The Hessian Penalty: A Weak Prior for Unsupervised Disentanglement

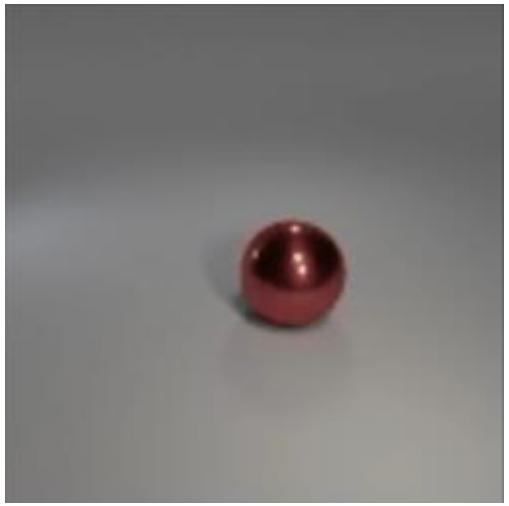
Чёлушкин Максим, БПМИ 172

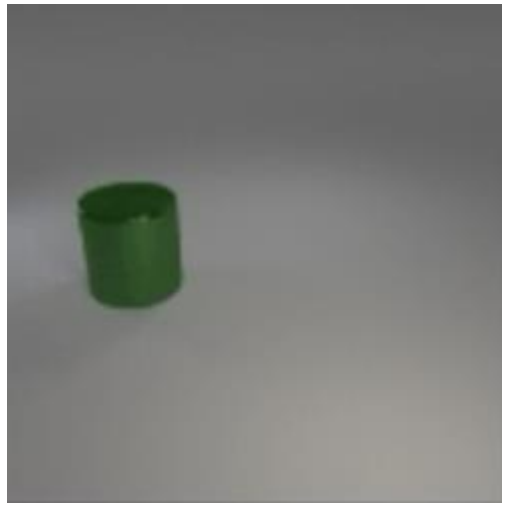
Проблема: Disentanglement(Распутывание)

- Давайте обучать нейронную сеть так чтобы она сама смогла различать понятия. При этом при обучении хотелось бы почти не подсказывать сети ключевые понятия, а чтобы она сама выделяла и группировала области.
- При обучении вводятся дополнительные критерии, которые позволяют выделять смыслы.

Пример

Датасет CLEVR(расположение, форма, цвет)



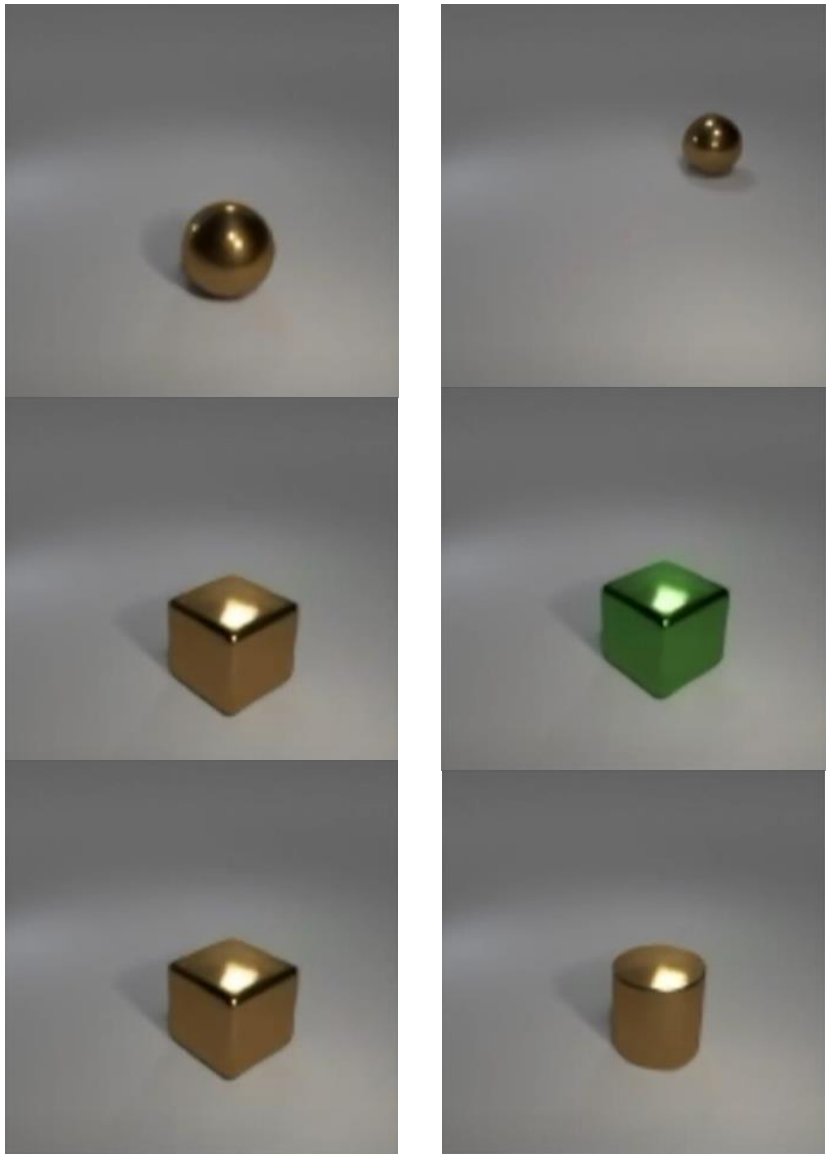


Пример

z_0(положение)

z_1(цвет)

z_2(форма)



$$\frac{\partial}{\partial z_1}\left(\frac{\partial G}{\partial z_0}\right) = \frac{\partial^2 G}{\partial z_1 \partial z_0}$$

Формулировка

Зависимость одной компоненты от другой:

$$\frac{\partial}{\partial z_1}\left(\frac{\partial G}{\partial z_0}\right) = \frac{\partial^2 G}{\partial z_1 \partial z_0}$$

$$H = \begin{bmatrix} \frac{\partial^2 G}{\partial z_0 \partial z_0} & \frac{\partial^2 G}{\partial z_0 \partial z_1} & \frac{\partial^2 G}{\partial z_0 \partial z_2} \\ \frac{\partial^2 G}{\partial z_1 \partial z_0} & \frac{\partial^2 G}{\partial z_1 \partial z_1} & \frac{\partial^2 G}{\partial z_1 \partial z_2} \\ \frac{\partial^2 G}{\partial z_0 \partial z_2} & \frac{\partial^2 G}{\partial z_1 \partial z_2} & \frac{\partial^2 G}{\partial z_2 \partial z_2} \end{bmatrix}$$

Hessian Penalty:

$$\mathcal{L}_H = \left(\frac{\partial^2 G}{\partial z_0 \partial z_1}\right)^2 + \left(\frac{\partial^2 G}{\partial z_1 \partial z_2}\right)^2 + \left(\frac{\partial^2 G}{\partial z_0 \partial z_2}\right)^2 = \sum_{i=1}^{|z|} \sum_{j \neq i}^{|z|} H_{ij}^2$$

Преобразование для использования

Теорема:

$$\mathcal{L}_H = \sum_{i=1}^{|z|} \sum_{j \neq i}^{|z|} H_{ij}^2 = 0.5\, Var_{\upsilon}(\upsilon^T H \upsilon)$$

, υ – Радамахерские вектора[P(υ_i = -1) = P(υ_i = 1) = 0.5]

$$\upsilon^T H \upsilon = \frac{1}{\epsilon^2}[G(z + \epsilon\upsilon) - 2\, G(z) + G(z - \epsilon\upsilon)]$$

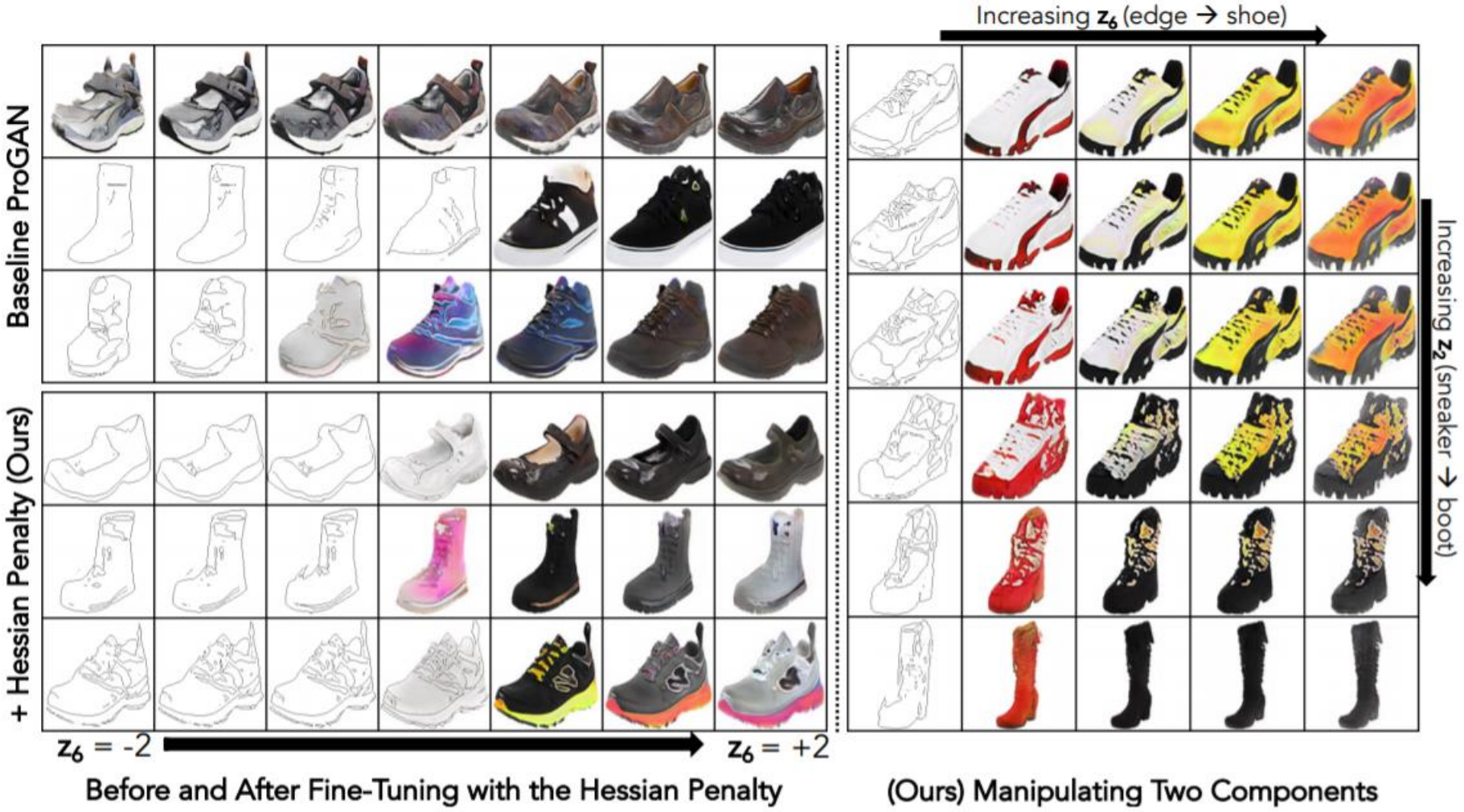
Эксперименты

- Обучение с Hessian Penalty
- Сокращение скрытого пространства
- Выделение направлений в скрытом пространстве

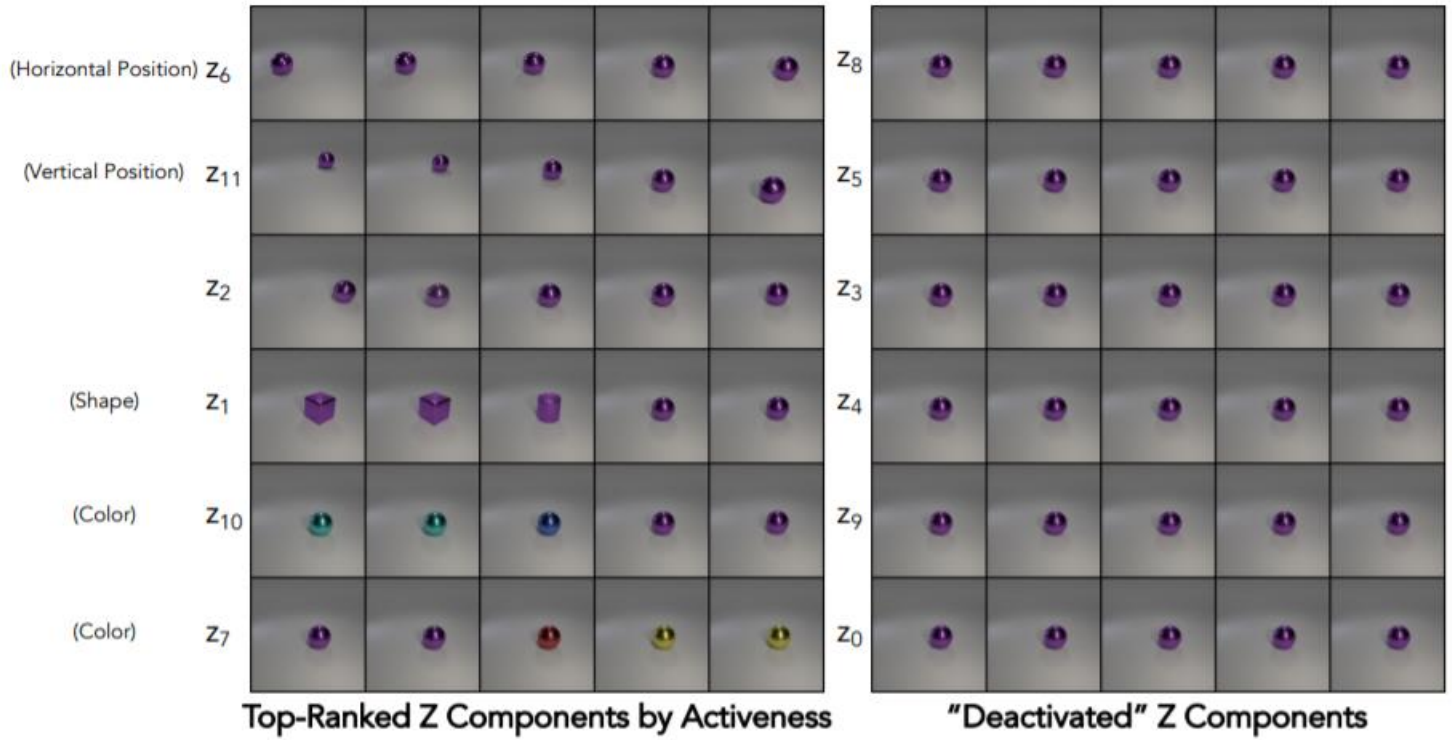
Обучение с Hessian Penalty

Для обучения был взят объединенный датасет Edges2Shoes.

В качестве бейзлайна был взят ProGAN.

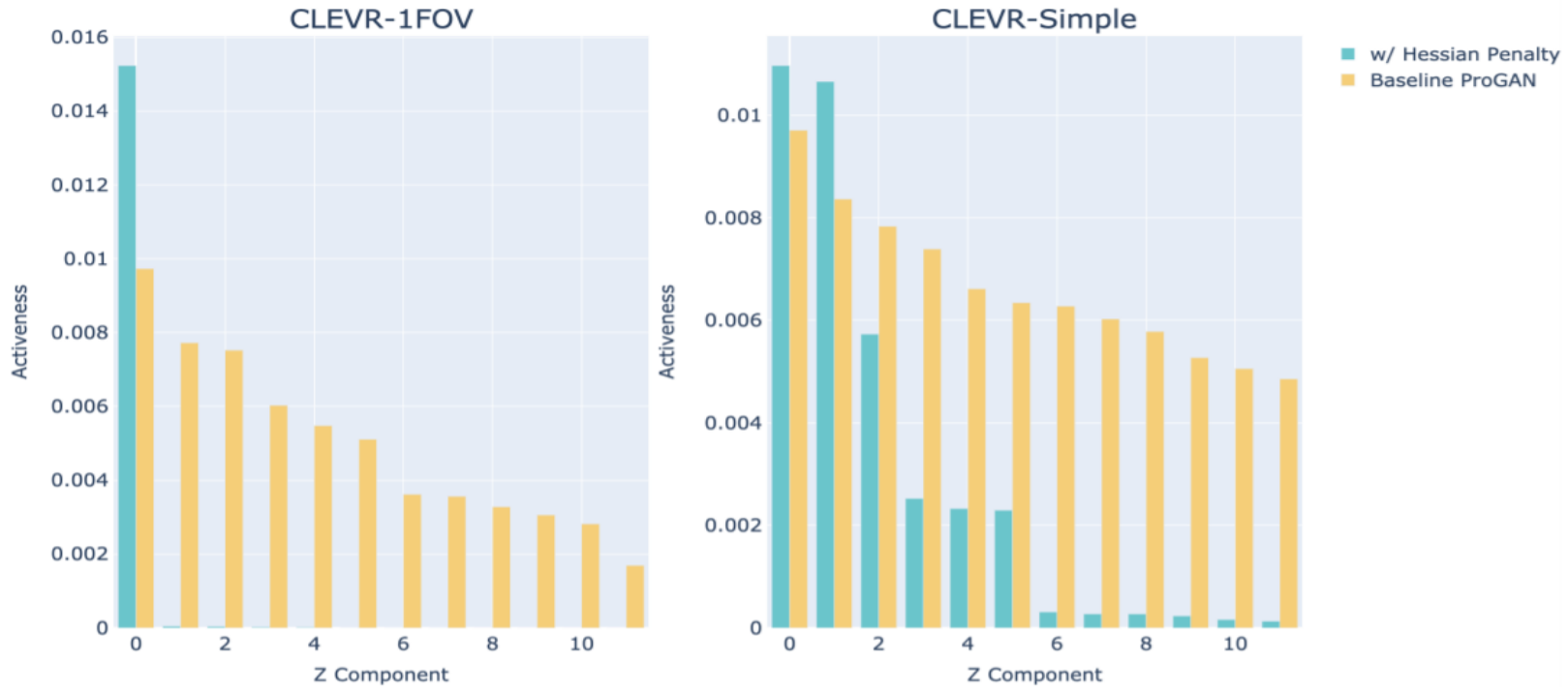


Сокращение скрытого пространства

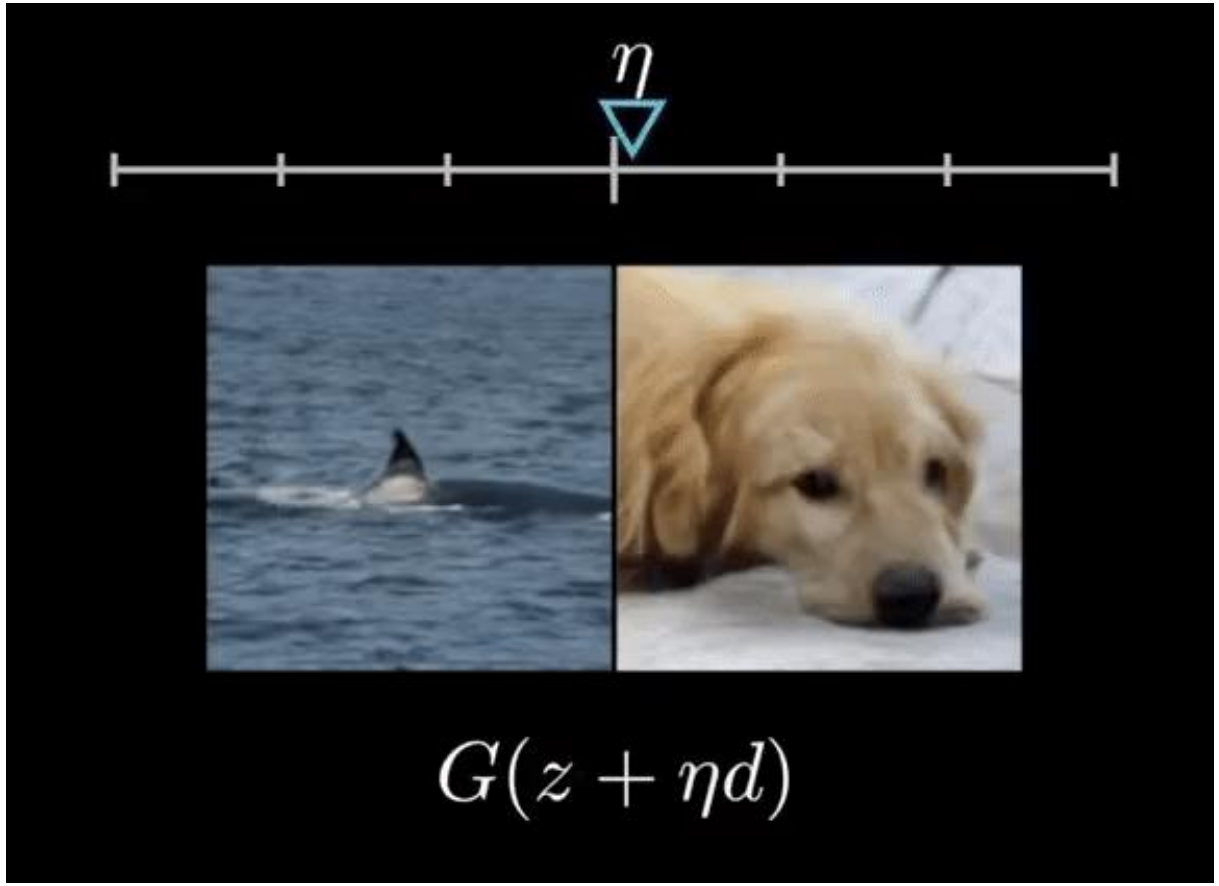


Сокращение скрытого пространства

$Activeness(z_i) = Mean(Var_{z_i}(G))$ – активность i-ой компоненты



Выделение направлений в скрытом пространстве

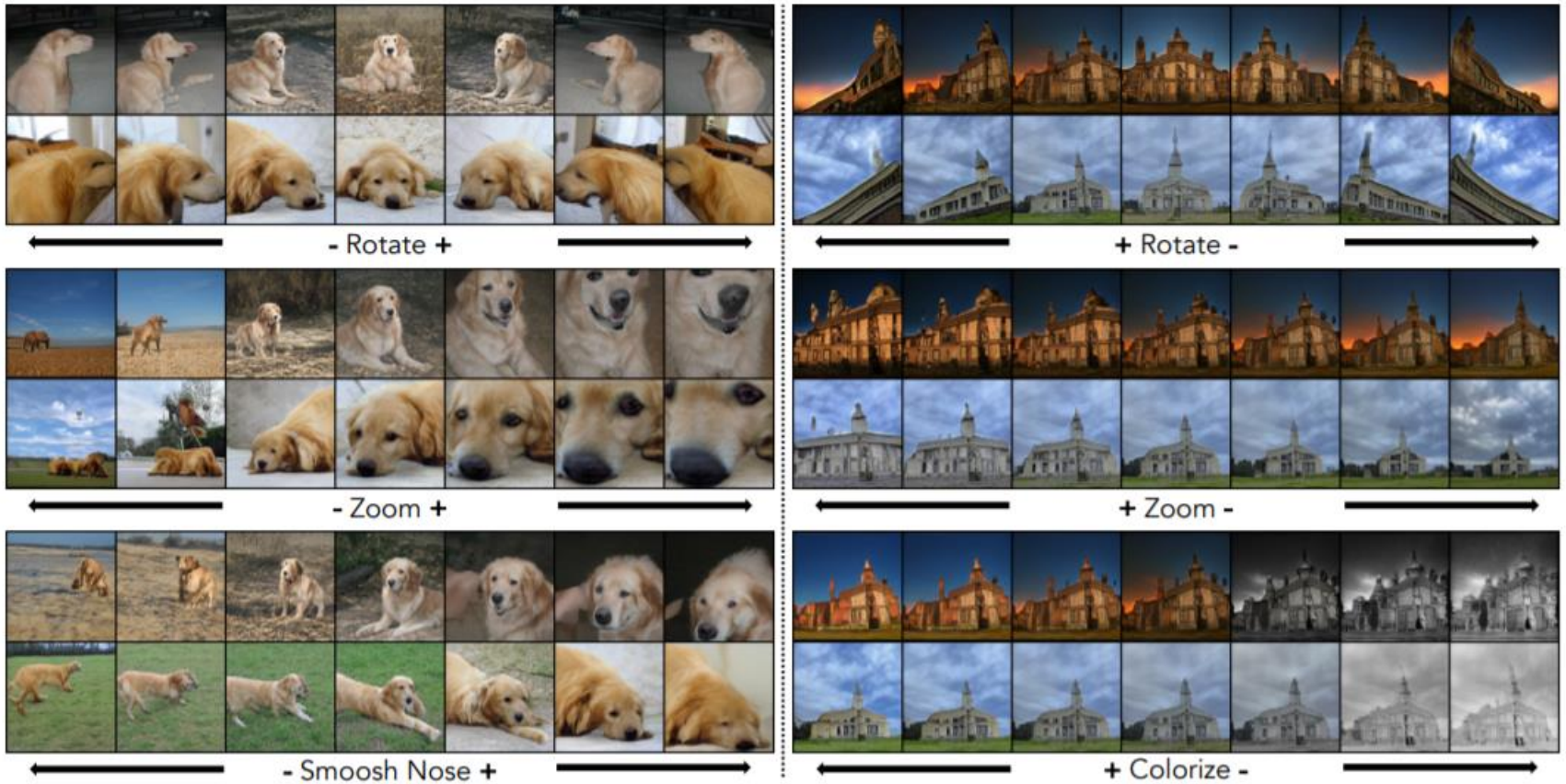


Выделение направлений в скрытом пространстве

Поиск d:

- $A = [d_0, d_1 \dots d_{N-1}]$
- $G(z + \eta A w_i), w_i - one\ hot$ vector
- $A^* = argmin_A \mathcal{L}_H(G(z + \eta A w_i))$

Выделение направлений в скрытом пространстве



Итог

Плюсы:

- Прост в использовании и не требует изменения архитектуры
- Хорошо показывает себя на простых датасетах

Минусы:

- Не достигает идеального disentanglement
- Ухудшение качества картинки
- Некорректная работа

Вопросы

- Выпишите формулу Hessian Penalty
- Как происходит поиск направлений в скрытом пространстве?
- Какие компоненты из скрытого пространства получилось выделить на примере Edges2Shoes?