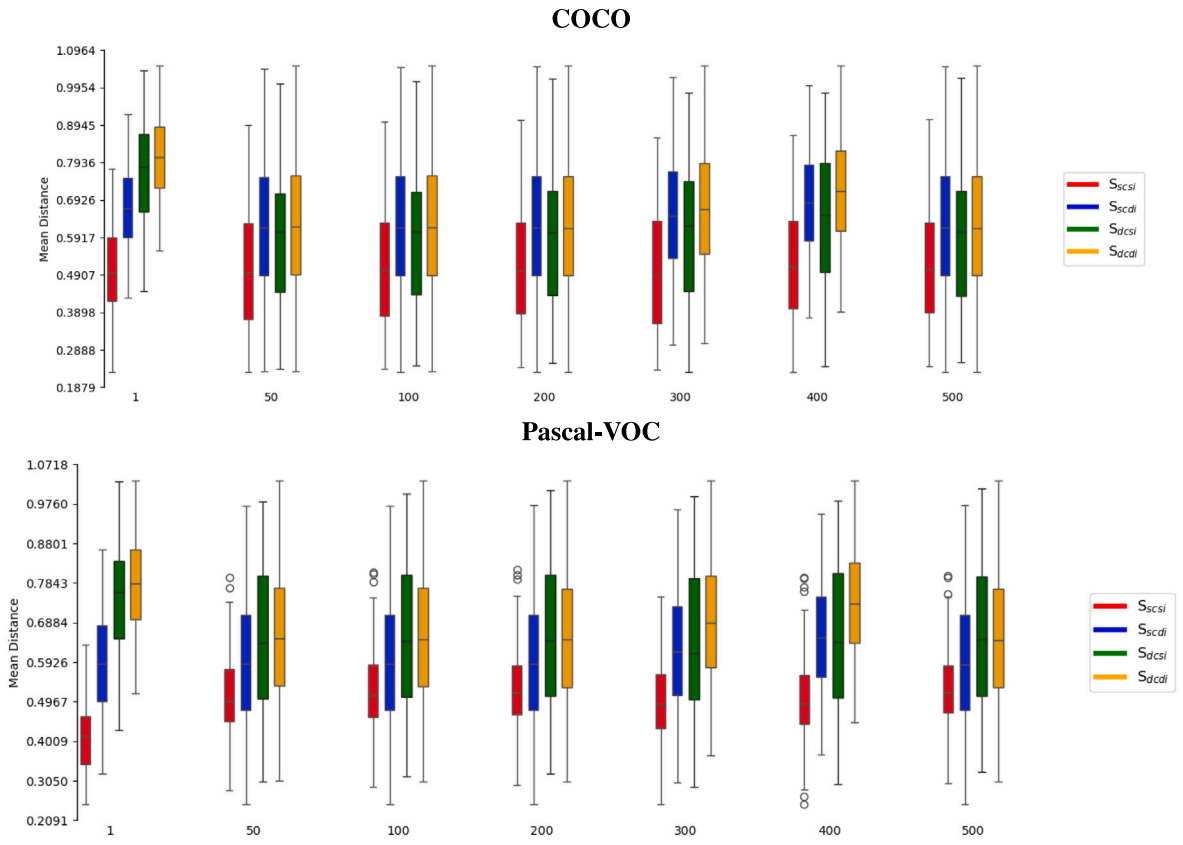


Fig. 7. Semantics preservation at different inference steps. Each box plot reports the distribution of the distances calculated on the following sets: *Same Class - Same Image* (S_{scsi}, red), *Same Class - Different Image* (S_{scdi}, blue), *Different Class - Same Image* (S_{dcsi}, green and *Different Class - Different Image* (S_{dcdi}, yellow). The evaluation is separately performed on COCO (top) and Pascal-VOC (bottom).

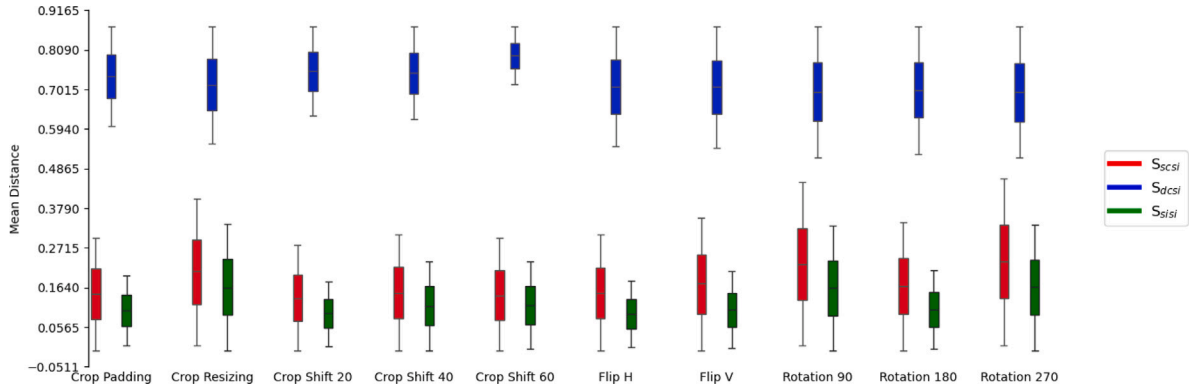


Fig. 8. Semantics preservation after various image transformations. Each box plot reports the distribution of the distances calculated on the following sets: *Same Class - Same Image* (S_{scsi}, red), *Different Class - Same Image* (S_{dcsi}, blue), *Same Instance - Different Image* (S_{sidi}, green). The evaluation is separately performed on COCO (top) and Pascal-VOC (bottom).

the same class from the same image and from different images (S_{scsi}, S_{scdi}). Moreover, we assess the impact of simple image transformations on the features of a given object. To do this, we compute the subset S_{sidi}, which contains the distances between the same instance in different images (the original image and its transformed version). To evaluate if the trend in Eq. (17) is respected after the transformations we use the box plots of the distribution of the three subsets in Fig. 8.

The higher mean values of the blue boxes suggest that, for all the types of transformations, it is always possible to distinguish objects of one class from the other classes.

However, not all transformations are capable of maintaining feature representations that can distinguish the same instance of an object from other objects of the same class. In particular, for each single transformation, we can make the following observations.

- Image Crop – When cropping (Crop Padding) an image, an instance of a class remains barely distinguishable from other instances of the same class within the same image.
- Image Resizing – When scaling an image (Crop Resizing), the average distance between the same instance and different instances of the same class becomes almost indistinguishable making it difficult to recognize features extracted from the same instance of an object.
- Image Occlusion (Crop Shift) – As expected, occlusion affects the ability to recognize objects of the same instance as the percentage of occlusion increases.

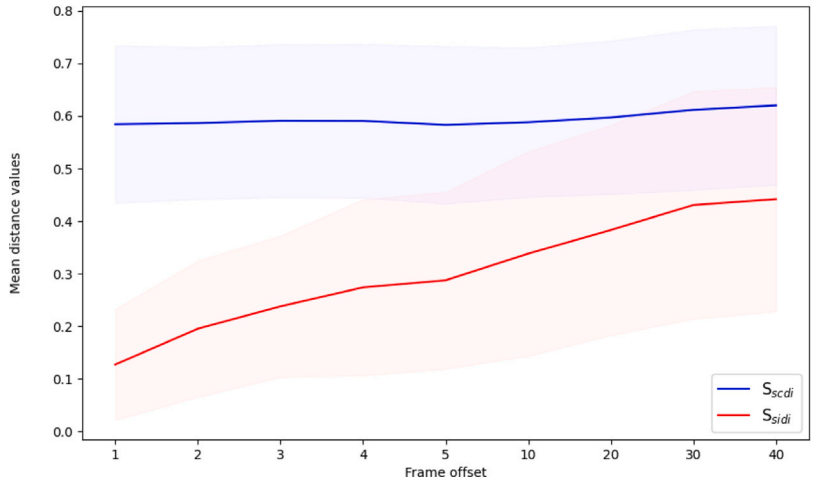


Fig. 9. Features distances across subsequent video frames Trend of the means and the standard deviations of the distance between objects in subsequent frames. *Same Instance - Different Image* (S_{sidi}, red), and *Same Class - Different Image* (S_{scdi}, blue).

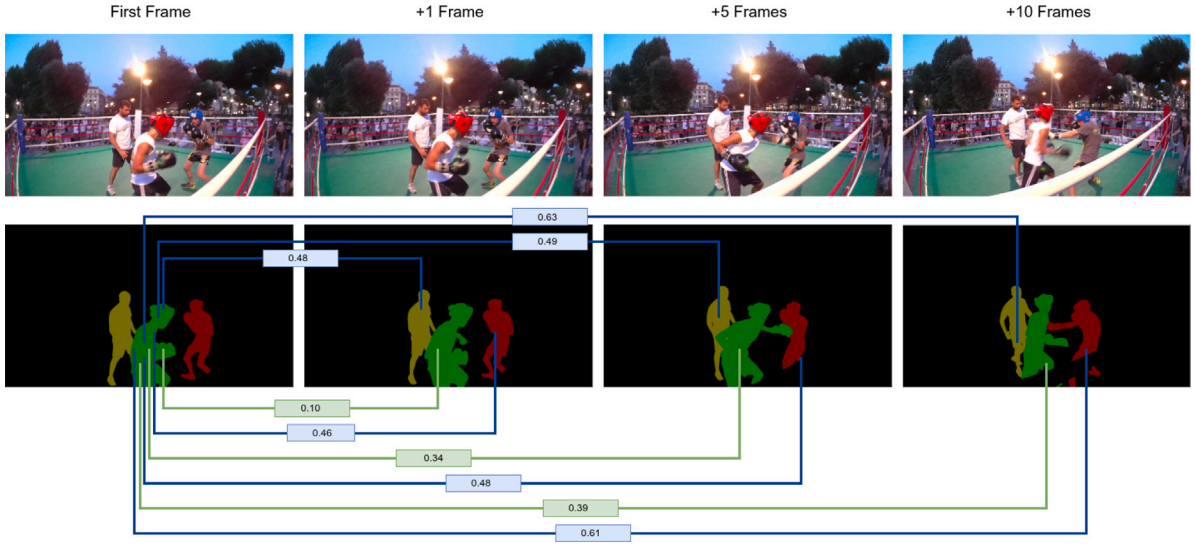


Fig. 10. Example of distances computed on different frames. In green the correlation distances between the same instance of a person in various frames of the video. In blue the distances between other people in the same video over multiple frames.

- Image Flip (Horizontal/Vertical) – Both Horizontal and Vertical flipping do not guarantee preservation of the representation. Although the means are distinct, the box plot shows a significant overlap between the distributions.
- Image Rotation – Rotating the image at 90°, 180°, and 270°, results in a loss of the ability to recognize the same instance of an object.

In the supplementary material, we report the mean and the standard deviation of each set at different inference steps (Table S.10) and the corresponding p-values (Table S.11).

6.5. Object's features in subsequent video frames

We have observed that, under various conditions and setups, the extracted features effectively retain semantic information. Now, we aim to investigate what happens when using a set of related but distinct images, such as those in a video. In our experiment, we keep a frame as a reference, and we measure the distances between object instances in the latent space over a range of +1 to +40 frames. In Fig. 9, it can be observed that the distance between different objects of the same class remains almost constant and it is not affected by the offset. Indeed, the mean value of S_{scdi} (blue line) is close to the distance computed in the same image ($\bar{S}_{scsi} = 0.58$). Instead, the distance between individual instances of the same object increases over successive frames (red line). This phenomenon could also be linked to the results obtained using transformations. In fact, after several frames, an object can undergo occlusions or configuration changes (*e.g.*, an object moving away from the camera appears at a smaller scale in the image), which we have demonstrated are types of transformations that gradually modify the internal representation of an object. Quantitative results and the p-values are reported in Table S.12 and in Table S.13, respectively.

In Fig. 10 we also report an example of distances calculated using the proposed method between a person (in green) and the people in the scene (comprising the same person) after several frames (+1, +5 and +10 frames). As we can observe, the distance between the same person is consistently smaller than the others. This result suggests that instances of the same object, if the configuration changes are not excessive, maintain their internal representation within several frames in the same video. This opens up new research opportunities

in tasks that go beyond generation, such as video surveillance, object re-identification, and tracking.

6.6. Discussion

Our experiments suggest that the first encoder layer at a resolution of 16×16 effectively preserves semantic information, demonstrating consistent results. We further explore how different numbers of inference steps and various transformations impact the internal feature representation of the diffusion U-Net. Our findings reveal that, while the first layer captures discriminative features, increasing the number of inference steps and applying transformations can hinder the model's ability to consistently recognize the same object instance.

For classification tasks, based on our results, we observe that even with substantial changes to the image (e.g., context or object appearance), the nature of the features remains preserved, and the classes remain distinguishable. However, when using the internal representation for instance-based tasks, such as identifying an object across successive frames, the object representation is preserved only for non-substantial transformations.

7. Conclusion

SDMs demonstrate surprising generalization capabilities, enabling their application to tasks beyond the image generation for which they were originally trained. Numerous studies have employed these models for downstream tasks in zero-shot setups. This raises the need for a deeper understanding of their internal properties to optimize their performance without retraining. This paper presents a novel approach to investigate the properties of Stable Diffusion's internal representations. We leverage the model's ability to maintain locality in its internal features, which allows generating prototypes for objects in a scene based on their corresponding segmentation masks. We then compare these prototypes by calculating their distances within the diffusion U-Net feature space. By examining various network layers, we aim to pinpoint those that best maintain object semantics. Ideally, objects with similar semantics should have distances in the feature space that are proportional to their similarity. Our experiments suggest that the first encoder layer at resolution 16×16 effectively preserves semantic information. Furthermore, we investigate if and how different amounts of inference steps and various kinds of transformations affect the internal feature representation of the diffusion U-Net. We show that increasing the number of inference steps (i.e., adding more noise) and applying various transformations to objects (e.g., resizing and rotating the images) adversely impact the preservation of the internal object representation. These results could open new research opportunities in using SDMs in tasks beyond image generation such as image segmentation and object recognition using zero-shot or few-shot approaches. Finally, we observe that the internal representation of an object remains consistent across several frames in the same video, provided the configuration changes are not excessive. This result provides an insightful indication that SDMs can be effectively used for tasks such as object re-identification, tracking, and video surveillance.

CRediT authorship contribution statement

Simone Bonechi: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Paolo Andreini:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Barbara Toniella Corradini:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Franco Scarselli:** Writing – review & editing, Validation, Supervision, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.neucom.2024.128846.

Data availability

Only public data is used for this study.

References

[1] H. Gong, J. Su, K.P. Seng, A. Nguyen, A. Liu, H. Liu, Film-GAN: towards realistic analog film photo generation, Neural Comput. Appl. 36 (8) (2024) 4281–4291.

[2] N.P. Toliya, N.B. Chadaga, R. Harshitha, M. Dhruva, N. Nagarathna, GAN based model for virtual try on of clothes, in: 2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications, ICETCS, IEEE, 2024, pp. 1–7.

[3] Y. Chen, H. Lin, W. Zhang, W. Chen, Z. Zhou, A.A. Heidari, H. Chen, G. Xu, ICycle-GAN: Improved cycle generative adversarial networks for liver medical image generation, Biomed. Signal Process. Control 92 (2024) 106100, http://dx.doi.org/10.1016/j.bspc.2024.106100, URL https://www.sciencedirect.com/science/article/pii/S1746809424001587.

[4] G. Ciano, P. Andreini, T. Mazzierli, M. Bianchini, F. Scarselli, A multi-stage GAN for multi-organ chest X-ray image generation and segmentation, Mathematics 9 (22) (2021).

[5] A. Beers, J. Brown, K. Chang, J.P. Campbell, S. Ostmo, M.F. Chiang, J. Kalpathy-Cramer, High-resolution medical image synthesis using progressively grown generative adversarial networks, 2018, arXiv preprint arXiv:1805.03144.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[7] D.P. Kingma, M. Welling, Auto-encoding variational bayes, 2013, arXiv preprint arXiv:1312.6114.

[8] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, 2015, arXiv preprint arXiv:1511.06434.

[9] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 214–223.

[10] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019.

[11] A. Brock, J. Donahue, K. Simonyan, Large scale GAN training for high fidelity natural image synthesis, in: International Conference on Learning Representations, 2018.

[12] A. Razavi, A. Van den Oord, O. Vinyals, Generating diverse high-fidelity images with vq-vae-2, Adv. Neural Inf. Process. Syst. 32 (2019).

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[14] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.

[15] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, A. Komatsuzaki, Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021, arXiv preprint arXiv:2111.02114.

[16] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al., Laion-5b: An open large-scale dataset for training next generation image-text models, Adv. Neural Inf. Process. Syst. 35 (2022) 25278–25294.

[17] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, Adv. Neural Inf. Process. Syst. 30 (2017).

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[19] Z. Shi, X. Zhou, X. Qiu, X. Zhu, Improving image captioning with better use of captions, 2020, arXiv preprint arXiv:2006.11807.

[20] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E.L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al., Photorealistic text-to-image diffusion models with deep language understanding, Adv. Neural Inf. Process. Syst. 35 (2022) 36479–36494.

[21] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684–10695.
[22] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
[23] U. Bhalla, A. Oesterling, S. Srinivas, F.P. Calmon, H. Lakkaraju, Interpreting CLIP with sparse linear concept embeddings (SpLiCE), 2024, arXiv preprint arXiv:2402.10376.
[24] A.C. Li, M. Prabhudesai, S. Duggal, E. Brown, D. Pathak, Your diffusion model is secretly a zero-shot classifier, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2206–2217.
[25] S. Mukhopadhyay, M. Gwilliam, V. Agarwal, N. Padmanabhan, A. Swaminathan, S. Hegde, T. Zhou, A. Shrivastava, Diffusion models beat gans on image classification, 2023, arXiv preprint arXiv:2307.08702.
[26] K. Clark, P. Jaini, Text-to-image diffusion models are zero shot classifiers, Adv. Neural Inf. Process. Syst. 36 (2024).
[27] L. Karazija, I. Laina, A. Vedaldi, C. Rupprecht, Diffusion models for zero-shot open-vocabulary segmentation, 2023, arXiv preprint arXiv:2306.09316.
[28] R. Burgert, K. Ranasinghe, X. Li, M.S. Ryoo, Peekaboo: Text to image diffusion models are zero-shot segmentors, 2022, arXiv preprint arXiv:2211.13224.
[29] J. Tian, L. Aggarwal, A. Colaco, Z. Kira, M. Gonzalez-Franco, Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 3554–3563.
[30] S. Bonechi, P. Andreini, B.T. Corradini, F. Scarselli, DiffProps: Is semantics preserved within a diffusion model?, in: Procedia Computer Science, 246C, 2024, pp. 5231–5240.
[31] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.
[32] M. Everingham, L. Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.
[33] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, L. Van Gool, The 2017 davis challenge on video object segmentation, 2017, arXiv preprint arXiv:1704.00675.
[34] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, Deep unsupervised learning using nonequilibrium thermodynamics, in: International Conference on Machine Learning, PMLR, 2015, pp. 2256–2265.
[35] T. Chen, On the importance of noise scheduling for diffusion models, 2023, arXiv preprint arXiv:2301.10972.
[36] A.Q. Nichol, P. Dhariwal, Improved denoising diffusion probabilistic models, in: International Conference on Machine Learning, PMLR, 2021, pp. 8162–8171.
[37] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, 2020, arXiv preprint arXiv:2010.02502.
[38] J. Ho, T. Salimans, Classifier-free diffusion guidance, 2022, arXiv:2207.12598.
[39] Y. Hao, Z. Chi, L. Dong, F. Wei, Optimizing prompts for text-to-image generation, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, Vol. 36, Curran Associates, Inc., 2023, pp. 66923–66939, URL https://proceedings.neurips.cc/paper_files/paper/2023/file/d346d91999074dd8d6073d4c3b13733b-Paper-Conference.pdf.
[40] C. Yu, J. Peng, X. Zhu, Z. Zhang, Q. Tian, Z. Lei, Seek for incantations: Towards accurate text-to-image diffusion synthesis through prompt engineering, 2024, arXiv preprint arXiv:2401.06345.
[41] D. Baranchuk, I. Rubachev, A. Voynov, V. Khrulkov, A. Babenko, Label-efficient semantic segmentation with diffusion models, 2021, arXiv preprint arXiv:2112.03126.
[42] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, S. De Mello, Open-vocabulary panoptic segmentation with text-to-image diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2955–2966.
[43] K. Pnvr, B. Singh, P. Ghosh, B. Siddiquie, D. Jacobs, Ld-znet: A latent diffusion approach for text-based image segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4157–4168.
[44] M. Kwon, J. Jeong, Y. Uh, Diffusion models already have a semantic latent space, 2022, arXiv preprint arXiv:2210.10960.
[45] Y.-H. Park, M. Kwon, J. Choi, J. Jo, Y. Uh, Understanding the latent space of diffusion models through the lens of riemannian geometry, Adv. Neural Inf. Process. Syst. 36 (2024).
[46] S. Luccioni, C. Akiki, M. Mitchell, Y. Jernite, Stable bias: Evaluating societal representations in diffusion models, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, Vol. 36, Curran Associates, Inc., 2023, pp. 56338–56351, URL https://proceedings.neurips.cc/paper_files/paper/2023/file/b01153e7112b347d8ed54f317840d8af-Paper-Datasets_and_Benchmarks.pdf.
[47] L. Zhang, A. Rao, M. Agrawala, Adding conditional control to text-to-image diffusion models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3836–3847.
[48] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, A. Sorkine-Hornung, A benchmark dataset and evaluation methodology for video object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 724–732.
[49] G.J. Székely, M.L. Rizzo, Distance correlation: a measure for dependence between multivariate random variables, Ann. Statist. 35 (6) (2007) 2769–2792.
[50] E.F. Krause, Taxicab geometry, Math. Teach. 66 (8) (1973) 695–706.
[51] J.C. Dunn, Well-separated clusters and optimal fuzzy partitions, J. Cybern. 4 (1) (1974) 95–104.
[52] W.H. Kruskal, W.A. Wallis, Use of ranks in one-criterion variance analysis, J. Am. Stat. Assoc. 47 (260) (1952) 583–621.
[53] E.R. Girden, ANOVA: Repeated Measures, vol. 84, Sage, 1992.
[54] S.S. Shapiro, M.B. Wilk, An analysis of variance test for normality (complete samples), Biometrika 52 (3–4) (1965) 591–611, http://dx.doi.org/10.1093/biomet/52.3-4.591.
[55] B. Yang, C. Liu, B. Li, J. Jiao, Q. Ye, Prototype mixture models for few-shot semantic segmentation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16, Springer, 2020, pp. 763–778.
[56] K. Wang, J.H. Liew, Y. Zou, D. Zhou, J. Feng, Panet: Few-shot image semantic segmentation with prototype alignment, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9197–9206.

Simone Bonechi graduated in Computer Science at the Department of Information Engineering and Mathematical Sciences of the University of Siena (2014) where he then obtained his PhD. In 2018 he spent a period as a visiting Ph.D. at the University of Copenhagen. After two years of PostDoc, first at the University of Tuscia, and then at the University of Pisa, he is now a researcher at the Department of Social, Political and Cognitive Sciences of the University of Siena. His research activity is focused on Deep Learning and Artificial Intelligence, with particular reference to computer vision, image processing, and image generation. He is the author of over 30 publications in international journals and conference proceedings. He also carries out editorial activities as Associated Editor for the journal Neurocomputing and he is Guest Editor for the Special Issue "Mathematical Modelling and Machine Learning Methods for Bioinformatics and Data Science Applications II" for the journal Mathematics.

Paolo Andreini is a postdoc at the University of Siena (Italy), he has a degree in computer science and a Ph.D. in Information Engineering and Science. His research activity focuses on machine learning and computer vision. He has been involved in different projects, often cooperating with clinicians and industries, with particular application to the automatic analysis of biomedical images.

His main research topics include semantic segmentation, object detection, bioinformatics, weakly-supervised and unsupervised learning and generative models. From a practical point of view, he worked on a variety of domains including agar plates, retinal images, skin lesions, brain NMR, oocites images, kidney histological images, text detection and recognition, ribo-seq profiles etc.

Barbara Toniella Corradini is a Ph.D. candidate in Smart Computing at the University of Siena and the University of Florence. Her research focuses on Deep Learning techniques for Computer Vision, with an emphasis on studying image-generative foundation models, ranging from GANs to more recent multimodal generative models such as Diffusion Models. Her work has been published in AI and computer vision conferences. In addition to her research, she has been involved in academic collaborations and actively contributes as a reviewer for sector-specific conferences and journals.

Prof. **Franco Scarselli** received the Laurea degree with honors in Computer Science from the University of Pisa, Italy, and the PhD degree in Computer Science and Automation Engineering from the University of Florence. From 1999, he is at the University of Siena, where he is currently full professor at Department of Information Engineering and Mathematics. Franco Scarselli is associate editor of IEEE Transactions on Neural Networks and Learning Systems and has been involved in the organization od several conferences and workshops on machine learning. Moreover, he has been involved in more than 30 research projects focused on machine learning and information retrieval, founded by the Italian Ministry of Education, by the Australian Research Council, by the University of Siena, and private companies. The research of Franco Scarselli is in the field of machine learning, with a particular focus on graph neural networks, deep learning, and generative models. His applicative interests include computer vision, bioinformatics, and information retrieval. His main contributions include the proposal of the graph neural network model, a theory explaining the advances of deep leaning architectures v.s. swallow ones, and studies on approximation and generalization capability of feedforward neural networks and neural networks for graphs.