

Wizualizowanie monosemantycznych cech w modelach dyfuzyjnych

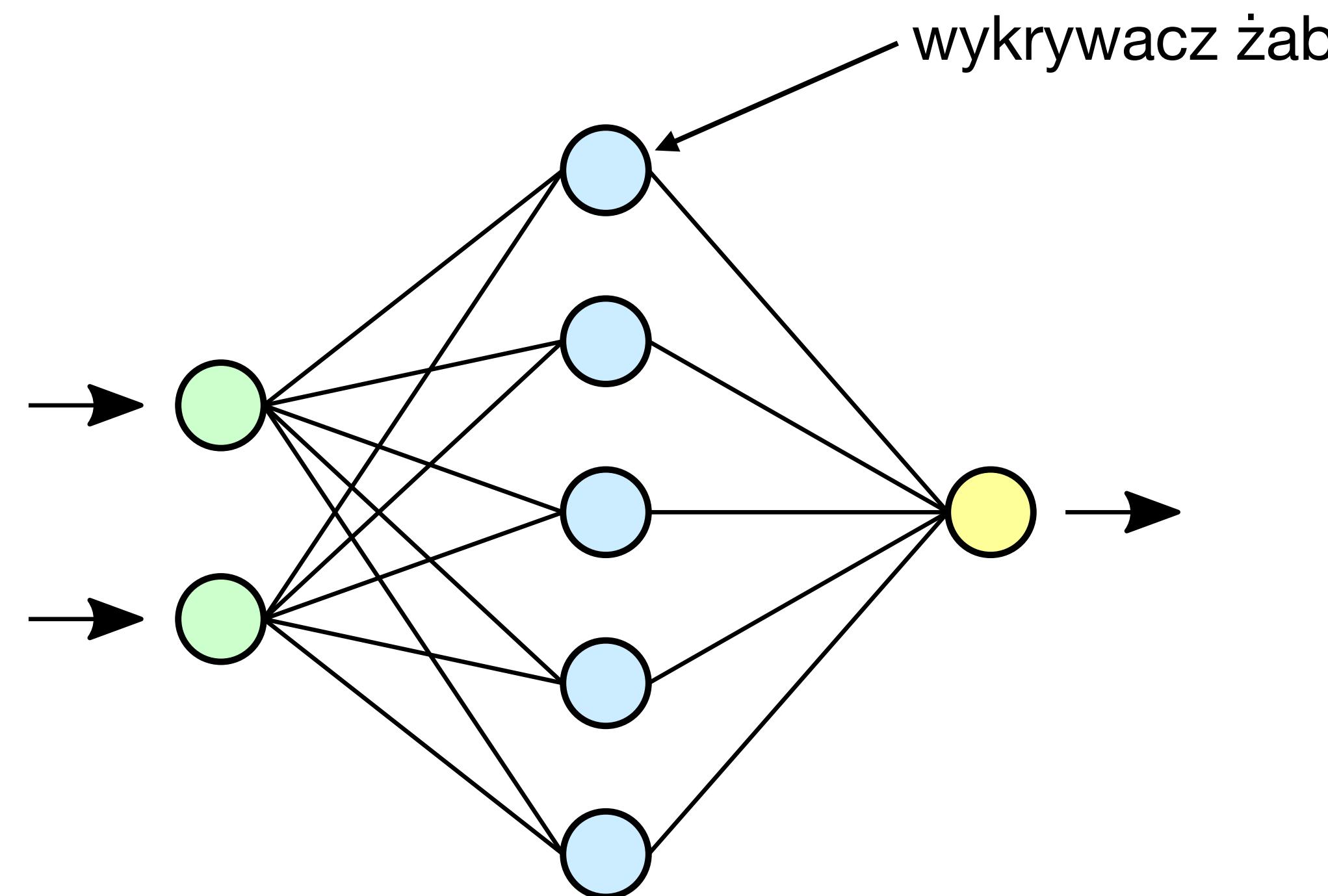
Adam Szokalski

Promotor: dr. inż. Mateusz Modrzejewski

Audio Intelligence Lab, Instytut Informatyki Politechniki Warszawskiej

Cel Pracy

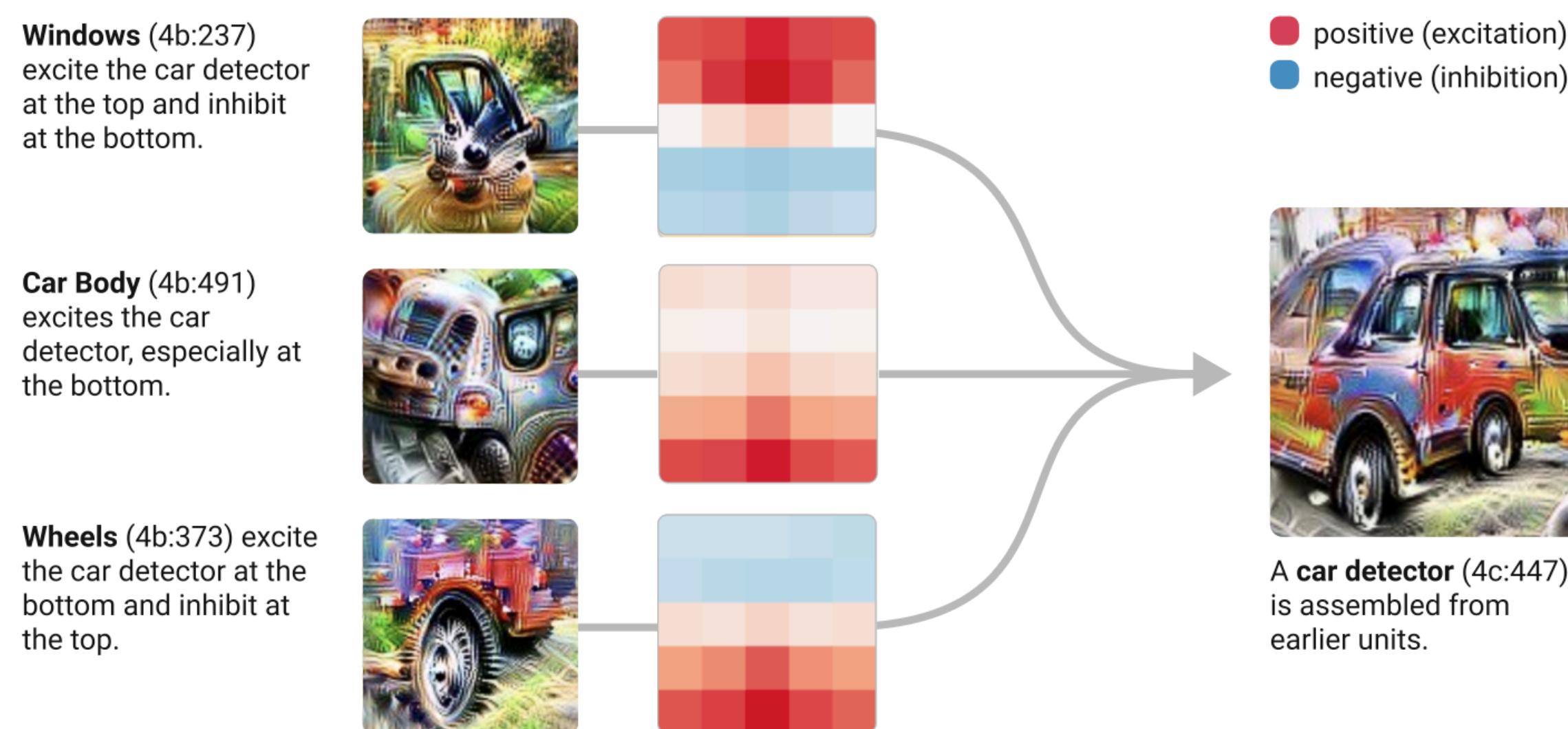
Stworzenie metody umożliwiającej interpretację funkcji poszczególnych neuronów w modelu dyfuzyjnym (Stable Diffusion 1.X).



Obrazek: Wikipedia

Motywacja

Taka metoda przydatna może być w badaniach nad **mechanistyczną interpretowalnością** - dziedziną próbującą dokonać inżynierii wstecznej sieci neuronowych w celu zrozumienia ich działania ([Olah, 2022](#)).



Obrazek: Olah et. al, 2020

Cecha

- W teorii mechanistycznej interpretowalności **cechy** to **fundamentalne jednostki sieci neuronowej**.
- Odpowiadają one **kierunkom w wektorowej przestrzeni aktywacji modelu**.
- Spekulacyjne założenie: cechy są **zrozumiałe dla ludzi**.

Feature #34M/31164353 **Golden Gate Bridge** feature example

The feature activates strongly on English descriptions and associated concepts

They also activate in multiple other languages on the same concepts

in the Presidio at the end (that's the huge park right next to the Golden Gate bridge), perfect. But not all people

ゴールデン・ゲート・ブリッジ、金門橋は、アメリカ西海岸のサンフランシスコ湾と太平洋が接続するゴールデンゲート海

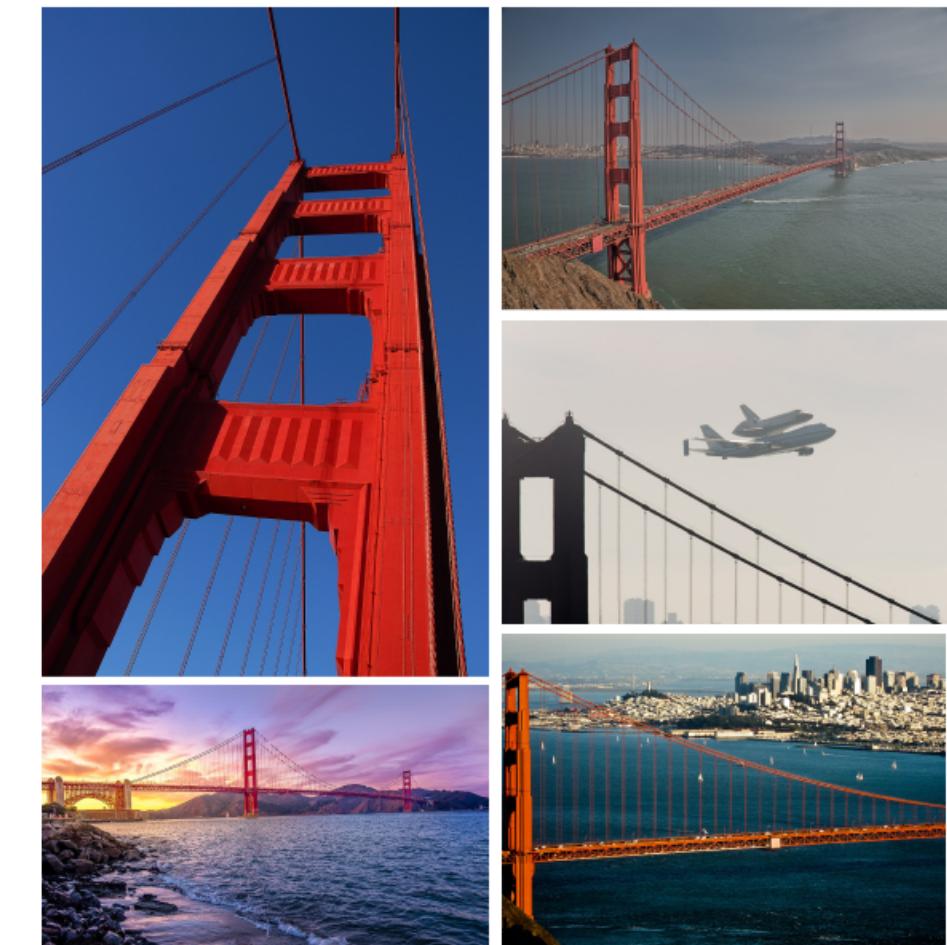
repainted, roughly, every dozen years." "while across the country in san francisco, the golden gate bridge was

골든게이트교 또는 금문교는 미국 캘리포니아주 골든게이트 해협에 위치한 현수교이다. 골든게이트 교는 캘리포니아주 샌프란시

it is a suspension bridge and has similar coloring, it is often compared to the Golden Gate Bridge in San Francisco, US

МОСТ золотые ворота – висячий мост через пролив золотые ворота. он соединяет город сан-фран

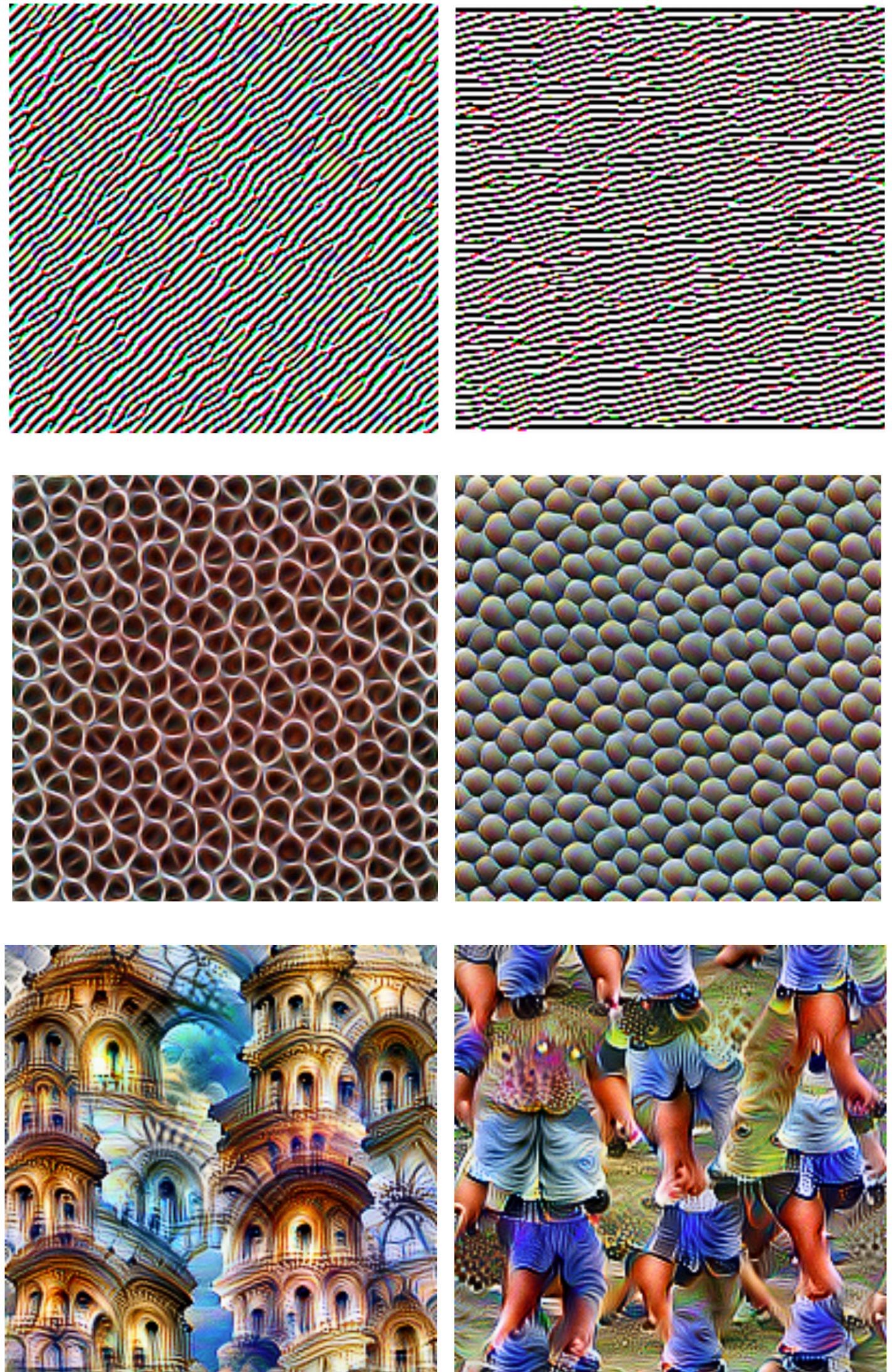
And on relevant images as well



Cecha *Golden Gate Bridge* znaleziona w Claude 3 Sonnet (Templeton et. al., 2024)

Wizualizacja cech

- Mordvintsev et. al. 2015 zaproponowali **metodę wizualizacji cech** w sieciach konwolucyjnych.
- Polega na optymalizacji gradientowej podanego na wejście szumu w celu **maksymalizacji aktywacji danej cechy** sieci.
- Wizualizacje podatne są na degenerację do postaci niezrozumiałego **adwersarialnego szumu wysokoczęstotliwościowego**.
- **Metody regularyzacji** pozwalają utrzymać wizualizację w formie zrozumiałej dla człowieka.

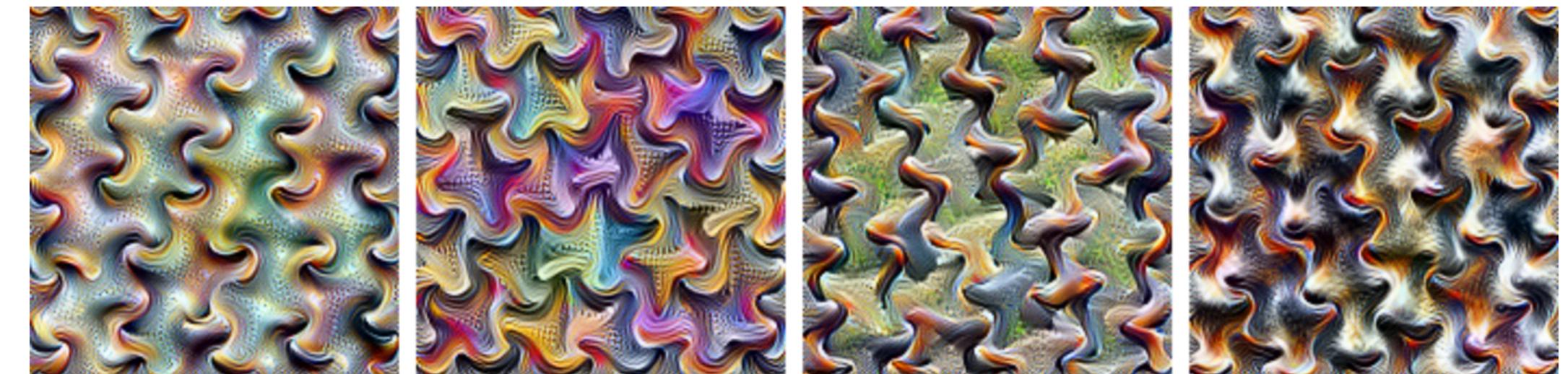


Wybrane wizualizacje DeepDream ([Olah et. al. 2017](#))

Superpozycja

- Neurony są **monosemantyczne** gdy jeden neuron reprezentuje jedną cechę.
- Niestety, w praktyce zdarza się to jedynie we wczesnych warstwach modeli.
- Głębsze warstwy zwykle są **polisemantyczne**, tzn. reprezentują wiele cech, są *splątane*.
- Zjawisko to nazywamy **superpozycją** i wynika ono z ograniczonych rozmiarów sieci neuronowych, które muszą *kompresować* informacje.
- Polisemantyczne neurony ciężej jest interpretować.

Neuron monosemantyczny

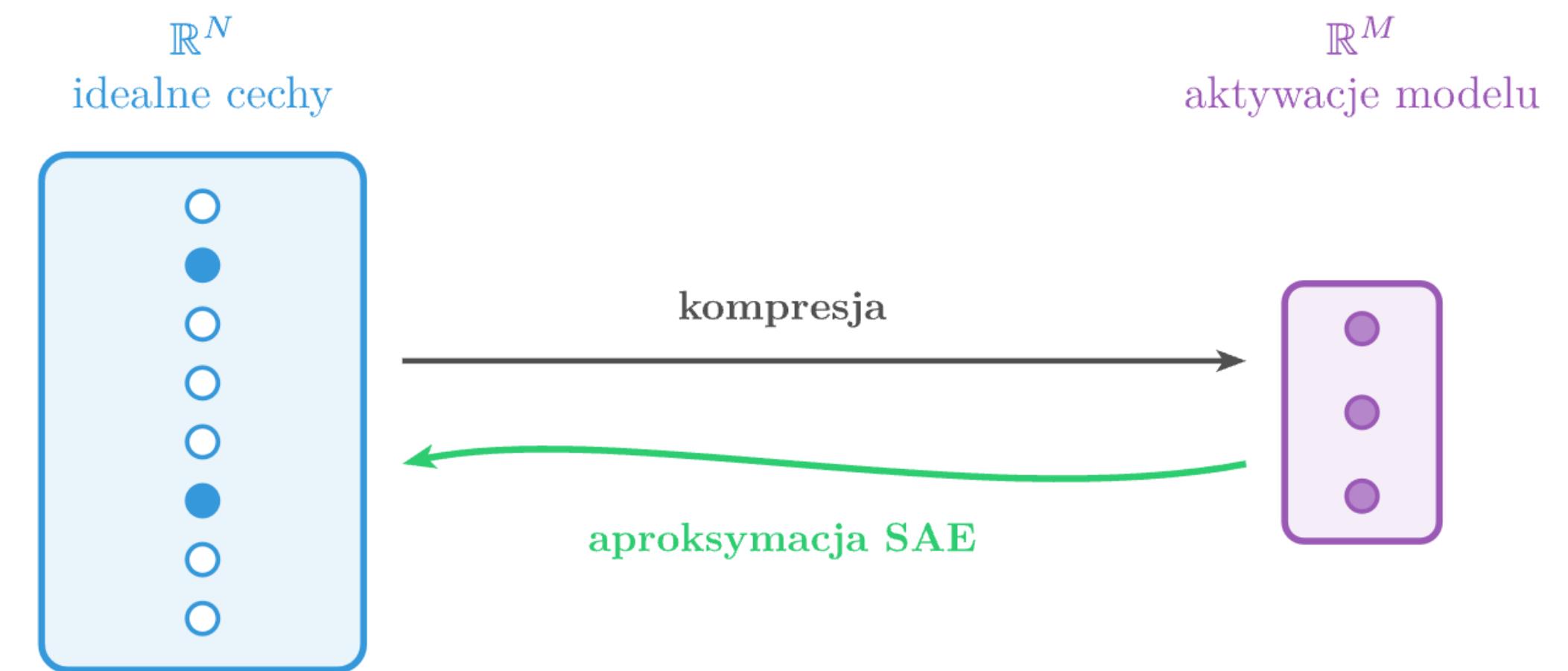


Neuron polisemantyczny



Rzadkie autoenkodery (SAE)

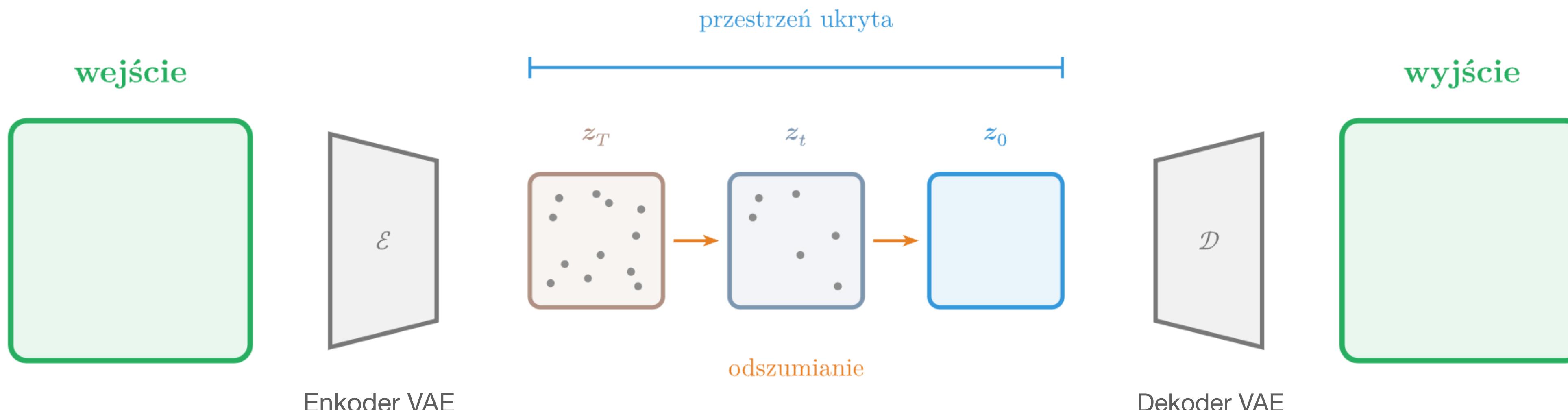
- SAE pozwalają na **odpłatanie polisemantycznych neuronów i analizę monosemantycznych cech.**
- SAE trenowane są do **aproksymacji odwrocenia stratnej kompresji** powodującej superpozycję.
- Jest to możliwe dzięki założeniu **rzadkości** rozpłatanych aktywacji - większość z nich równa jest 0.



Wizualizacja superpozycji i jej rozpłatania przez SAE

Modele dyfuzji w przestrzeni ukrytej

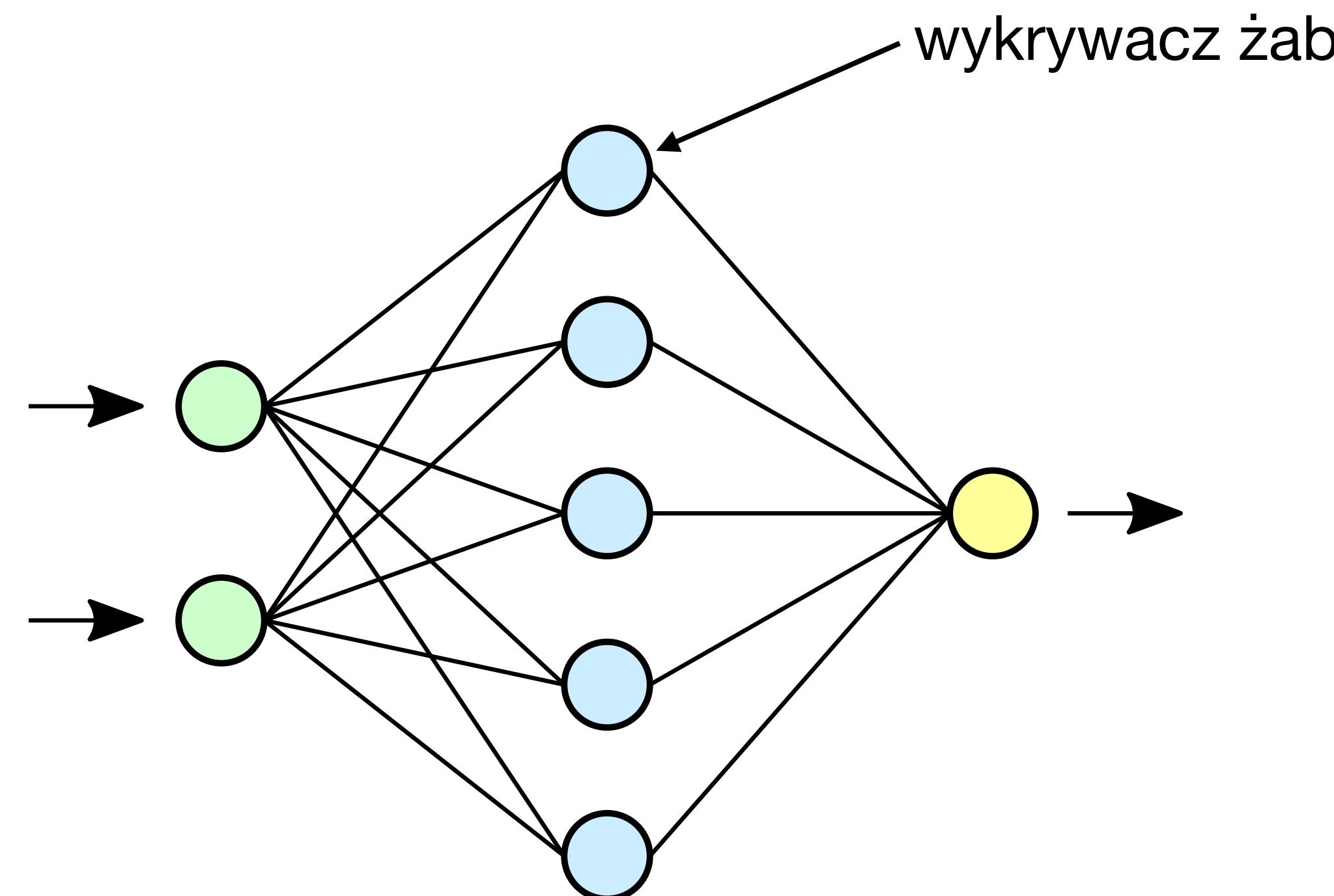
- Modele dyfuzyjne to sieci typu U-Net trenowane do **stopniowego odszumiania obrazu** pod wpływem danego warunku (np. prompta).
- Działają na kolejnych krokach czasowych. Najczęściej od 1000 do 0.
- Modele state-of-the-art takie jak Stable Diffusion 1.X **wykonują dyfuzję w skompresowanej przestrzeni ukrytej** autoencodera wariacyjnego (VAE).



Wizualizacja modelu dyfuzji w przestrzeni ukrytej.

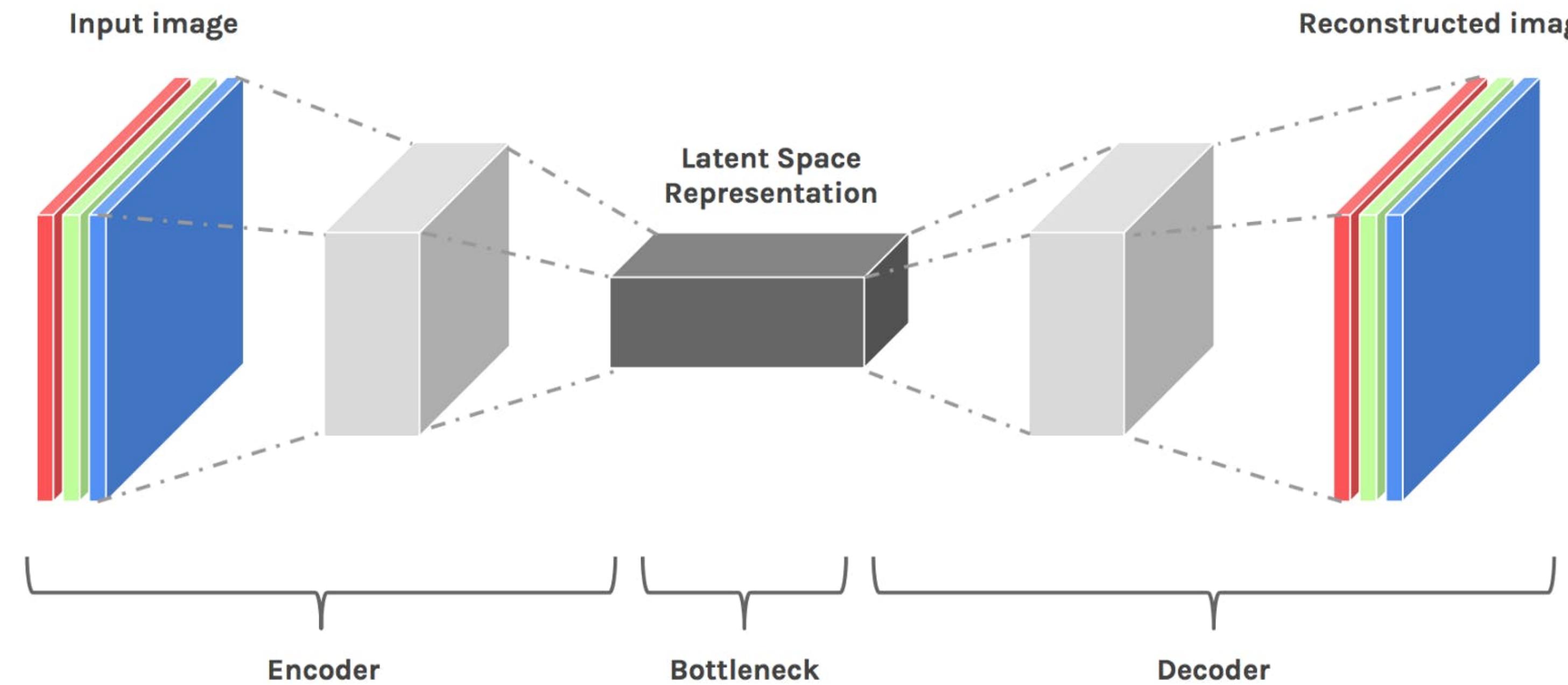
Cel Pracy – rozwiniecie

Stworzenie metody umożliwiającej interpretację funkcji poszczególnych neuronów w modelu dyfuzyjnym.



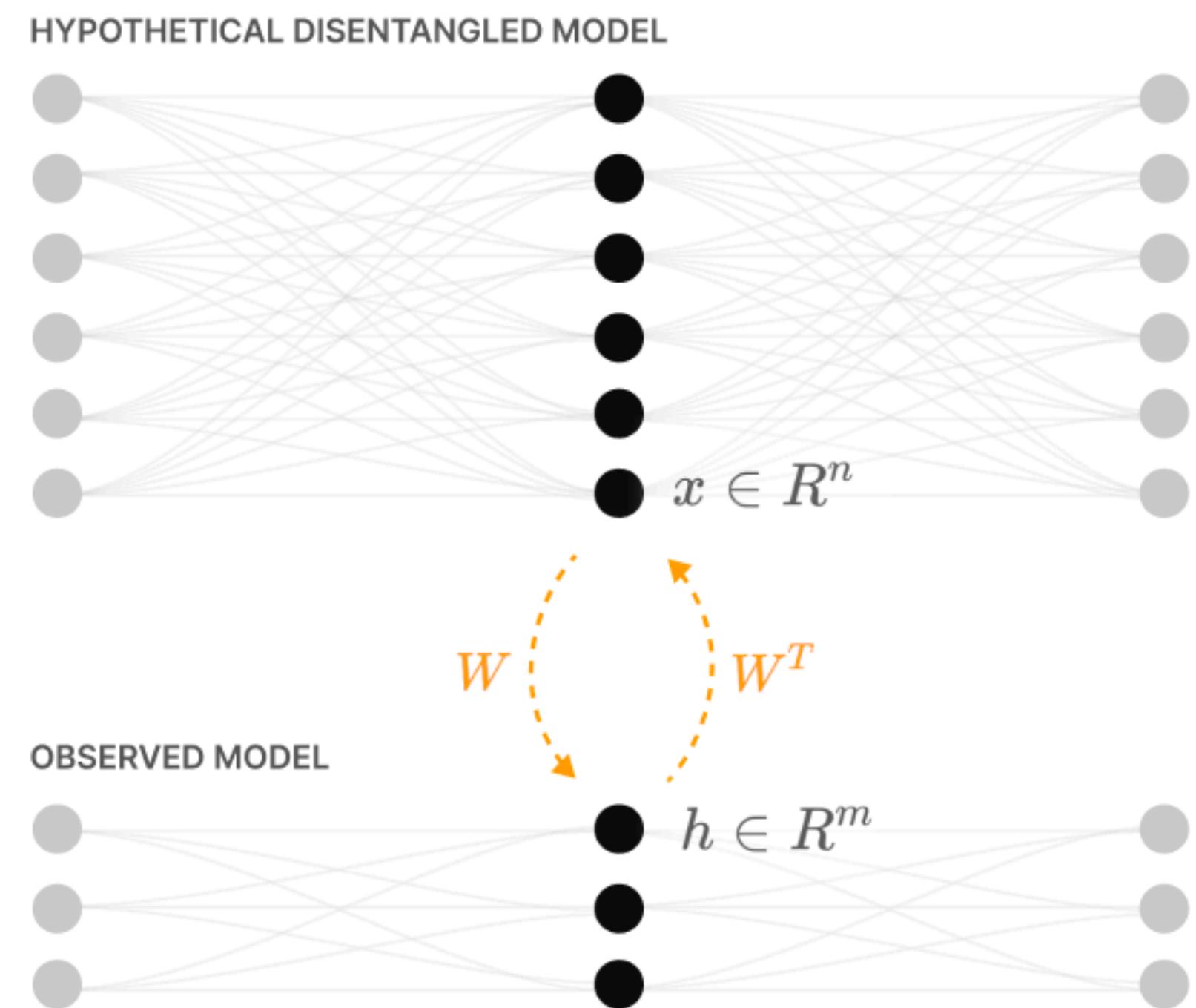
Cel Pracy: rozwinięcie

Stworzenie metody umożliwiającej interpretację funkcji poszczególnych neuronów w modelu dyfuzyjnym **działającym w skompresowanej przestrzeni ukrytej.**



Cel Pracy: rozwinięcie

Stworzenie metody umożliwiającej interpretację funkcji poszczególnych neuronów **zakodowanych w superpozycji** w modelu dyfuzyjnym **działającym w skompresowanej przestrzeni ukrytej**.



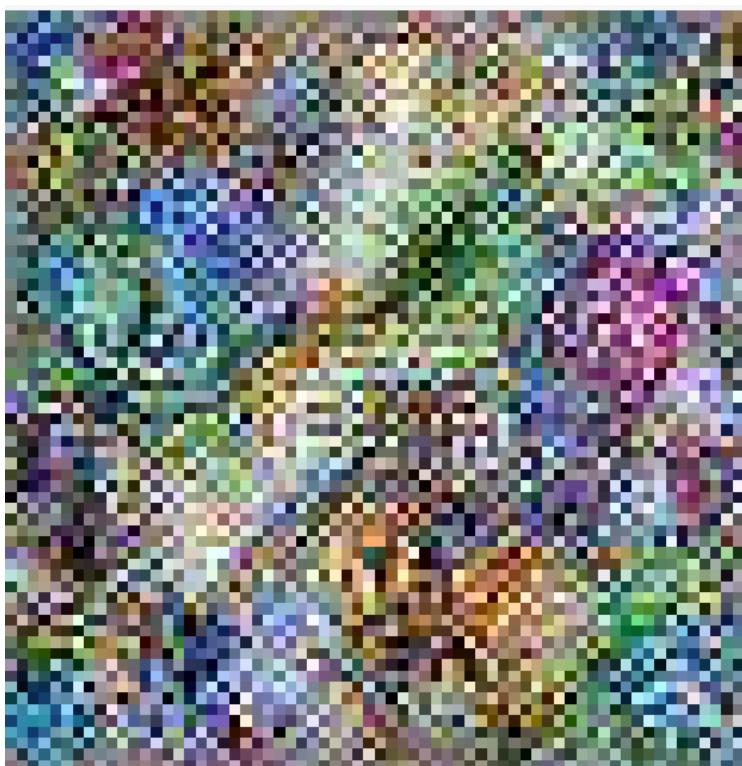
Obrazek: Anthropic

Cel Pracy: rozwinięcie

Stworzenie metody umożliwiającej interpretację funkcji poszczególnych neuronów **zakodowanych w superpozycji** w modelu dyfuzyjnym **działającym w skompresowanej przestrzeni ukrytej w sposób zrozumiały dla człowieka.**



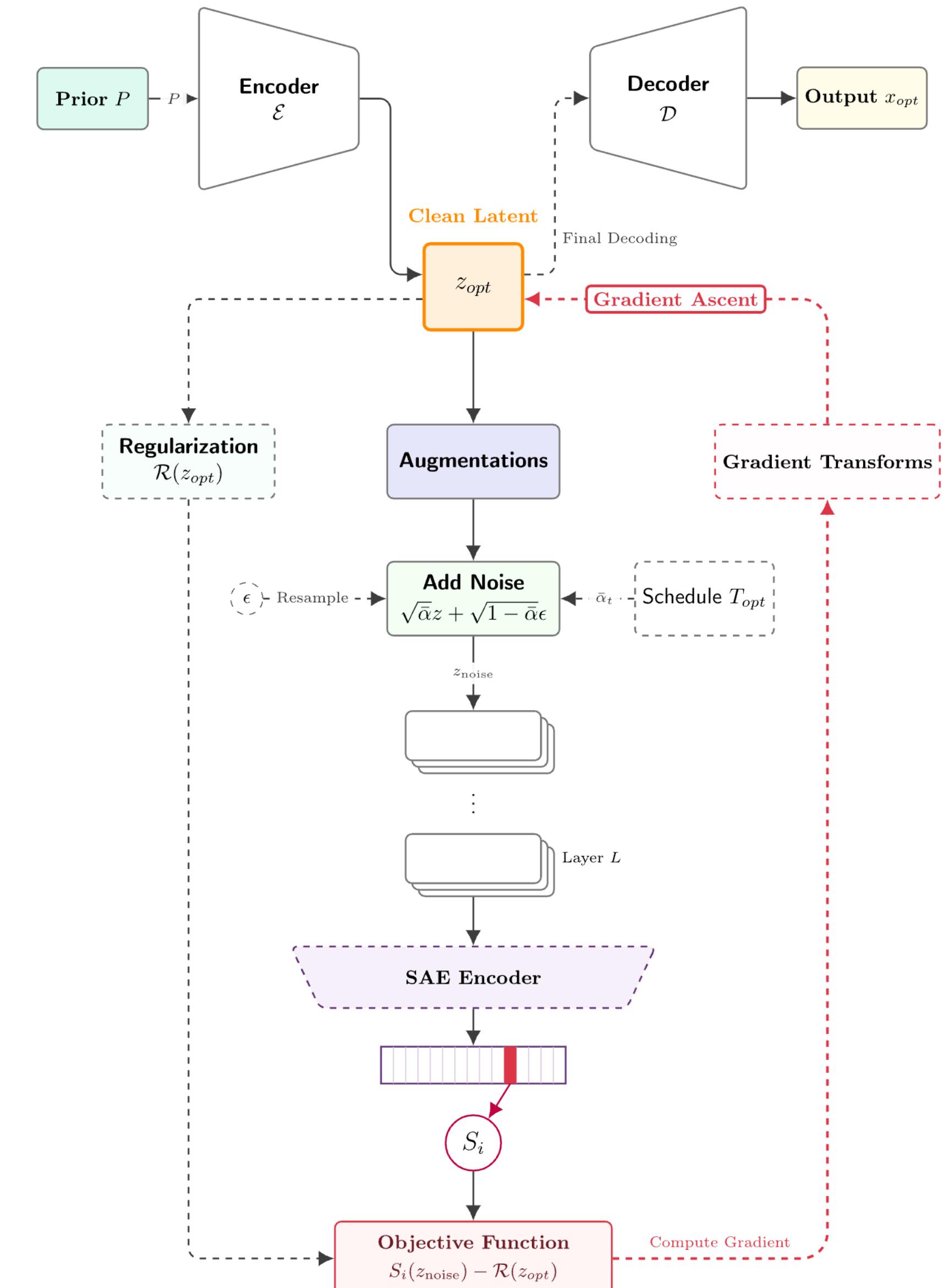
Kwiaty



Też Kwiaty

Architektura rozwiązania

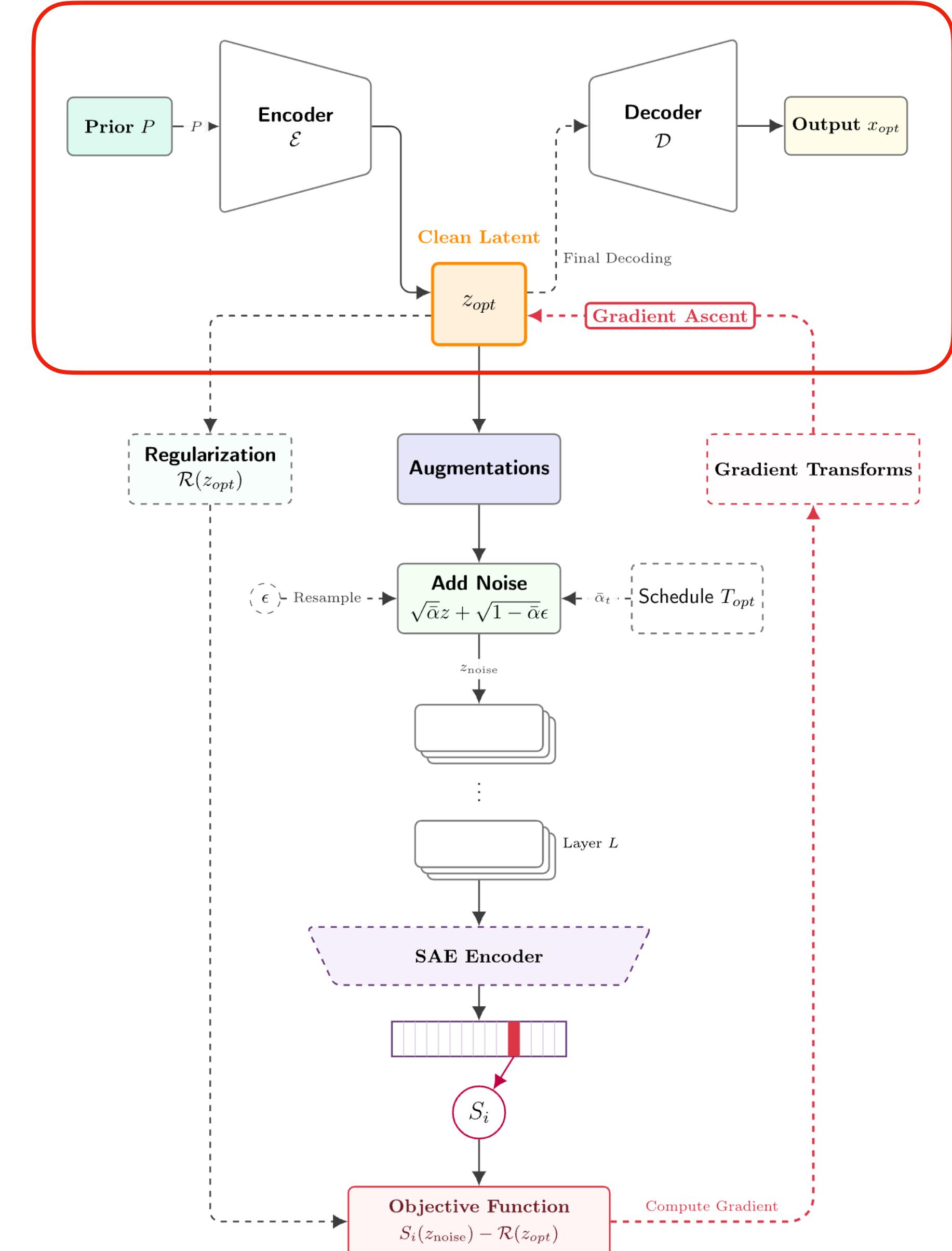
- **Dostosowanie metody** wizualizacji cech do potrzeb modeli dyfuzyjnych działających w przestrzeni ukrytej.
- **Pokazuje przyczynę** - wejście zoptymalizowane tak aby ekscytowało neuron.
- Wymaga optymalizacji na **konkretnym kroku czasowym**.



Schemat architektury rozwiązania.

Optymalizacja w przestrzeni ukrytej

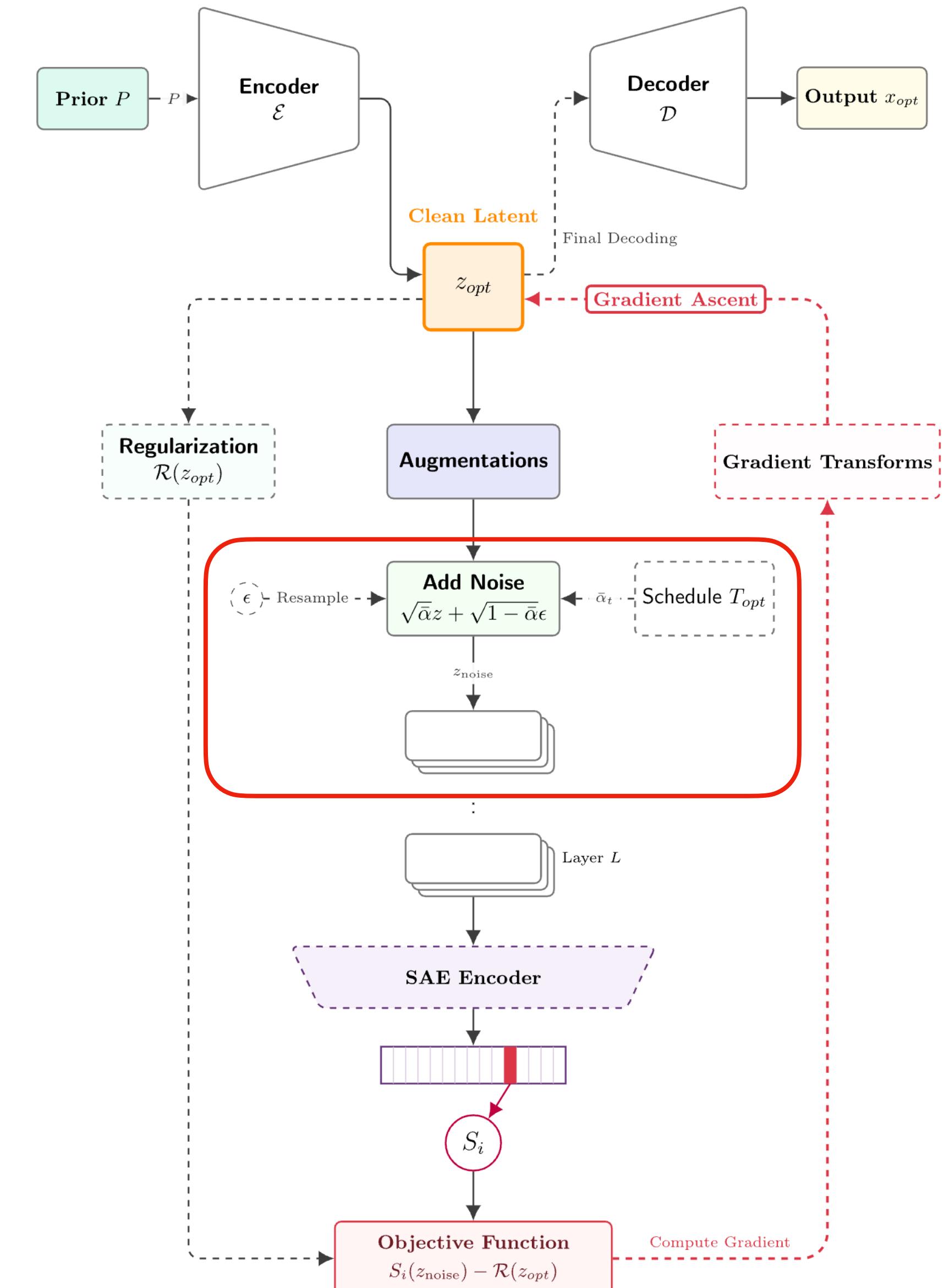
- Optymalizacja obrazu sprzed enkodera jest wielokrotnie kosztowniejsza obliczeniowo (>10 -krotnie) wyjątkowo podatna na szum wysokoczęstotliwościowy.
- Moja metoda optymalizuje skompresowane wejście w przestrzeni ukrytej po czym dekoduje wynik do przestrzeni obrazu.



Schemat architektury rozwiązania.

Widzenie przez szum

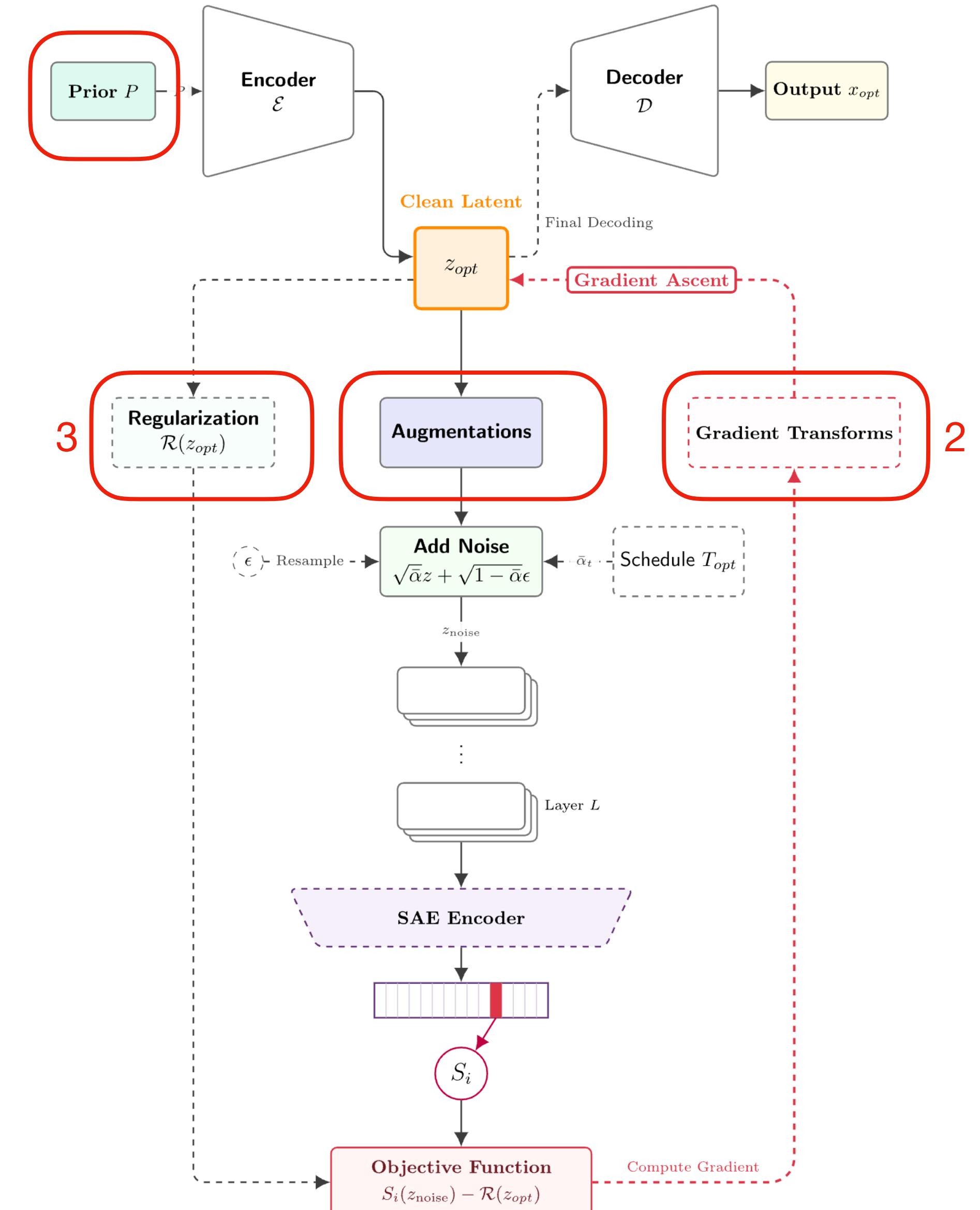
- U-Net modelu dyfuzyjnego w konkretnym kroku czasowym spodziewa się **odpowiednio zaszumionego wejścia** - tak jak był trenowany.
- Moja metoda podczas optymalizacji **używa harmonogramu szumu** (jest on częścią modelu) i sztucznie **zaszumia czysty cel optymalizacji** przed każdym przejściem przez model.



Schemat architektury rozwiązania.

Zaawansowana regularyzacja

- Aby zapobiec powstawaniu szumu adwersarialnego zastosowałem **7 technik regularyzacji**.
- Zbadałem wpływ i optymalne parametry każdej z nich **metodą jakościową** na pierwszych 30 cechach SAE.
- Ich użycie znaczaco **poprawiło jakość wyników** - od szumu do zrozumiałych konceptów.



Schemat architektury rozwiązania.

Wyniki

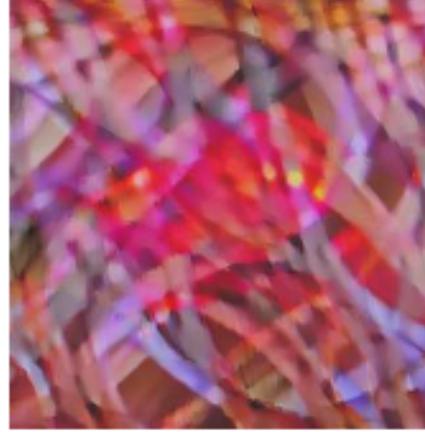
- Przeanalizowałem warstwę atencji **up1.1** modelu **StableDiffusion 1.4 (Style50)** rozplątaną SAE z **SAeUron** na krokach czasowych wykazujących silną aktywację danej cechy.
- Używając mojej metody udało mi się znaleźć **wiele zrozumiałych monosemantycznych cech**.
- Wizualizacje są **spójne z przykładami ze zbioru danych**, które wyjątkowo silnie aktywowały daną cechę podczas generacji.



Wizualizacje cechy 10331: człowiek



Przykłady ze zbioru danych cechy 10331.

Feature	Description	Visualization	Dataset Example	Example prompt¹³
14	cables			An Butterfly image in Pointillism style
10204	roses			An Flowers image in Winter style
10272	inward perspective			An Towers image in Surrealism style
10538	foam			An Waterfalls image in French style

Inne wybrane wyniki.

Wnioski i dalsze prace

- **Metoda wizualizacji cech przez optymalizację jest możliwa w nowoczesnych modelach dyfuzyjnych.**
- **Nie zależy ona od zbioru danych**, więc umożliwia analizę nawet kiedy badana cecha nie jest obecna w dostępnych przykładach.
- Użyta w badaniu metoda jakościowa stanowi ograniczenie i możliwą drogą rozwoju jest **użycie modeli językowo-wizyjnych do stworzenia ilościowej metryki zrozumiałości konceptów na obrazie**.
- Wraz z dr inż. Modrzejewskim planujemy przeanalizować działanie tej metody również w **modalności audio**.

Podziękowania

- Chciałbym podziękować **mgr inż Bartoszowi Cywińskiemu i dr inż Kamilowi Deji** za pomoc z projektem i konsultacje w sprawie ich pracy: SAeUron.
- Również dziękuję mojemu promotorowi **dr Mateuszowi Modrzejewskiemu** za wsparcie merytoryczne i nakierowanie mnie na odpowiednią literaturę.
- A Państwu, dziękuję za uwagę ;)