DIPARTIMENTO
DI INFORMATICA
SISTEMISTICA
E COMUNICAZIONE

Classificazione della

Composizione Fotografica

delle immagini

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Relatore: *Dott. Luigi Celona*

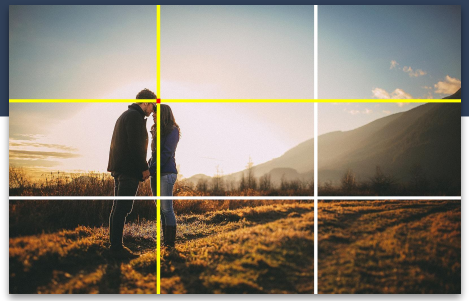
Correlatori: *Prof. Raimondo Schettini*, *Prof. Gianluigi Ciocca*

Candidato: *Alessandro Locatelli, 879362*

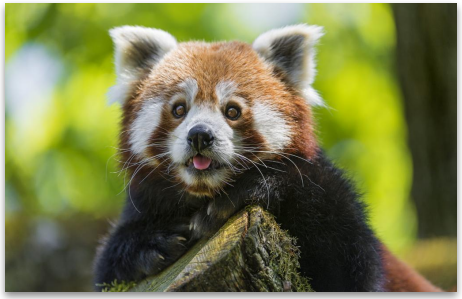
Anno Accademico 2023/2024

Composizione Fotografica

Posizione dei soggetti

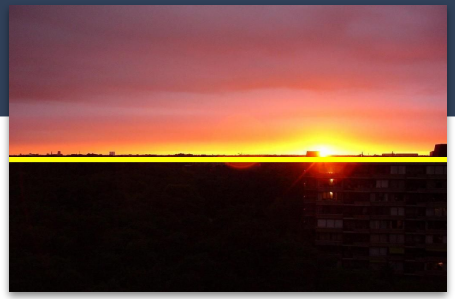


Regola dei terzi (RoT)



Centrato

Leading Lines



Orizzontale

Incidenti

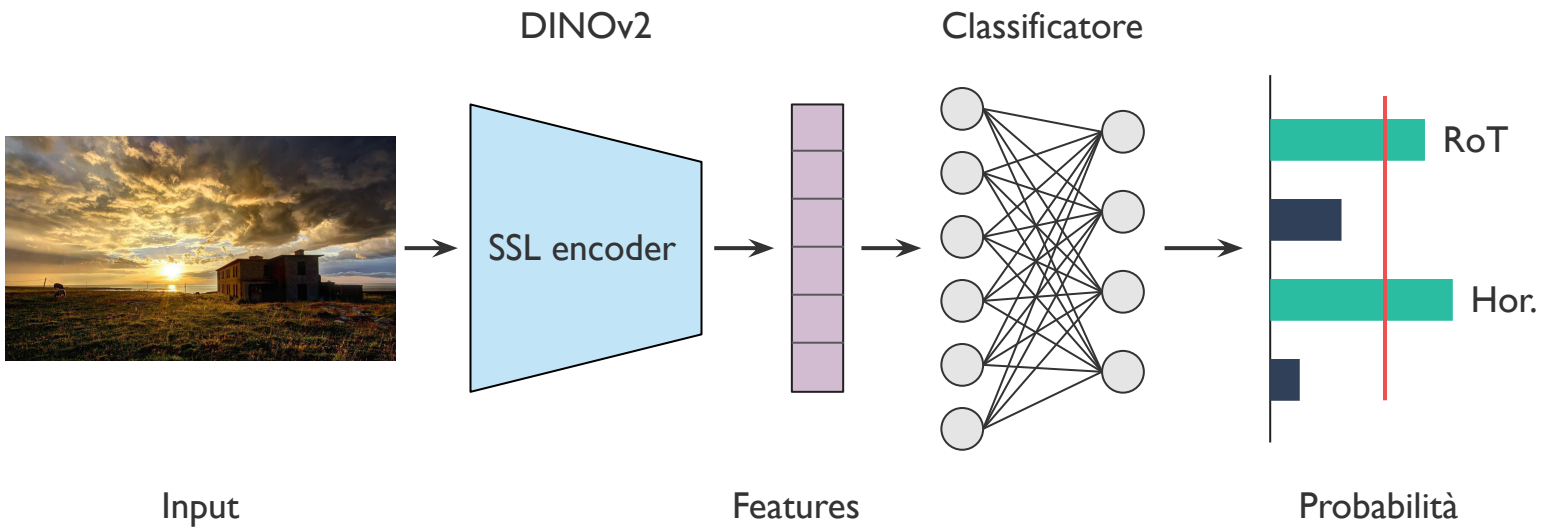
Forme

Triangolo

Pattern

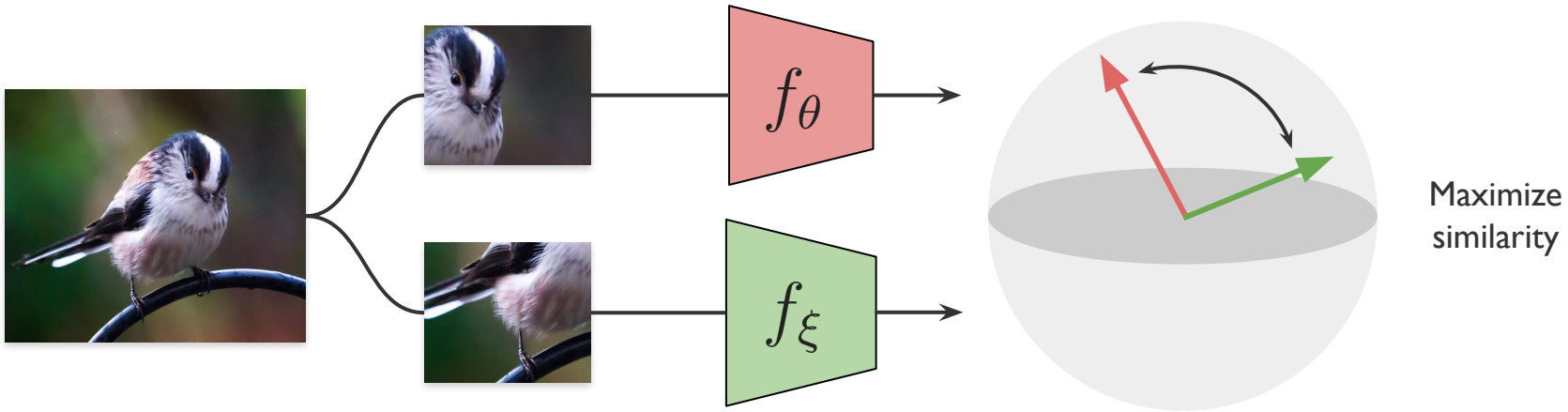
Obiettivo

Valutare le capacità di generalizzazione di DINOv2 nella classificazione della composizione



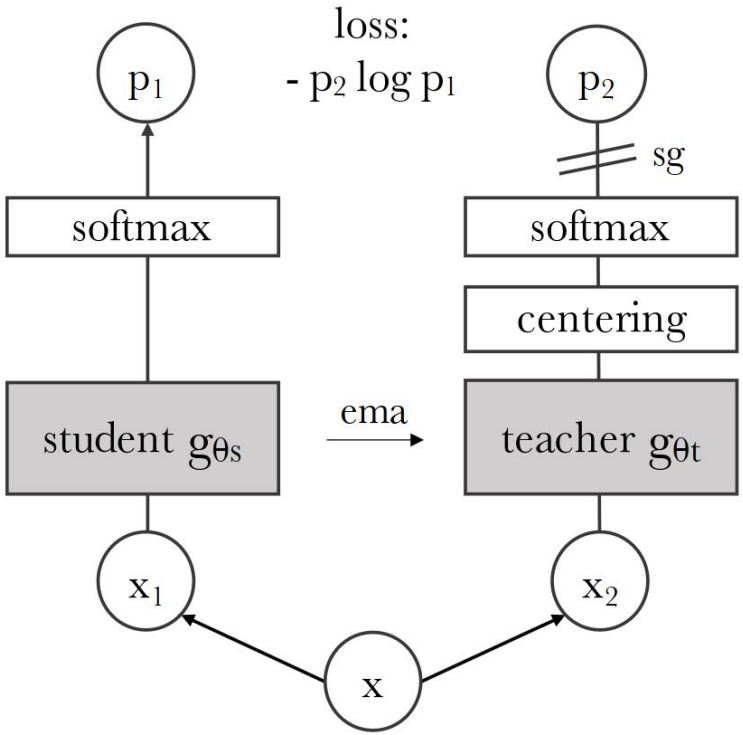
Self-Supervised Learning (SSL)

Produrre rappresentazioni descrittive e intelligibili di dati non etichettati



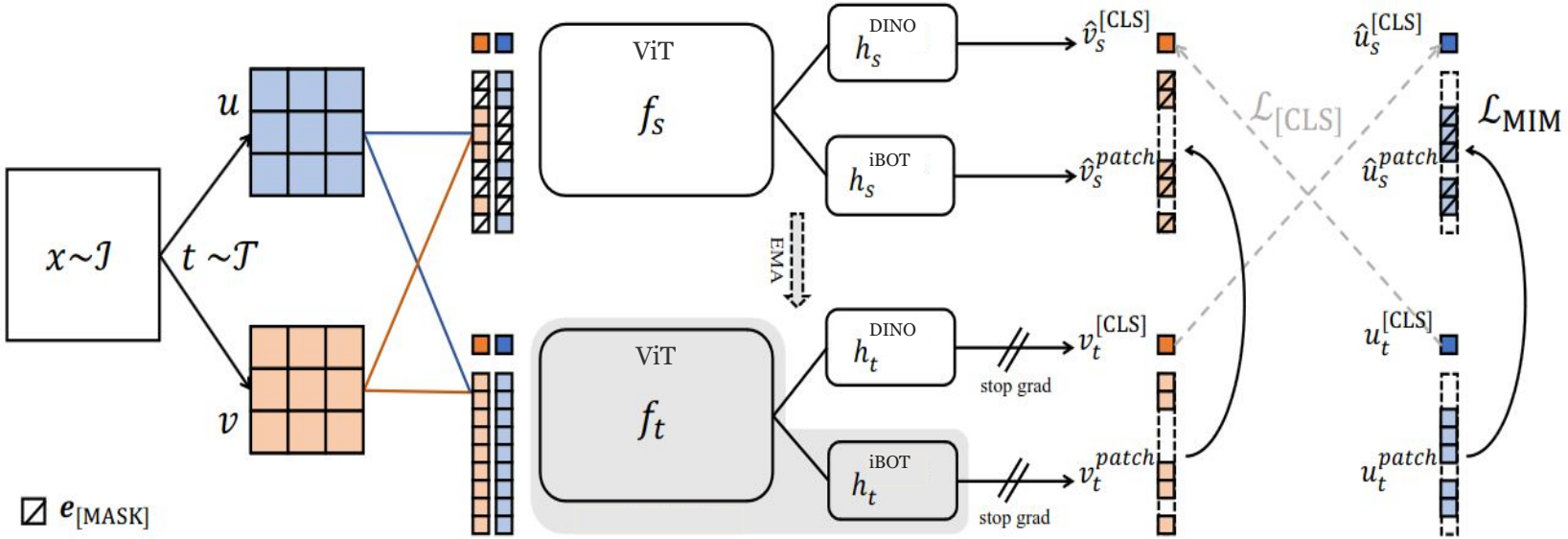
DINO

Self-distillation with NO labels



[1] *"Emerging Properties in Self-Supervised Vision Transformers" (2021), MetaAI*

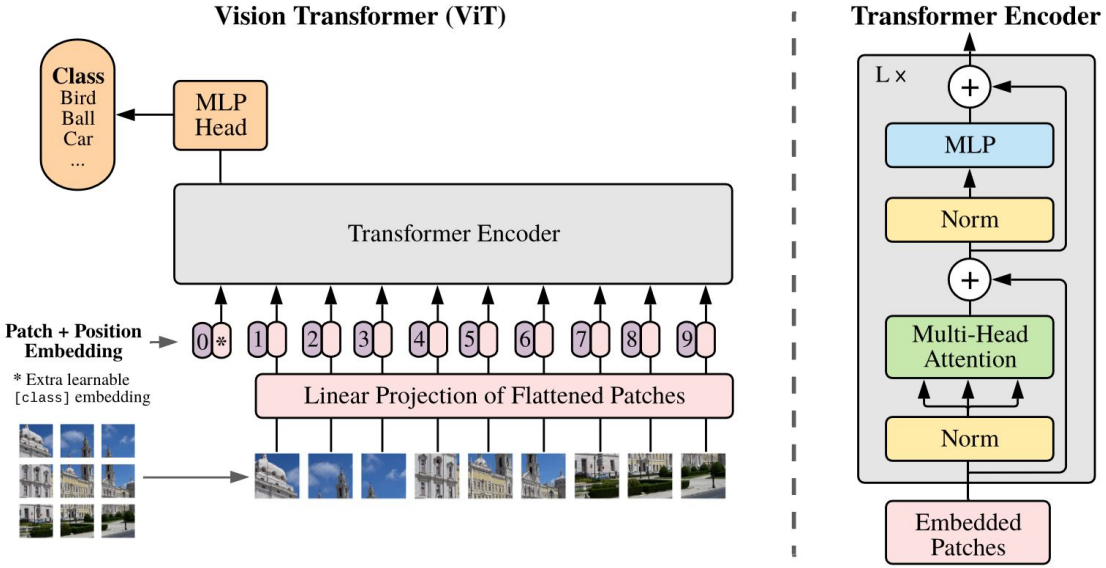
DINOv2



[2] *"DINOv2: Learning robust visual features without supervision" (2023), MetaAI*

Vision Transformers

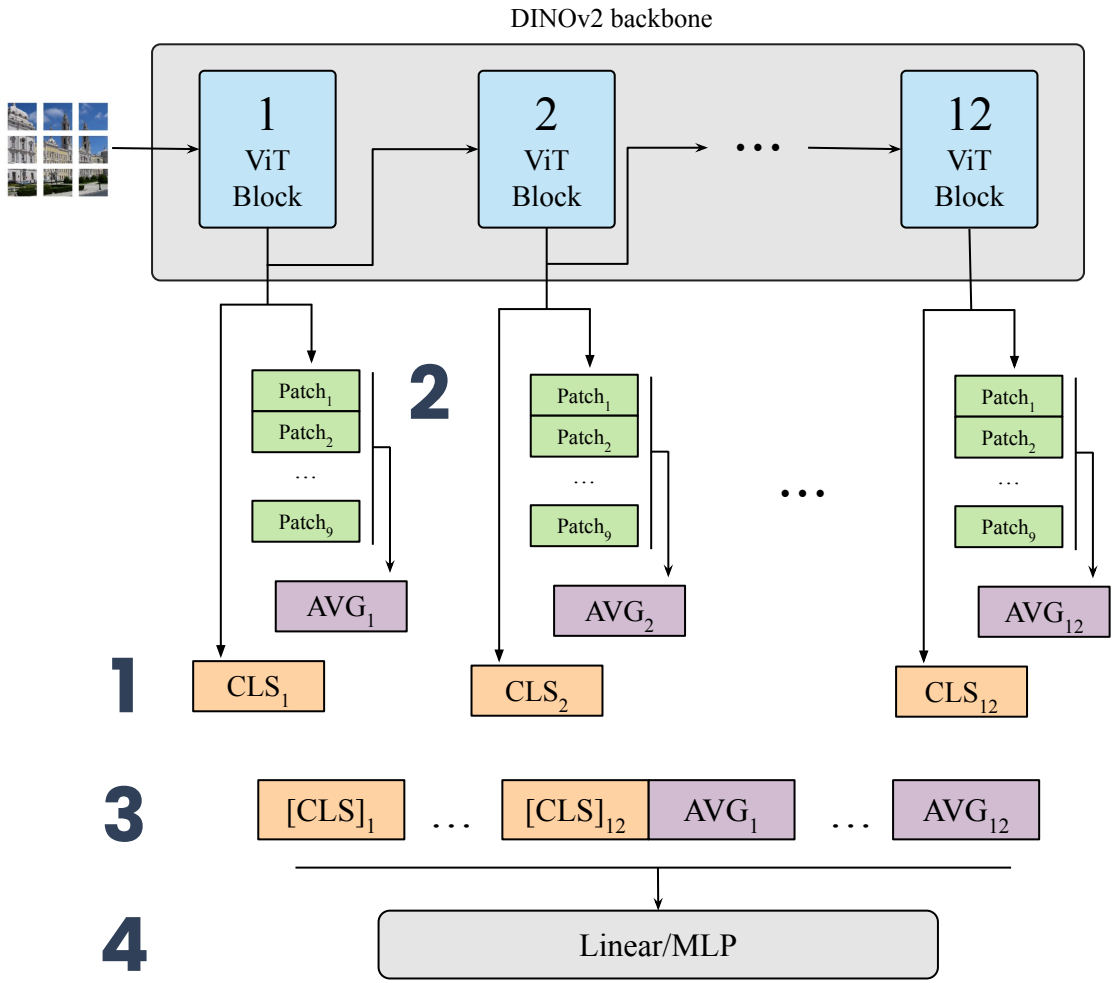
- Dividere l'immagine in patches
- Aggiungere token [CLS]
- Encoder produce vettori di features per i patch



[3] *"An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" (2020)*

Metodo

1. Estrazione [CLS]
2. Mediare features dei patch
3. Concatenazione
4. Classificatore



KU-PCP

Dataset standard nella composizione fotografica

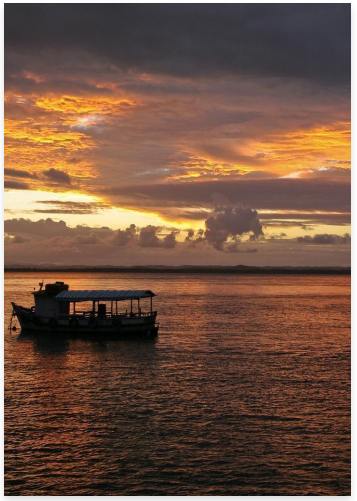
- 9 classi, multi-label
- 4.244 immagini etichettate
- ~20% doppia o tripla label

(a) *RoT*

(b) *Center*

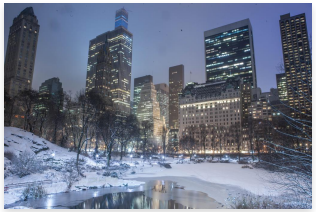
(c) *Symmetric*

(d) *Diagonal*

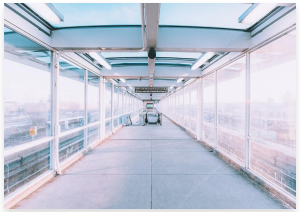


(e) *Horizontal*

(f) *Curved*



(g) *Vertical*



(h) *Triangle*

(i) *Pattern*

[4] *"Photographic Composition Classification and Dominant Element Detection for Outdoor scenes" (2018)*

LODB

Nato per correggere i difetti di KU-PCP

- 17 classi, multi-label
- 6.029 immagini etichettate
- ~200 immagini con doppia label

(a) *O2DiaR.*

(b) *RoT.R*

(c) *O3Tri.*

(d) *O3Li.*

(f) *O2DiaL.*

(g) *RoT.L*

(h) *O2Hor.*

(i) *Oline*

[5] "*Self-supervised Photographic Image Layout Representation Learning*" (2024)

Risultati

1. Linear Layer
2. MLP (1 hidden, ReLU)

BCE (binary cross-entropy)

Batch size 1024

Encoder ViT-S/14

K-fold cross-validation (5)

NVIDIA Tesla K80

Le features di DINOv2 non sono adatte al task

	Dataset	Epoche	lr	Acc. (Any)	Acc. (All)	MP	mAP	MF1
1	KUPCP	300	0.003	0.83	0.63	0.84	0.81	0.79
	LODB	200	0.003	0.56	0.52	0.66	0.65	0.49
2	KUPCP	200	0.005	0.84	0.54	0.83	0.76	0.77
	LODB	300	0.002	0.51	0.50	0.68	0.67	0.44

4 [CLS] + 4 AVG

Confronto con altri metodi in letteratura su KU-PCP:

Metodo	Acc. (Any) %
FT_CNN [4]	88.8
RSTN [6]	90.9
DINOv2 + linear	82.9

Per classe

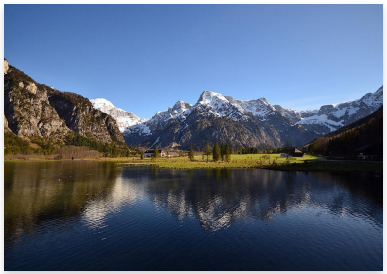
Scarsa varianza semantica intraclasse in KU-PCP influenza i risultati

Classe	**Precision**	**Recall**	**F1-score**
RoT	0.6866	0.7480	0.7160
Vertical	0.8571	0.6364	0.7304
Horizontal	0.8390	0.8190	0.8289
Diagonal	0.7372	0.5344	0.6196
Curved	0.7722	0.6630	0.7135
Triangle	0.8487	0.7544	0.7988
Center	0.9135	0.7814	0.8423
Symmetric	0.9770	0.8586	0.9140
Pattern	0.9683	0.9683	0.9683

Metriche per classe in KU-PCP, linear layer

Varianza semantica intraclasse

L'ipotesi è che l'informazione semantica influenzi le predizioni del classificatore



Symmetric: scarsa varianza, metriche migliori



RoT: alta varianza, metriche peggiori

Per classe

Sbilanciamento di LODB peggiora ulteriormente le prestazioni

Classe	Precision	Recall	F1-score
DiaL.	0.7333	0.3438	0.4681
O2DiaR.	0.2381	0.0847	**0.1250**
O2Hor.	0.4545	0.1923	**0.2703**
Hor.	0.9630	0.8254	0.8889
Cent.	0.8129	0.7079	0.7568
Pat.	0.9091	0.7843	0.8421
Radi.	0.5294	0.4286	0.4737
RoT.R	0.8333	0.5263	0.6452
Ver.	0.9615	0.8242	0.8876
O3Li.	0.4444	0.1143	**0.1818**
RoT.L	0.7333	0.2500	0.3729
DiaR.	0.7895	0.3261	0.4615
DiaX.	0.9308	0.8605	0.8943
O3Tri.	0.3333	0.0417	**0.0741**
O2DiaL.	0.1429	0.0159	**0.0286**
Oline	0.6250	0.2174	**0.3226**
Tri.	0.8182	0.7079	0.7590

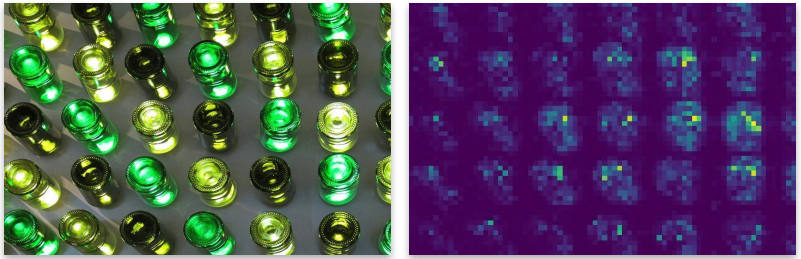
Metriche per classe su LODB con linear layer

Nome	Descrizione	Numero
Cent.	Soggetto centrato nell'immagine	947
RoT.L	Regola dei terzi, soggetto allineato con la griglia a sinistra	214
RoT.R	Regola dei terzi, soggetto allineato con la griglia a destra	275
O2Dia.L	Due oggetti disposti diagonalmente, dall'angolo in alto a sinistra	301
O2Dia.R	Due oggetti disposti diagonalmente, dall'angolo in alto a destra	293
O2Hor.	Due oggetti posizionati su una linea orizzontale	386
03Li.	Tre oggetti posizionati sulla stessa linea	157
03Tri.	Composizione triangolare formata da 3 oggetti	115
Oline.	Più di tre oggetti posizionati sulla stessa linea	110
Pat.	L'immagine contiene dei patterns	255
DiaL.	Composizione a diagonale, dall'angolo in alto a sinistra	446
DiaR.	Composizione a diagonale, dall'angolo in alto a destra	236
Hor.	Struttura orizzontale	574
Tri.	Struttura triangolare	451
Ver.	Struttura verticale	455
Radi.	Struttura radiale	187
DiaX.	Struttura comprendente entrambe le diagonali	861

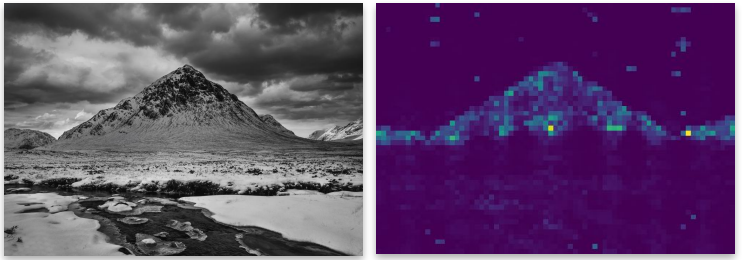
Composizione del dataset LODB

Sviluppi Futuri

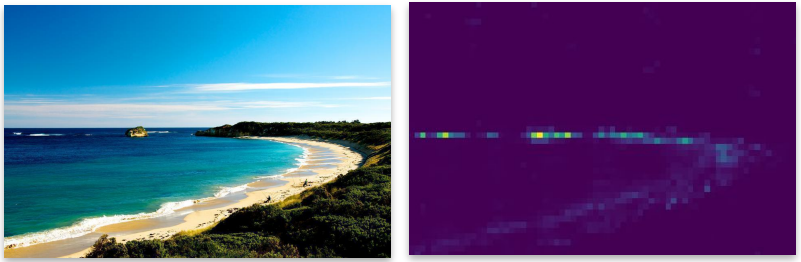
Usare una CNN sulle mappe di self-attention del ViT



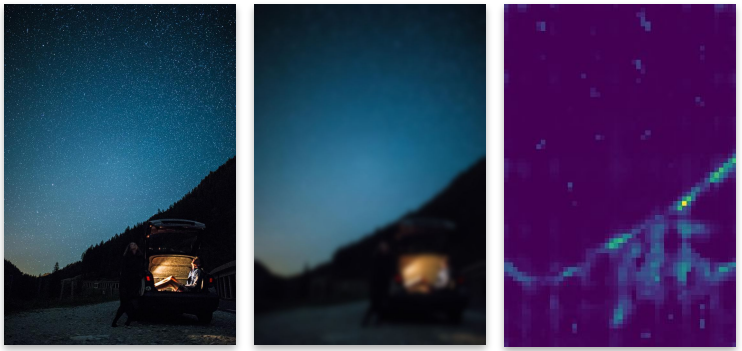
(a) *Pattern*



(b) *Triangle*



(c) *Curved, Horizontal*



(d) *RoT, Diagonal*

Riferimenti

[1] Mathilde Caron et al. "Emerging properties in self-supervised vision transformers". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 9650–966

[2] Maxime Oquab et al. "DINOv2: Learning robust visual features without supervision". In: *arXiv preprint arXiv:2304.07193* (2023)

[3] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020)

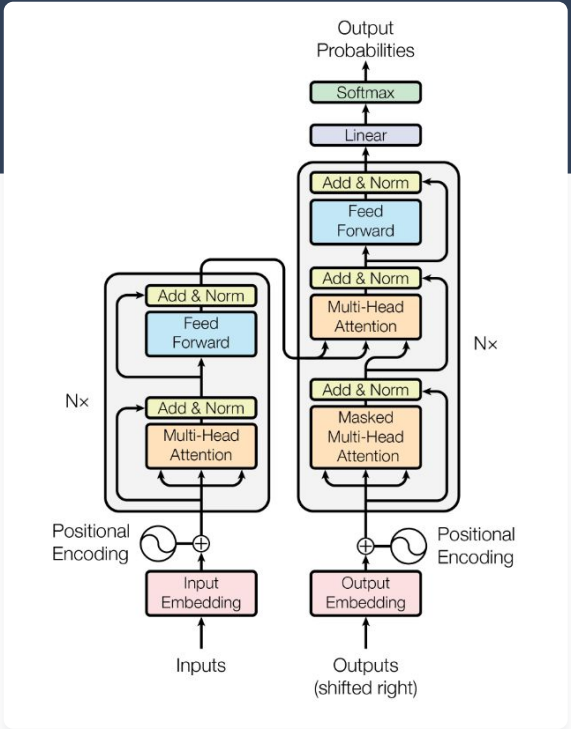
[4] Jun-Tae Lee et al. "Photographic Composition Classification and Dominant Geometric Element Detection for Outdoor Scenes". In: *Journal of Visual Communication and Image Representation 55* (mag. 2018)

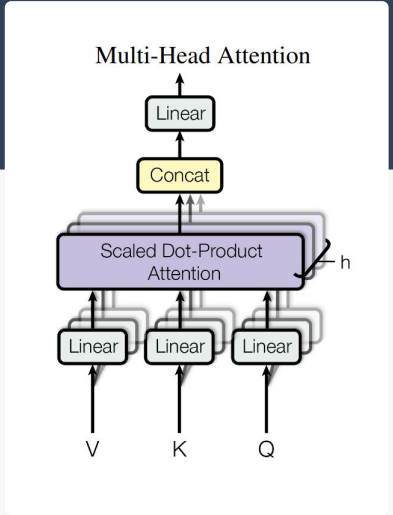
[5] Zhaoran Zhao et al. "Self-supervised Photographic Image Layout Representation Learning". In: *arXiv preprint arXiv:2403.03740* (2024)

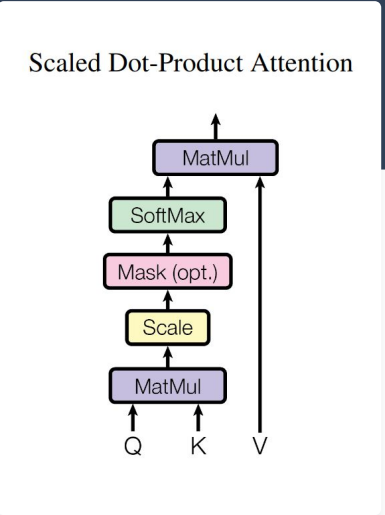
[6] Yaoting Wang et al. "Spatial-invariant convolutional neural network for photographic composition prediction and automatic correction". In: *Journal of Visual Communication and Image Representation 90* (2023), p. 103751.

Extra

Transformer







$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Loss

$$\text{BCEwithLogits}(x, y) = BCE(\sigma(x), y)$$

y = classificazione reale (0, 1)

x = logit

Sigmoide:

$$p = \sigma(x) = \frac{1}{1 + e^{-x}}$$

Binary cross-entropy:

$$BCE(p, y) = -(y \log(p) + (1 - y) \log(1 - p))$$

Loss (MLP)

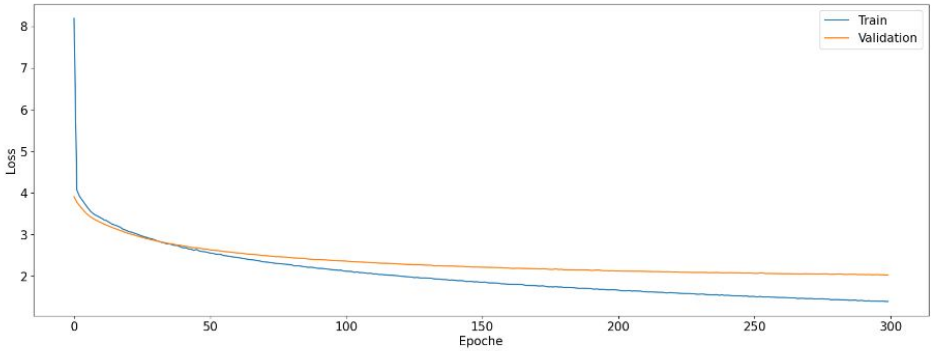


Figura 6.4: Grafico della loss in validation (quinto fold) su LODB e MLP per la configurazione 4 token CLS e 4 AVG patch, 300 epoche, learning rate 0.002, che viene usata in fase di test. Non ci sono segnali di overfitting.

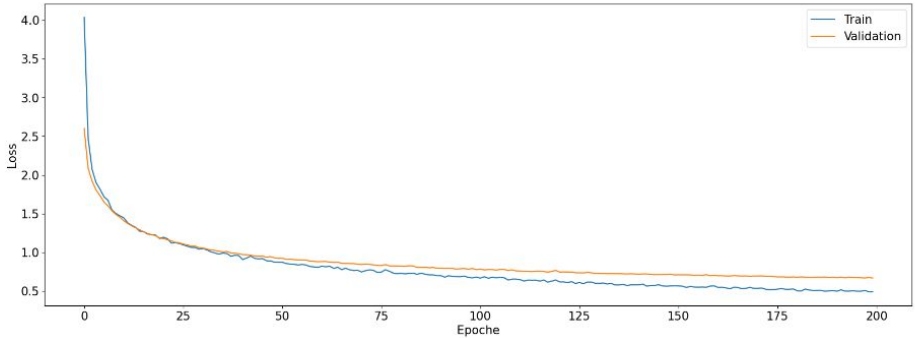


Figura 6.3: Grafico della loss in validation (quinto fold) su KU-PCP e MLP per la configurazione 4 token CLS e 4 AVG patch, 200 epoche, learning rate 0.005, che viene usata in fase di test. Non ci sono segnali di overfitting.

Loss (Linear)

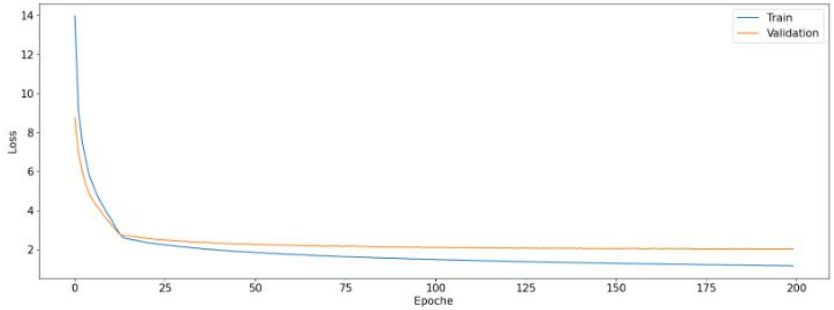


Figura 6.1: Grafico della loss in validation (quinto fold) su LODB e linear layer per la configurazione 4 token CLS e 4 AVG patch, 200 epoche, learning rate 0.003, che viene usata in fase di test. Non ci sono segnali di overfitting.

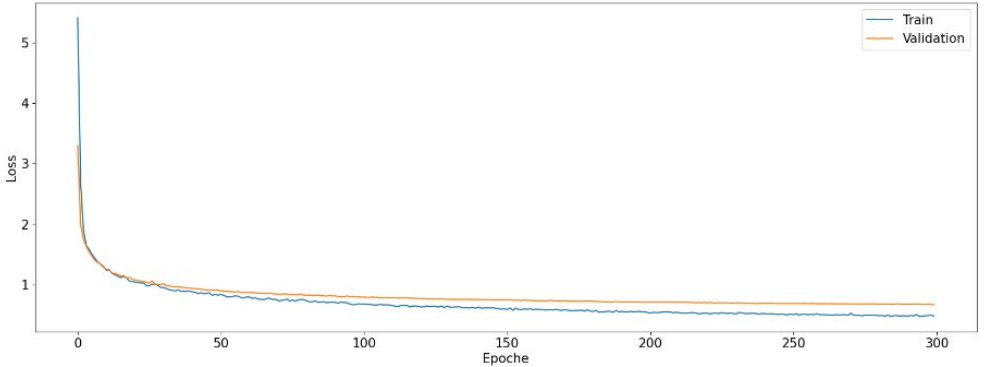


Figura 6.2: Grafico della loss in validation (quinto fold) su KU-PCP e linear layer per la configurazione con 4 token CLS e 4 AVG patch, 300 epoche e learning rate 0.003, che viene usata in fase di test. La discesa della loss non mostra segni di overfitting.

Metriche

$$Precision = \frac{TP}{TP + FP}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

$$Recall = \frac{TP}{TP + FN}$$

$$AP = \sum_{n} (R_n - R_{n-1}) P_n$$

R = recall, P = precision, n-esimo threshold

Matrici di confusione

	$\sim DiaX$	$DiaX$
$\sim DiaX$	805	**13**
$DiaX$	19	**111**

	$\sim O3Tri$	$O3Tri$
$\sim O3Tri$	937	**0**
$O3Tri$	11	**0**

Tabella 6.3: Matrici di confusione del quinto fold per *DiaX* e *O3Tri* nella configurazione 8 CLS e lr = 0.002 e nel fold 5.

Features

[CLS] e AVG usati singolarmente

Features	lr	**Acc. (Any)**	**Acc. (All)**	**MP**	**mAP**
CLS_{12}	0.05	**0.52**	0.42	0.46	0.49
CLS_9 ... CLS_{12}	0.006	0.45	0.44	0.65	0.58
CLS_5 ... CLS_{12}	0.002	0.38	0.37	0.58	0.56
CLS_1 ... CLS_{12}	0.0005	0.20	0.19	0.33	0.47

Tabella 6.1: Un confronto fra le prestazioni del layer lineare utilizzando diverse combinazioni di layer di token [CLS]

Features	lr	**Acc. (Any)**	**Acc. (All)**	**MP**	**mAP**
AVG_{12}	0.06	**0.52**	**0.49**	**0.61**	**0.59**
AVG_9 ... AVG_{12}	0.005	0.43	0.42	0.60	0.59
AVG_5 ... AVG_{12}	0.002	0.41	0.41	0.55	0.56
AVG_1 ... AVG_{12}	0.001	0.35	0.34	0.51	0.54

Tabella 6.4: Un confronto fra le prestazioni del layer lineare utilizzando diverse combinazioni di layer di AVG patch

Features

[CLS] e AVG usati in coppia

Features	lr	**Acc. (Any)**	**Acc. (All)**	**MP**	**mAP**
$CLS_{12}AVG_{12}$	0.01	0.49	0.46	0.58	0.56
CLS_9 ... CLS_{12} AVG_9 ... AVG_{12}	0.005	**0.51**	**0.49**	**0.64**	**0.63**
CLS_5 ... $CLS_{12}AVG_5$ ... AVG_{12}	0.001	0.42	0.41	0.63	0.60
CLS_1 ... CLS_4CLS_9 ... CLS_{12} AVG_1 ... AVG_4AVG_9 ... AVG_{12}	0.001	0.39	0.38	0.56	0.58

Tabella 6.5: Un confronto fra le prestazioni del layer lineare utilizzando diverse combinazioni di layer di token [CLS] e AVG patch.

Predizioni

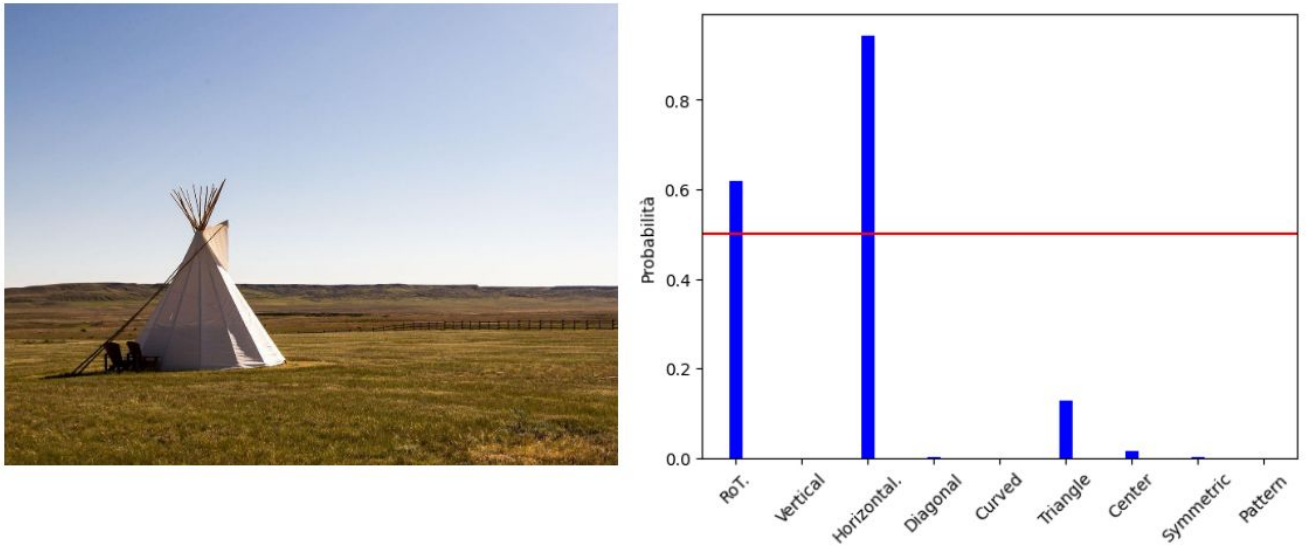


Figura 6.5: Esempio di predizione corretta della rete con linear classifier su KU-PCP. Le classi di ground truth sono *RoT.* e *Horizontal.* In rosso è segnata la soglia di threshold (posta a 0.5) utilizzata per il calcolo delle metriche.

Predizioni

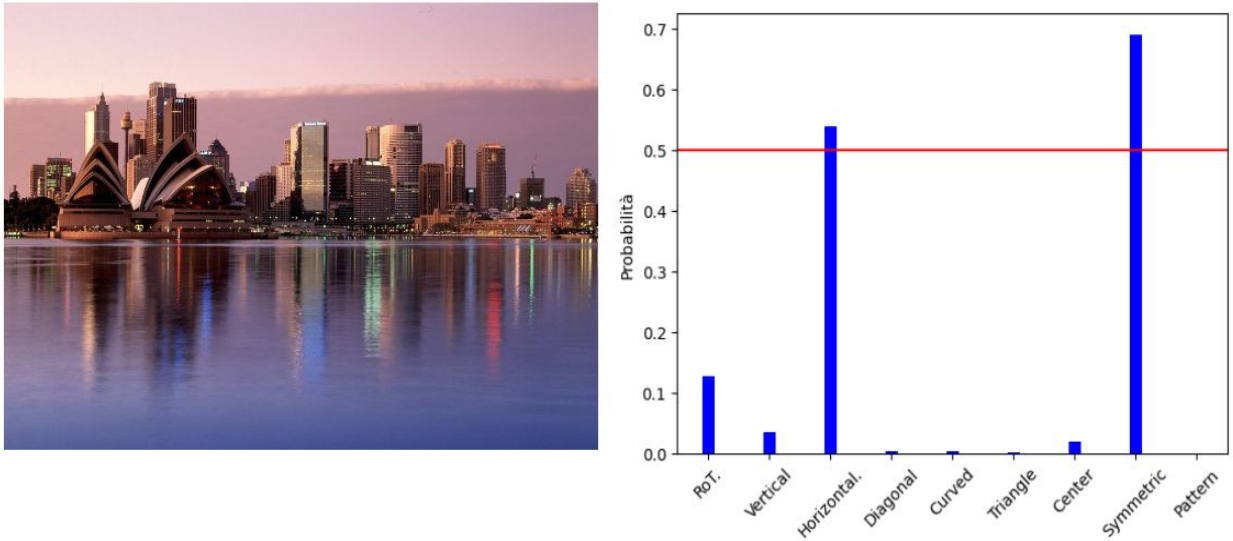


Figura 6.6: Predizione con classificatore lineare su KU-PCP. Le classi di ground truth sono *Vertical* e *Symmetric*, mentre la nostra rete predice *Horizontal* e *Symmetric*.

Predizioni

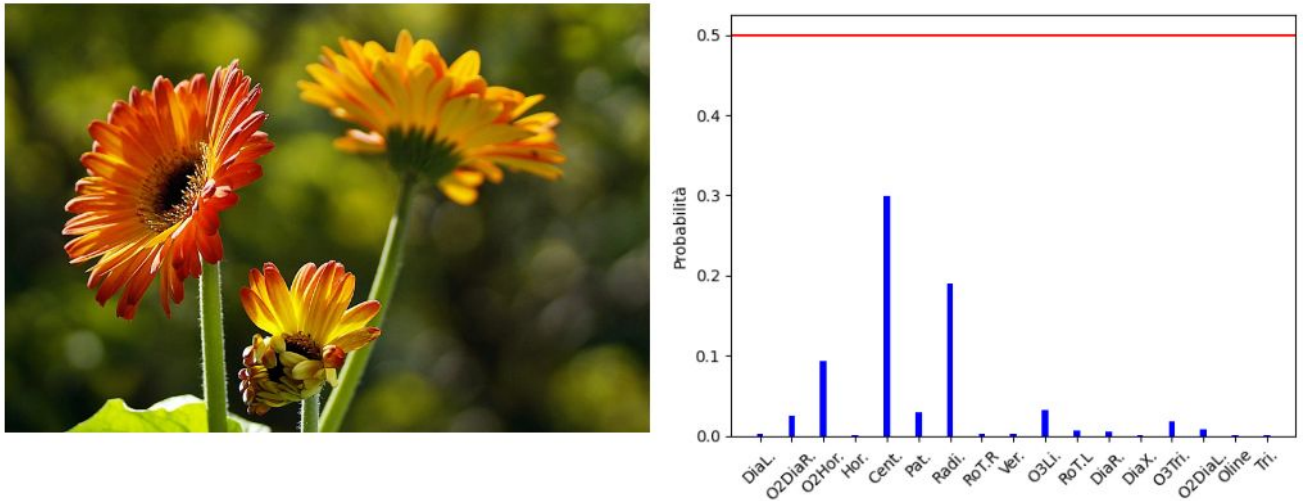


Figura 6.8: Predizione con classificatore lineare su LODB, ground truth *O3Tri.*.

Predizioni

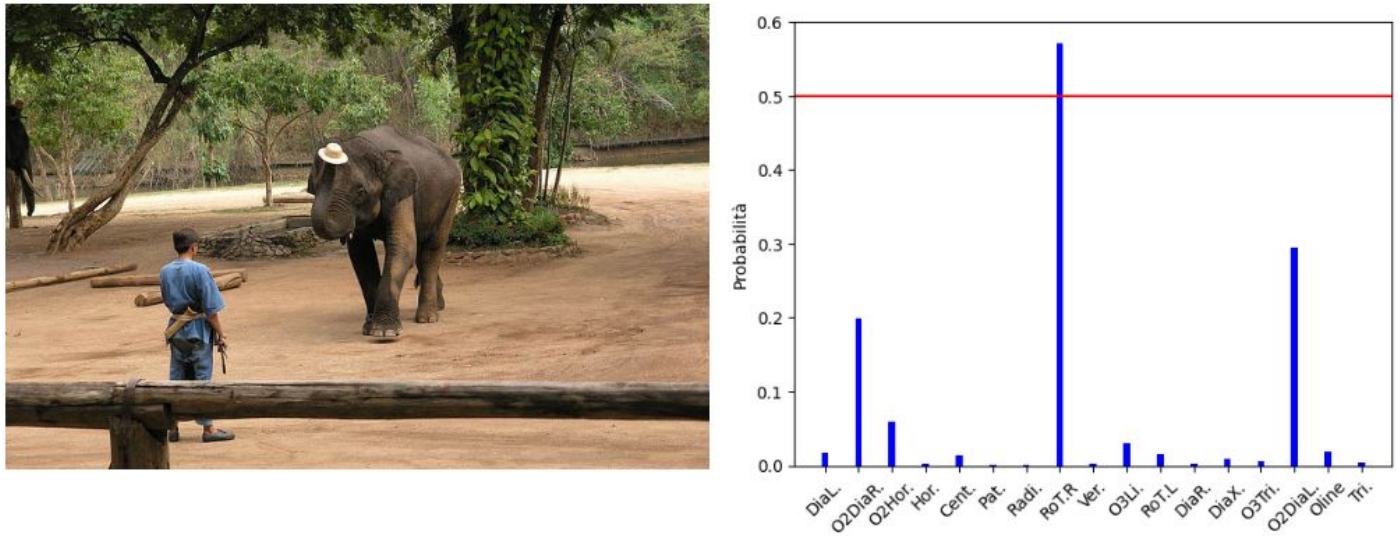


Figura 6.9: Predizione con classificatore lineare su LODB, ground truth *O2DiaR*.

Predizioni

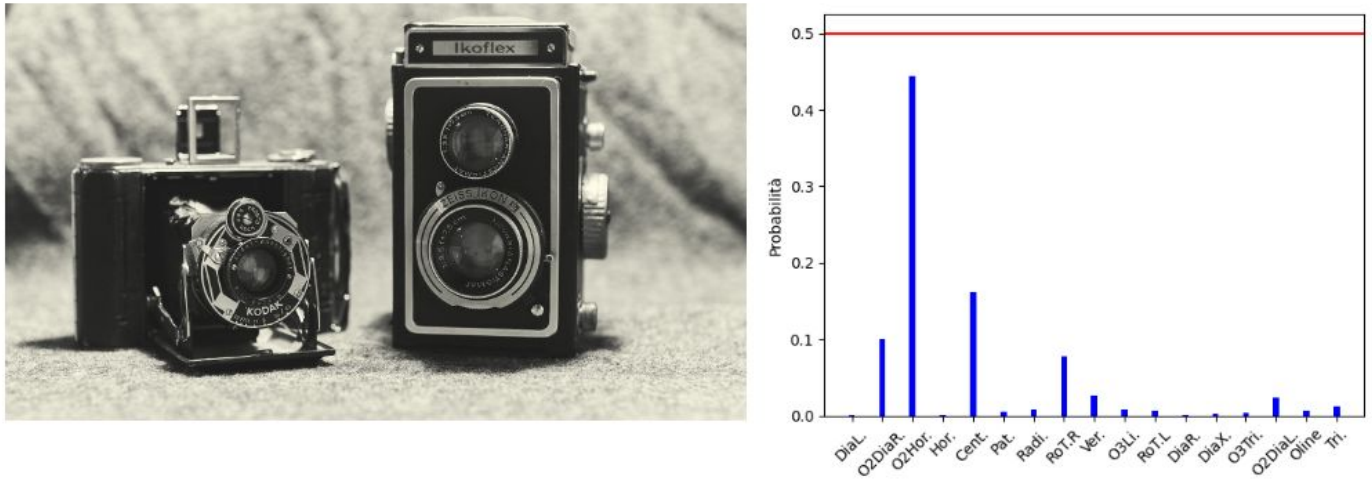


Figura 6.10: Predizione con classificatore lineare su LODB, ground truth *O2Hor.*. Anche su immagini la cui composizione risulta essere piuttosto chiara all'occhio umano, il modello fatica a riconoscerla e spalma le sue predizioni su diverse classi, facendo fatica a prendere una decisione.

Misclassification

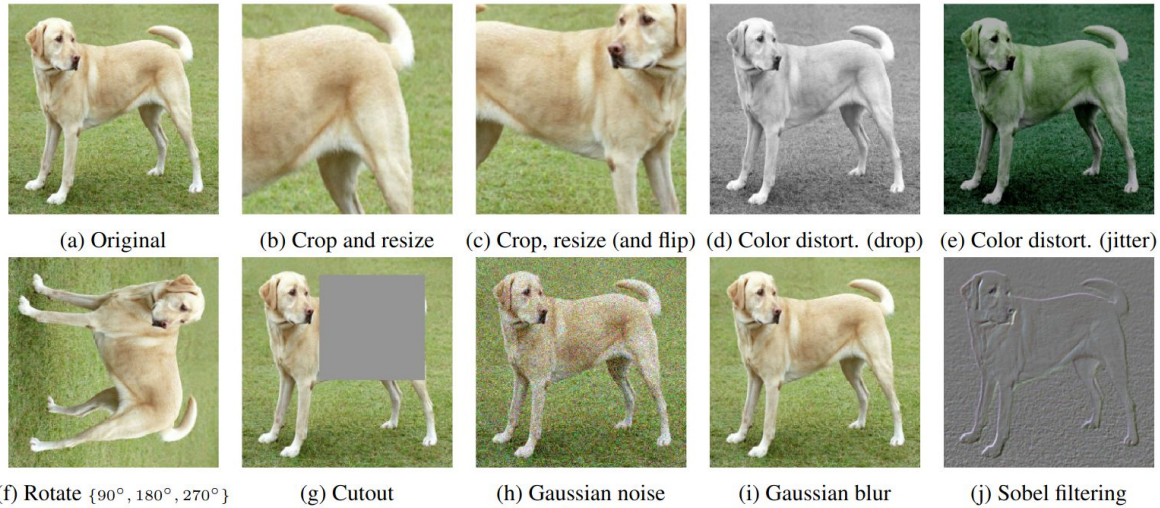
(a) Vertical

(b) O2DiaL.

(c) O3Tri.

Figura 7.4: Esempi di misclassification/forzatura della composizione. Sono indicate le etichette che LODB assegna a ciascuna.

Augmentations



(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

(f) Rotate $\{90°, 180°, 270°\}$ (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

Momentum Encoder

Aggiornamento dei pesi della rete Teacher con Exponential Moving Average dei pesi dello Student

$$\xi \leftarrow \tau\xi + (1-\tau)\theta$$

- ξ pesi del Teacher
- θ pesi dello Student
- τ momentum coefficient in [0, 1]

  inzia a 0.996, aumenta a 1 durante training

Stop Gradient (SG) evita che il Teacher venga aggiornato tramite backpropagation