AET vs. AED: Unsupervised Representation Learning by Auto-Encoding Transformations rather than Data

CVPR2019 Oral

2019.07.25

발표자 박성현

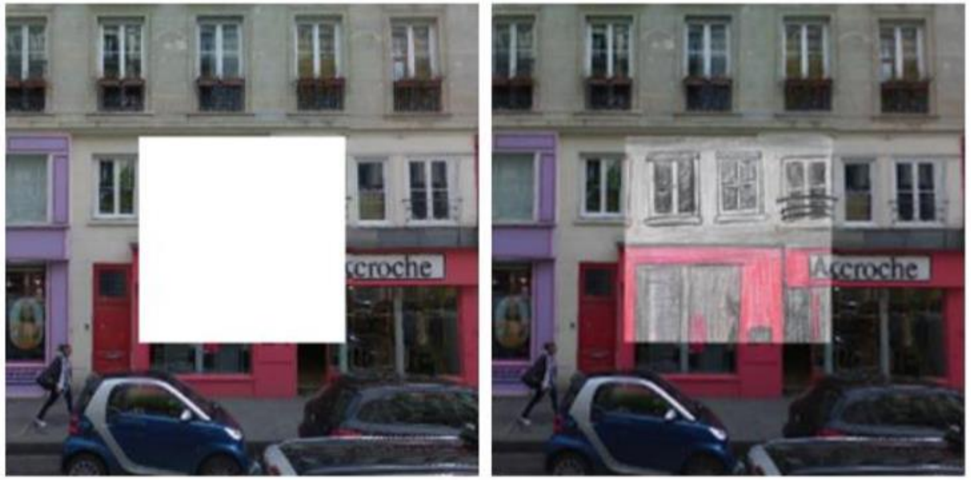
DAVIAN
Data and Visual Analytics Lab

KOREA
UNIVERSITY

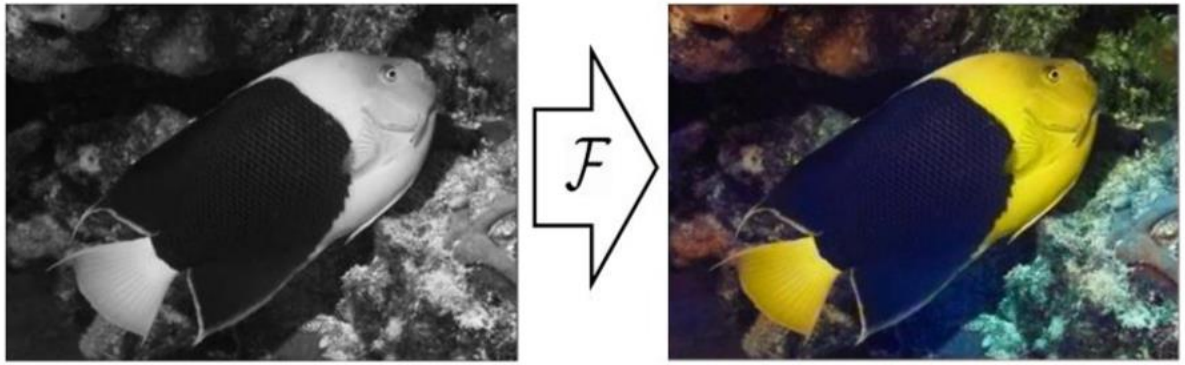
1 Introduction

Self-Supervision

- A form of unsupervised learning where the data provides the supervision
- In general, withhold some part of the data, and task the network with predicting it

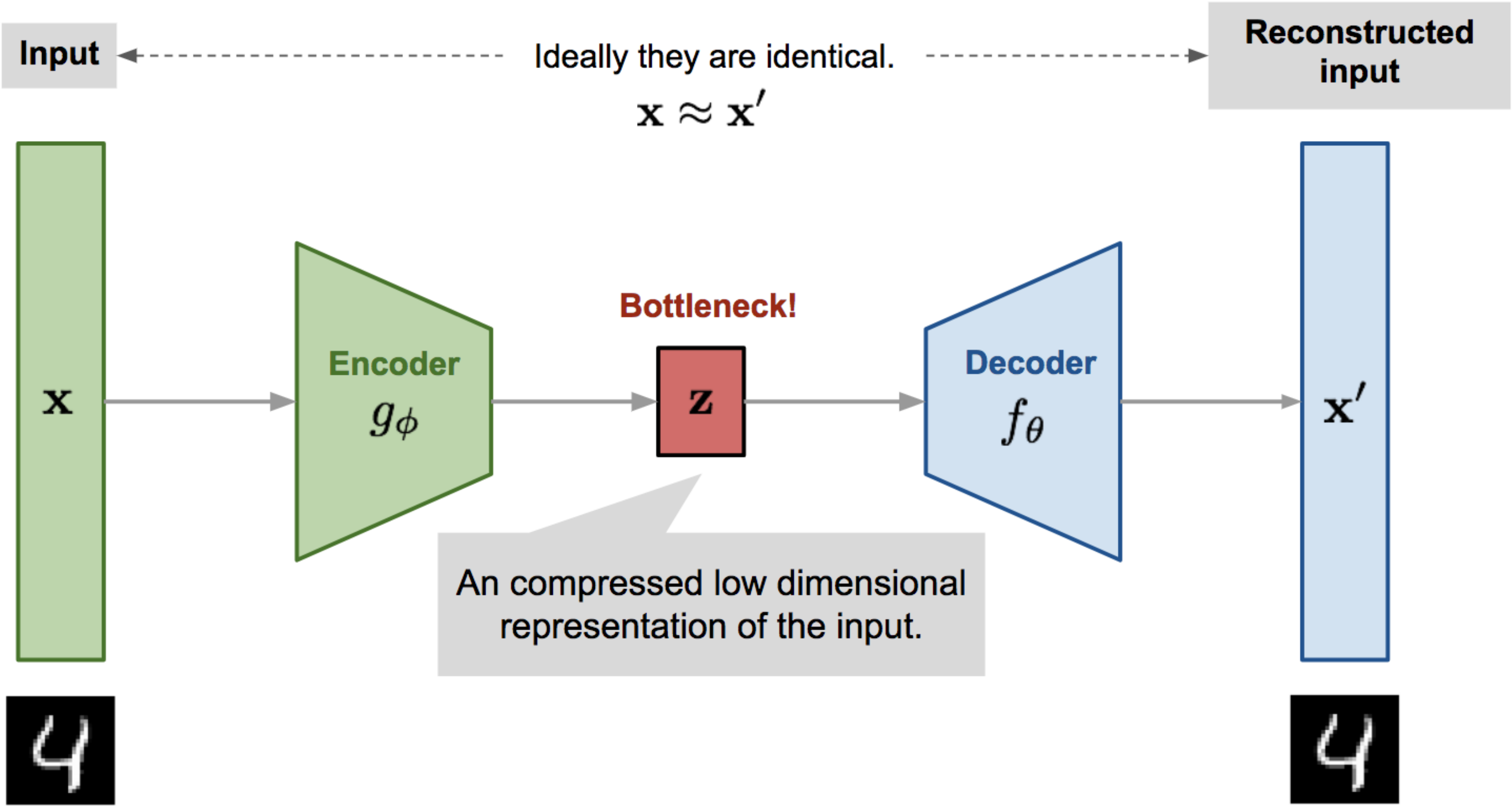


(a) Input context (b) Human artist



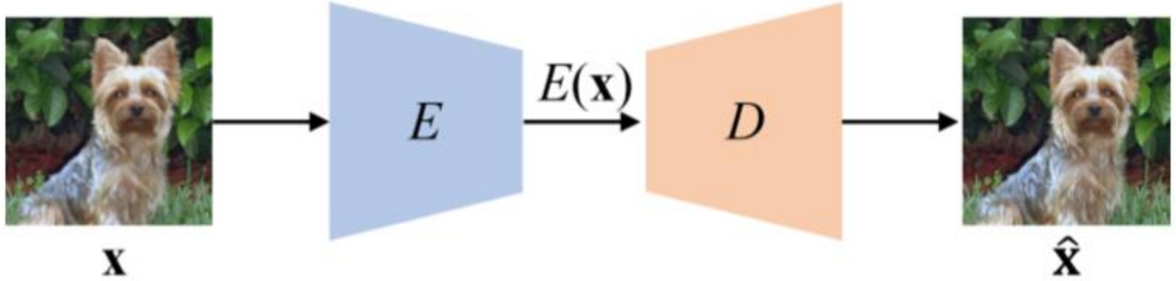
1 Introduction

Representation Learning

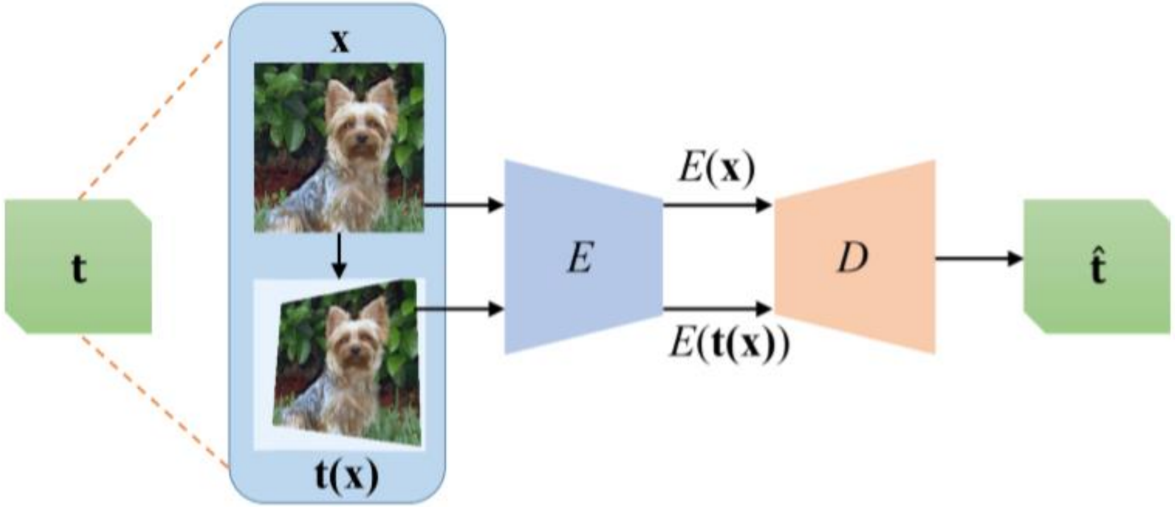

Input
Ideally they are identical.
$\mathbf{x} \approx \mathbf{x}'$
Reconstructed input
$\mathbf{x}$
Encoder
g_ϕ
Bottleneck!
$\mathbf{z}$
Decoder
f_θ
$\mathbf{x}'$
An compressed low dimensional representation of the input.

2 Model

AED & AET



(a) Auto-Encoding Data (AED)



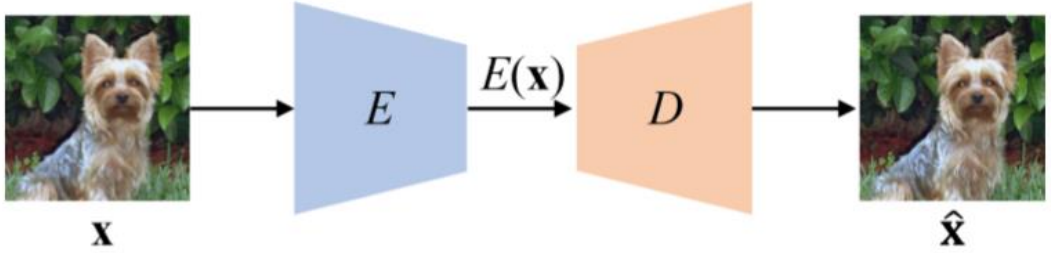
(b) Auto-Encoding Transformation (AET)

$$\min_{E,D} \mathop{\mathbb{E}}_{\mathbf{t}\sim\mathcal{T},\mathbf{x}\sim\mathcal{X}} \ell(\mathbf{t},\hat{\mathbf{t}})$$

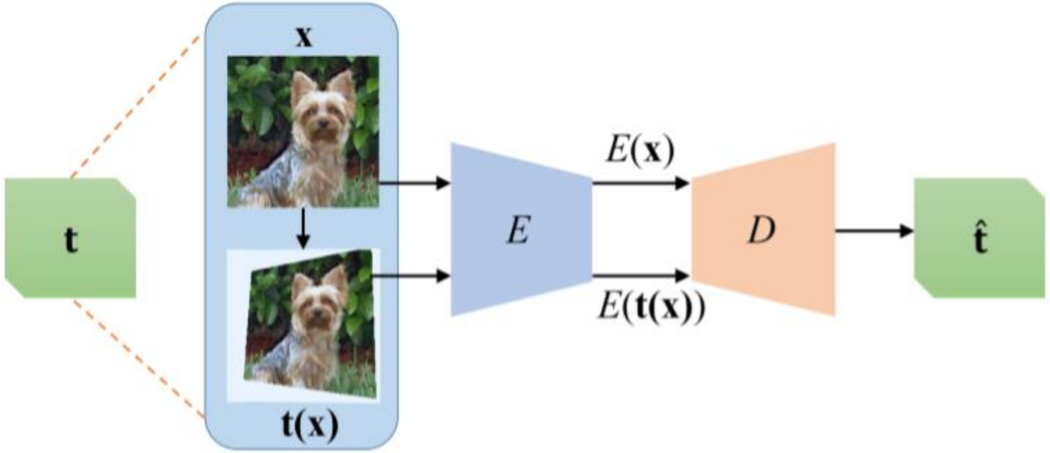
$$\hat{\mathbf{t}} = D\left[E(\mathbf{x}), E(\mathbf{t}(\mathbf{x}))\right]$$

2 Model

The AET Family



(a) Auto-Encoding Data (AED)



(b) Auto-Encoding Transformation (AET)

- **Parameterized Transformations**

Parameterized matrix $M(\boldsymbol{\theta}) \in \mathbb{R}^{3\times 3}$

$$\ell(\mathbf{t}_{\boldsymbol{\theta}}, \mathbf{t}_{\hat{\boldsymbol{\theta}}}) = \frac{1}{2}\|M(\boldsymbol{\theta}) - M(\hat{\boldsymbol{\theta}})\|_2^2$$

- **GAN-Induced Transformations**

$$\mathbf{t}_{\mathbf{z}}(\mathbf{x}) = G(\mathbf{x}, \mathbf{z})$$

$$\ell(\mathbf{t}_{\mathbf{z}}, \mathbf{t}_{\hat{\mathbf{z}}}) = \frac{1}{2}\|\mathbf{z} - \hat{\mathbf{z}}\|_2^2$$

- **Non-Parametric Transformations**

$$\ell(\mathbf{t}, \hat{\mathbf{t}}) = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{X}} \operatorname{dist}(\mathbf{t}(\mathbf{x}), \hat{\mathbf{t}}(\mathbf{x}))$$

3 Experiments

Training and Evaluating AET

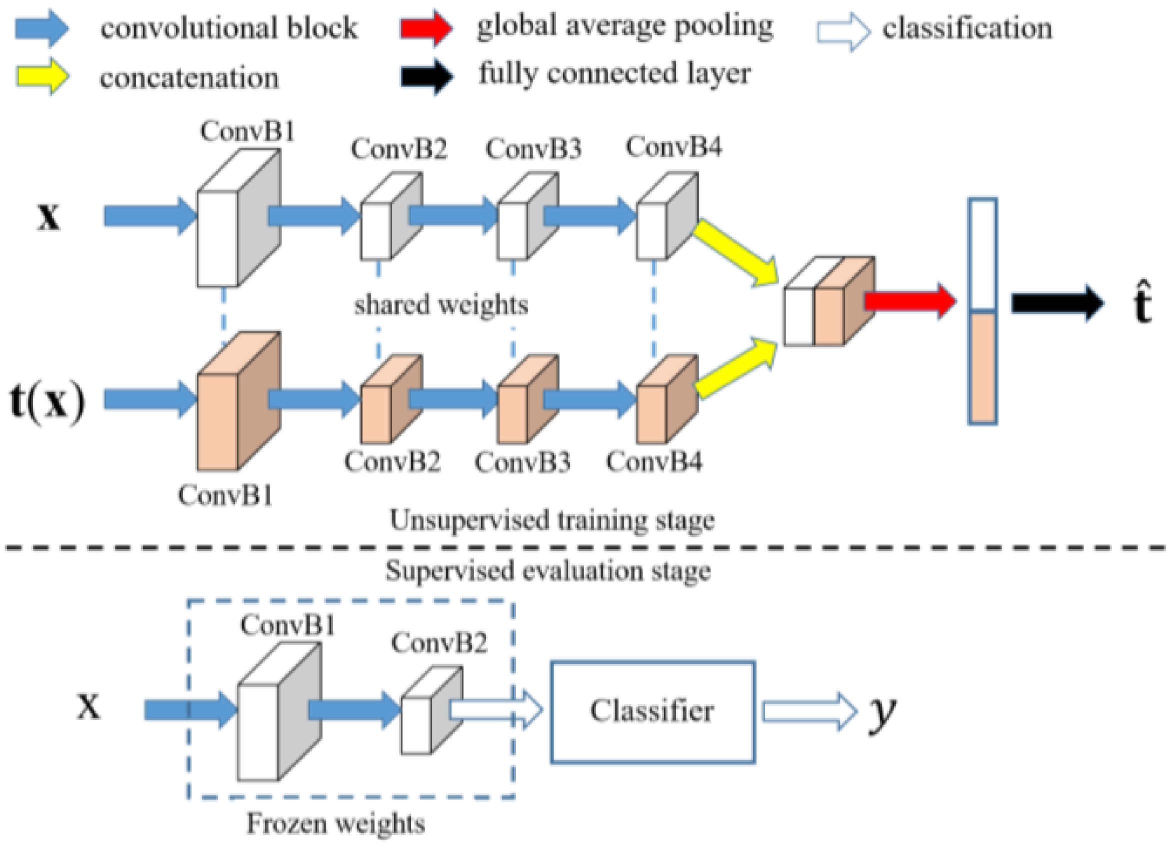


Figure 2: An illustration of the network architectures for training and evaluating AET on the CIFAR-10 dataset.

3 Experiments

CIFAR-10 Experiments

- **AET-affine**

Random rotation [−180, 180]
Random translation ±0.2
Random scaling [0.7, 1.3]
Random shearing [−30, 30]

- **AET-projective**

(Four corners of an image)
Random translate ±0.125
Random scaling [0.8, 1.2]
Random rotation 0, 90, 180, 270

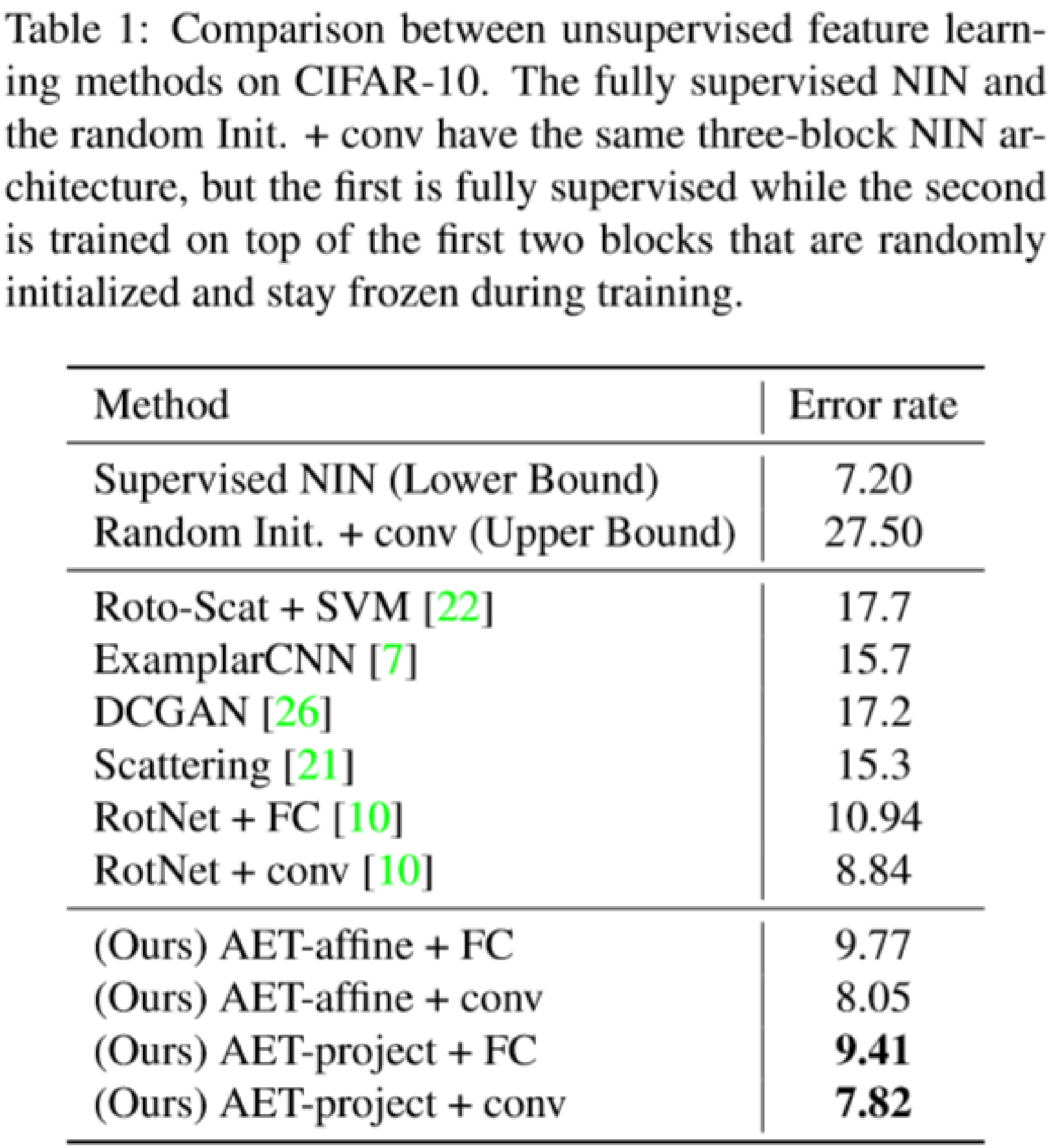


Table 1: Comparison between unsupervised feature learning methods on CIFAR-10. The fully supervised NIN and the random Init. + conv have the same three-block NIN architecture, but the first is fully supervised while the second is trained on top of the first two blocks that are randomly initialized and stay frozen during training.

Method	Error rate
Supervised NIN (Lower Bound)	7.20
Random Init. + conv (Upper Bound)	27.50
Roto-Scat + SVM [22]	17.7
ExamplarCNN [7]	15.7
DCGAN [26]	17.2
Scattering [21]	15.3
RotNet + FC [10]	10.94
RotNet + conv [10]	8.84
(Ours) AET-affine + FC	9.77
(Ours) AET-affine + conv	8.05
(Ours) AET-project + FC	**9.41**
(Ours) AET-project + conv	**7.82**

3 Experiments

CIFAR-10 Experiments

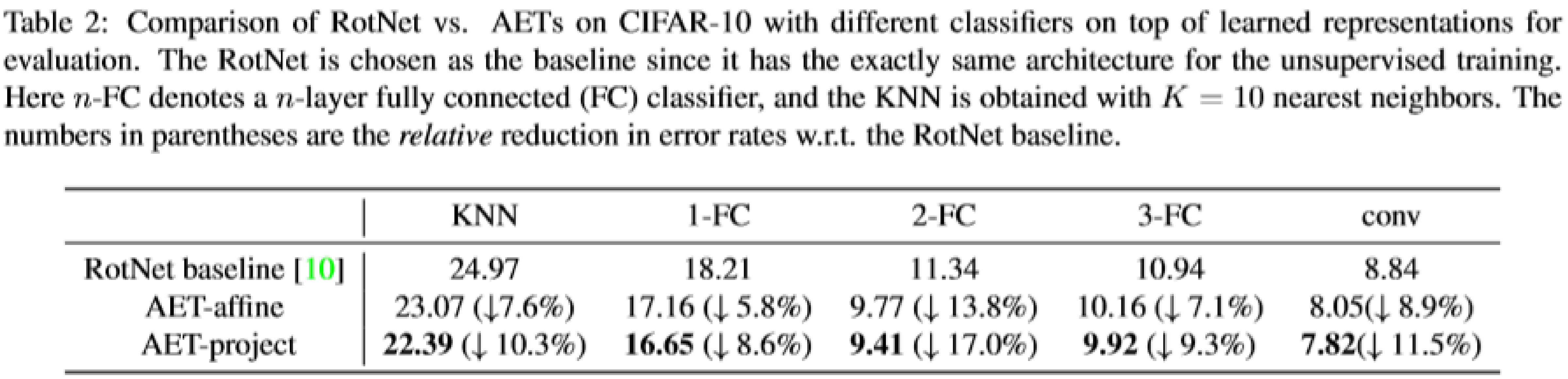


Table 2: Comparison of RotNet vs. AETs on CIFAR-10 with different classifiers on top of learned representations for evaluation. The RotNet is chosen as the baseline since it has the exactly same architecture for the unsupervised training. Here n-FC denotes a n-layer fully connected (FC) classifier, and the KNN is obtained with $K = 10$ nearest neighbors. The numbers in parentheses are the *relative* reduction in error rates w.r.t. the RotNet baseline.

	KNN	1-FC	2-FC	3-FC	conv
RotNet baseline [10]	24.97	18.21	11.34	10.94	8.84
AET-affine	23.07 (↓7.6%)	17.16 (↓ 5.8%)	9.77 (↓ 13.8%)	10.16 (↓ 7.1%)	8.05(↓ 8.9%)
AET-project	**22.39** (↓ 10.3%)	**16.65** (↓ 8.6%)	**9.41** (↓ 17.0%)	**9.92** (↓ 9.3%)	**7.82**(↓ 11.5%)

3 Experiments

CIFAR-10 Experiments

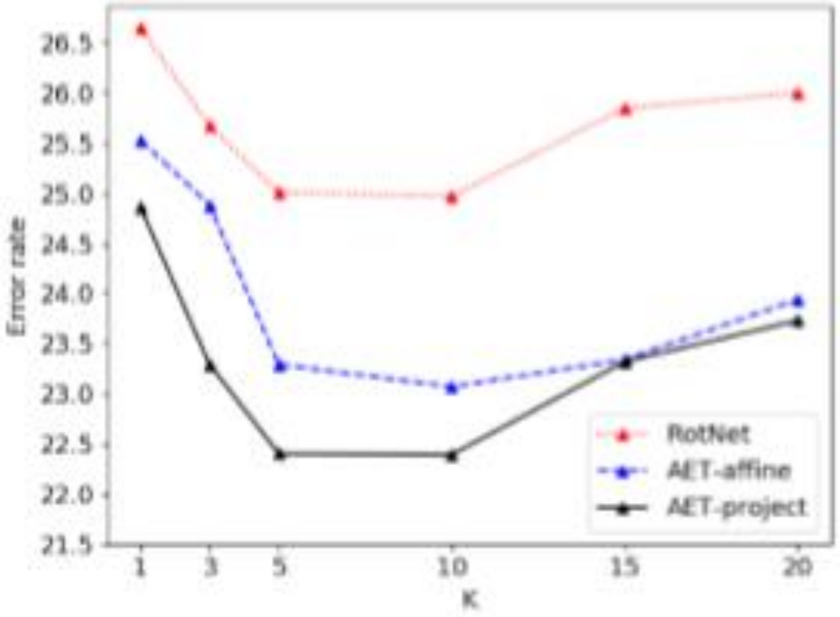


Figure 3: The comparison of the KNN error rates by different models with varying numbers K of nearest neighbors on CIFAR-10.

3

Experiments

ImageNet Experiments

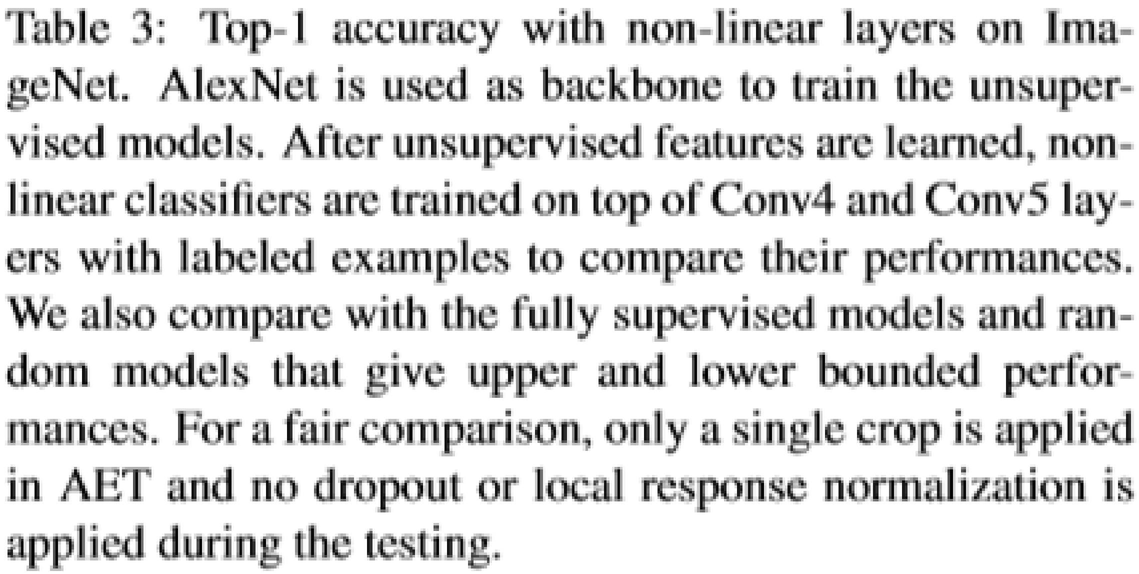
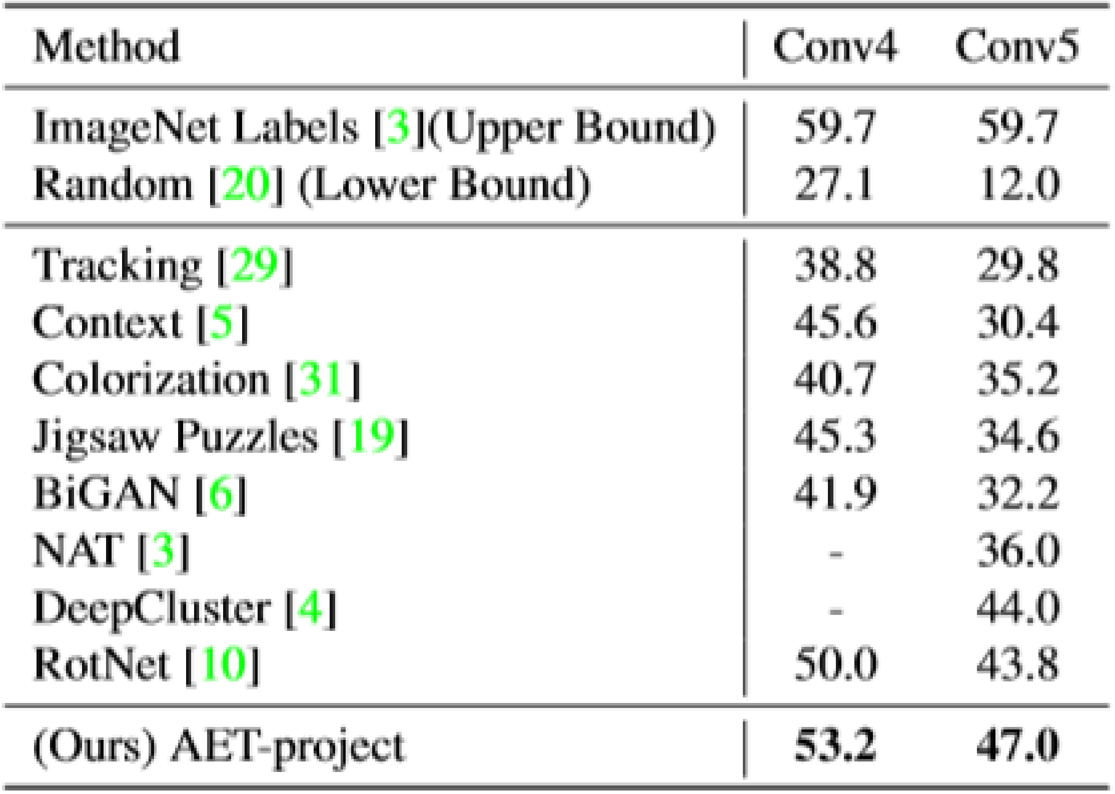


Table 3: Top-1 accuracy with non-linear layers on ImageNet. AlexNet is used as backbone to train the unsupervised models. After unsupervised features are learned, non-linear classifiers are trained on top of Conv4 and Conv5 layers with labeled examples to compare their performances. We also compare with the fully supervised models and random models that give upper and lower bounded performances. For a fair comparison, only a single crop is applied in AET and no dropout or local response normalization is applied during the testing.



Method	Conv4	Conv5
ImageNet Labels [3](Upper Bound)	59.7	59.7
Random [20] (Lower Bound)	27.1	12.0
Tracking [29]	38.8	29.8
Context [5]	45.6	30.4
Colorization [31]	40.7	35.2
Jigsaw Puzzles [19]	45.3	34.6
BiGAN [6]	41.9	32.2
NAT [3]	-	36.0
DeepCluster [4]	-	44.0
RotNet [10]	50.0	43.8
(Ours) AET-project	**53.2**	**47.0**

3 Experiments

ImageNet Experiments

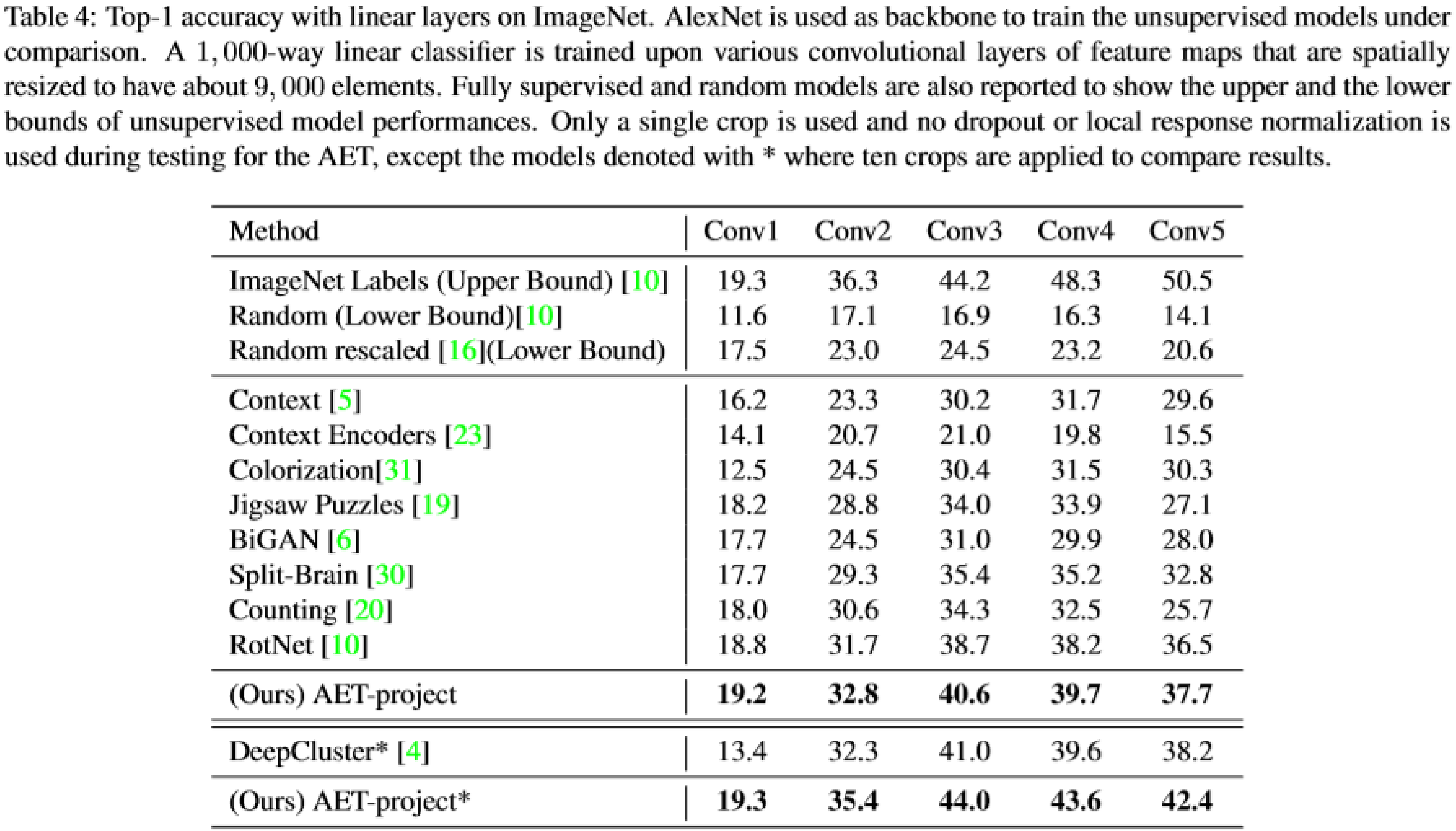


Table 4: Top-1 accuracy with linear layers on ImageNet. AlexNet is used as backbone to train the unsupervised models under comparison. A $1,000$-way linear classifier is trained upon various convolutional layers of feature maps that are spatially resized to have about $9,000$ elements. Fully supervised and random models are also reported to show the upper and the lower bounds of unsupervised model performances. Only a single crop is used and no dropout or local response normalization is used during testing for the AET, except the models denoted with * where ten crops are applied to compare results.

Method	Conv1	Conv2	Conv3	Conv4	Conv5
ImageNet Labels (Upper Bound) [10]	19.3	36.3	44.2	48.3	50.5
Random (Lower Bound)[10]	11.6	17.1	16.9	16.3	14.1
Random rescaled [16](Lower Bound)	17.5	23.0	24.5	23.2	20.6
Context [5]	16.2	23.3	30.2	31.7	29.6
Context Encoders [23]	14.1	20.7	21.0	19.8	15.5
Colorization[31]	12.5	24.5	30.4	31.5	30.3
Jigsaw Puzzles [19]	18.2	28.8	34.0	33.9	27.1
BiGAN [6]	17.7	24.5	31.0	29.9	28.0
Split-Brain [30]	17.7	29.3	35.4	35.2	32.8
Counting [20]	18.0	30.6	34.3	32.5	25.7
RotNet [10]	18.8	31.7	38.7	38.2	36.5
(Ours) AET-project	**19.2**	**32.8**	**40.6**	**39.7**	**37.7**
DeepCluster* [4]	13.4	32.3	41.0	39.6	38.2
(Ours) AET-project*	**19.3**	**35.4**	**44.0**	**43.6**	**42.4**

3 Experiments

Places Experiments

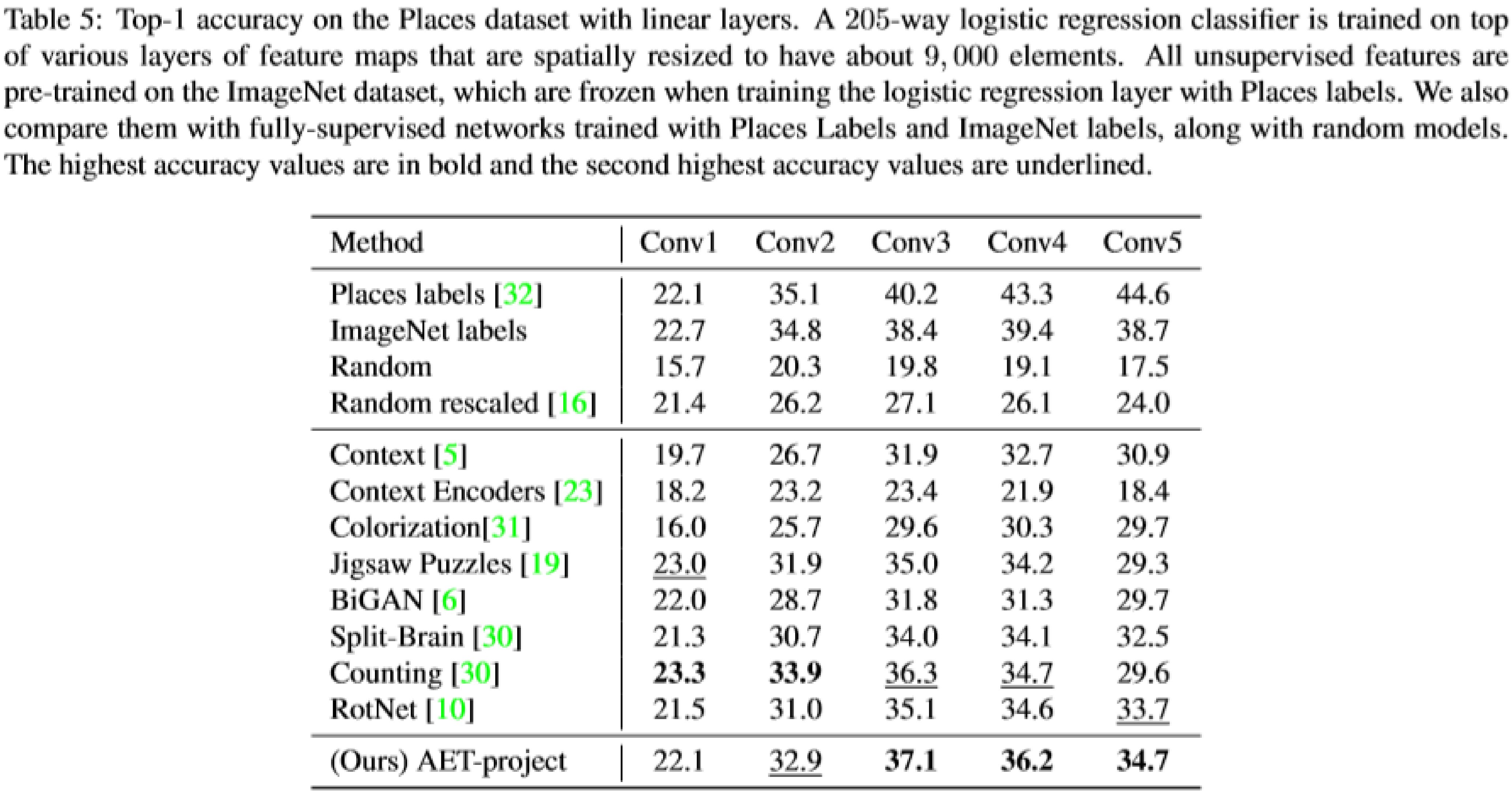
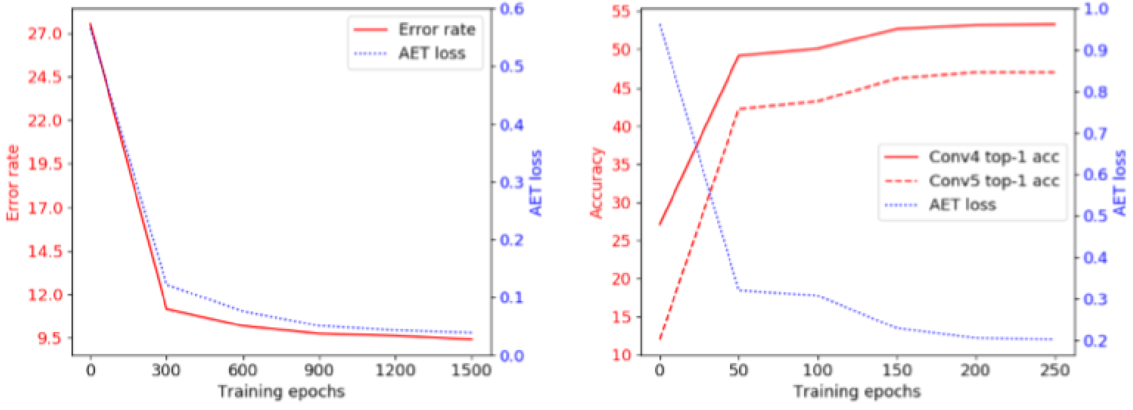


Table 5: Top-1 accuracy on the Places dataset with linear layers. A 205-way logistic regression classifier is trained on top of various layers of feature maps that are spatially resized to have about $9,000$ elements. All unsupervised features are pre-trained on the ImageNet dataset, which are frozen when training the logistic regression layer with Places labels. We also compare them with fully-supervised networks trained with Places Labels and ImageNet labels, along with random models. The highest accuracy values are in bold and the second highest accuracy values are underlined.

Method	Conv1	Conv2	Conv3	Conv4	Conv5
Places labels [32]	22.1	35.1	40.2	43.3	44.6
ImageNet labels	22.7	34.8	38.4	39.4	38.7
Random	15.7	20.3	19.8	19.1	17.5
Random rescaled [16]	21.4	26.2	27.1	26.1	24.0
Context [5]	19.7	26.7	31.9	32.7	30.9
Context Encoders [23]	18.2	23.2	23.4	21.9	18.4
Colorization[31]	16.0	25.7	29.6	30.3	29.7
Jigsaw Puzzles [19]	23.0	31.9	35.0	34.2	29.3
BiGAN [6]	22.0	28.7	31.8	31.3	29.7
Split-Brain [30]	21.3	30.7	34.0	34.1	32.5
Counting [30]	**23.3**	**33.9**	36.3	34.7	29.6
RotNet [10]	21.5	31.0	35.1	34.6	33.7
(Ours) AET-project	22.1	32.9	**37.1**	**36.2**	**34.7**

3 Experiments

Analysis of Predicated Transformations



(a) CIFAR-10

(b) ImageNet

Figure 4: Error rate(top-1 accuracy) vs. AET loss over epochs on the CIFAR-10 and ImageNet datasets.

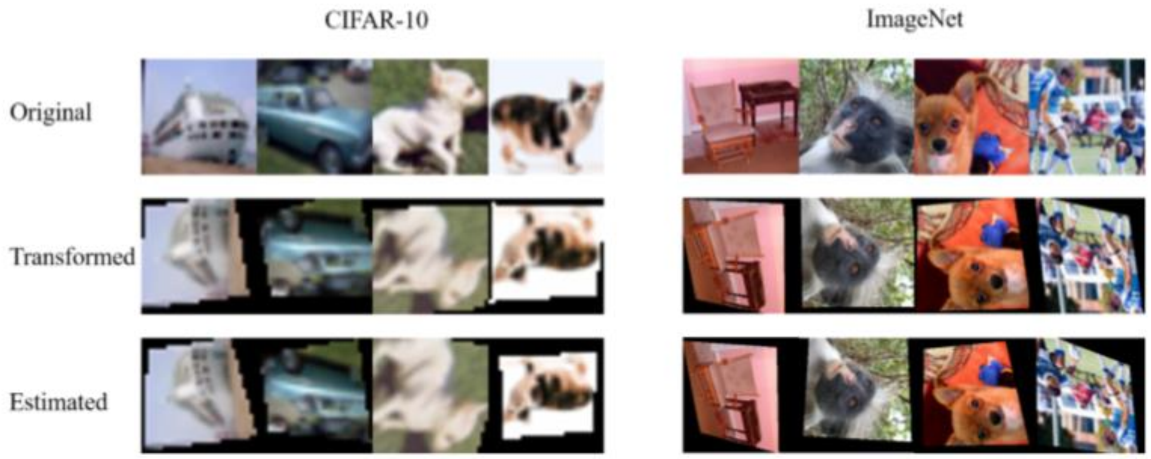


Figure 5: Some examples of original images (top), along with the counterparts of input (middle) and predicted (bottom) transformations by the AET model.