Methods	top-5 err. (%)	CPU (ms)	CPU speed-up rate	GPU (ms)	GPU speed-up rate
VGG16-Net	11.65	1289.28	1×	6.15	1×
Wavelet+CNN	14.42	392.24	3.29×	2.30	2.67×
Decomposition+CNN	12.98	411.63	3.13×	2.37	2.59×
Taylor-1	13.00	-	1.70×	-	2.20×
Taylor-2	15.50	-	2.10×	-	3.40×
ThiNet-Tiny	18.03	116.25	11.25×	1.32	4.66×
ThiNet-GAP	12.08	442.807	2.91×	2.52	2.44×
Ours	11.87	411.63	3.13×	2.37	2.59×

Table 2: Comparison of the top-5 error rate, execution time and speed-up rate on CPU and GPU of VGG16-Net, the two baseline methods and the previous state of the art methods on the ImageNet dataset. The error rate is measured on single-view without data augmentation.

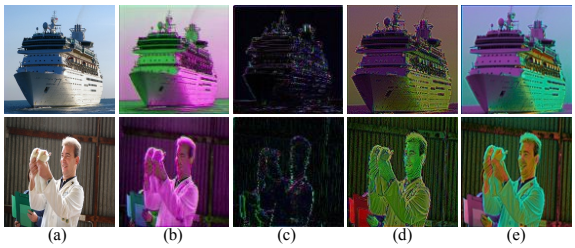

Figure 3: Visualization results of the input image (a), the sub-images (b) and (c) produced by our method and the sub-images (d) and (e) produced by the "Decomposition+CNN".

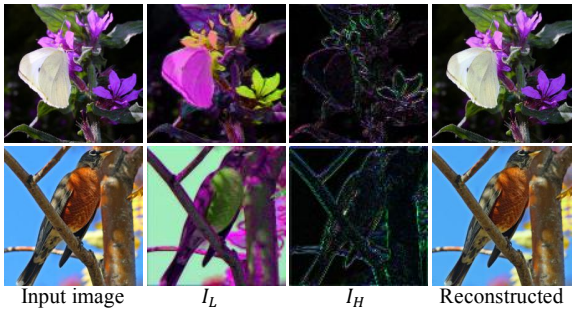

Figure 4: Visualization results of the I_L, I_H and the reconstructed image.

Input	cA	I_R	I_L	I_L+I_H
top-5 err.	15.92	15.73	14.20	11.87

Table 3: Comparison of the top-5 error rate using I_L+I_H, I_L, I_R and cA for classification on the ImageNet dataset.

the classification result. Our model also decomposes the input image into two channels, but it pushes most information to the I_L via minimizing the energy of I_H and are jointly trained to better adapt for classification. We will conduct experiments to analyze the classification performance merely using I_L and cA to give a deeper comparison later. To compare the difference between our model and the "Decomposition+CNN", we visualize the decomposed channels generated by this method and ours in Figure 3. Without the constraints, the two decomposed channels share identical appearance, and fusing the classification results of them can be regarded as the model ensemble. Conversely, the channels generated by our model are somehow complementary, as I_L retains the main content, while I_H preserves the subtle details. These comparisons well prove the proposed WAE can achieve a better balance between speed and accuracy.

Analysis on the decomposed channels Some examples of the decomposed channels and the reconstructed images are visualized in Figure 4. We can observe that I_L indeed contains the main content of the input image, while I_H preserves the details, e.g., edges and contours. It also shows excellent reconstructed results. These visualization results finely accord with the assumption of our method.

To provide deeper analysis of the decomposed channels, we present the performance using the I_L for classification. We first exclude the fusing with I_H and re-train the classification network with the parameters of WAE fixed. The top-5 error rate is depicted in Table 3. It is not surprising that the performance drops, as I_H preserves the image details and can provide auxiliary information for classification. We also conduct experiments that use cA generated by DWT, and I_R generated by directly resizing the image to a given size, for classification. Specifically, we first resize cA and I_R to 128×128, and randomly crop the patches of size 112×112 (and their horizontal reflections) for training. During testing, we crop the center patch with a size of 112×112 for fair comparisons. Although I_L, I_R and cA are all assumed to possess the main content of the original image, the classification result using I_L is obviously superior to those using I_R and cA. One possible reason is that the constraint on minimizing the energy of I_H explicitly pushes most content to I_L so that I_L contains much more discriminative information than I_R and cA. These comparisons can also give a possible explanation that our approach outperforms the "Wavelet+CNN".

Contribution of joint fine tuning step We evaluated the contribution of joint fine tuning by comparing the performance with and without it, as reported in Table 4. The top-5 error rates with fine tuning decreases by 0.34%. This suggests fine tuning the network jointly can adapt the decom-

posed image for better classification.

Methods	w/o FT	w/ FT
top-5 err. (%)	12.21	11.87

Table 4: Comparison of the top-5 error rate with and without joint fine tuning (FT) on the ImageNet dataset.

ImageNet classification with ResNet-50

In this part, we further evaluate the performance of our proposed method on ResNet. Without loss of generalization, we select ResNet-50 from the ResNet family and simply use it to replace the VGG-Net as the baseline network. Then it is trained from scratch using a similar process as described in the Sec. of Learning. Because ResNet is a recently proposed network architecture, few works are proposed to accelerate this network. Thus, we simply compared with the standard ResNet-50, ThiNet in Table 5. ResNet is a more compact model, and accelerating this network is even more difficult. However, our method can still achieve $1.88\times$ speed-up with merely 0.8% increase in top-5 error rate, surpassing ThiNet on both accuracy and efficiency.

Methods	top-5 err. (%)	GPU SR	CPU SR
ResNet-50	8.86	$1\times$	$1\times$
ThiNet-30	11.70	$1.30\times$	-
Ours	9.66	$1.73\times$	$1.88\times$

Table 5: Comparison of the top-5 error rate and speed-up rate (SR) of our model and ThiNet on ResNet-50 on the ImageNet dataset.

CACD face identification

CACD is a large-scale and challenging dataset for face identification. It contains 163,446 images of 2,000 identities collected from the Internet that vary in age, pose and illumination. A subset of 56,138 images that cover 500 identities are manually annotated (**?**). We randomly select 44,798 images as the training set and the rest as the test set. All the models are trained on the training set and evaluated on the test set. Table 6 presents the comparison results. Note that the execution times are the same as Table 2. In this dataset, our model outperforms the VGG16-Net (0.22% increase in accuracy) and meanwhile achieves a speed-up rate of $3.13\times$. Besides, our method also beats the baseline methods. These comparisons again demonstrate the superiority of our proposed WAE. Remarkably, the images on CACD are far different from those on ImageNet, and our method still achieves superior performance on both accuracy and efficiency. It suggests our model can generalize to diverse datasets for accelerating the deep CNNs.

Noisy image classification

Generally, the high-frequency part of an image contains more noise. Our model may implicitly remove some high-frequency part by minimize the energy of I_H, so it may be inherently more robust to the noise. To validate this assumption, we add Gaussian noise of mean zero and different variances V to the test images, and present the accuracy of our method and the original VGG16-Net on these noisy images in Table 7. Note that both our model and the VGG16-Net is trained with the clean images. Our model performs consistently better than VGG16-Net over different noise levels. Remarkably, the superiority of our model is more evident when adding larger noise. For example, when adding noise with a variance of 0.05, our model outperforms the VGG16-Net by 10.81% in accuracy. These comparisons suggest our method is more robust to noise compared to VGG16-Net.

Methods	acc. (%)
VGG16-Net	95.91
Wavelet+CNN	94.99
Decomposition+CNN	95.20
Ours	96.13

Table 6: Comparison of the accuracy of our model, VGG16-Net and the baseline methods on the CACD dataset.

Methods	VGG16-Net	Ours
V=0	95.91	96.13
V=0.01	90.22	91.16
V=0.02	80.00	83.85
V=0.05	45.10	55.91
V=0.1	14.31	23.88

Table 7: Comparison of accuracy (in %) on the image of our model and VGG16-Net with gaussian noise of zero mean and different variances on the CACD dataset.

Conclusion

In this paper, we learn a Wavelet-like Auto-Encoder, which decomposes an input image into two low-resolution channels and utilizes the decomposed channels as inputs to the CNN to reduce the computational complexity without compromising the accuracy. Specifically, the WAE consists of an encoding layer to decompose the input image into two half-resolution channels and a decoding layer to synthesize the original image from the two decomposed channels. A transform loss, which combines a reconstruction error that constrains the two low-resolution channels to preserve all the information of the input image, and an energy minimization loss that constrain one channel contains minimum energy, are further proposed to optimize the network. In future work, we will conduct experiments to decompose the image into sub-images of lower resolution to explore a better trade-off between accuracy and speed.

Autem quis tempora, porro eligendi quis alias quisquam nisi quod sit provident, aperiam pariatur reiciendis perspiciatis voluptatibus natus, soluta accusamus qui delectus aliquam similique quibusdam velit beatae cupiditate nisi maxime?Dolore tempore consequuntur beatae, voluptates asperiores labore veritatis corporis aut dolores, incidunt provident amet enim inventore magnam?Hic vel at delectus nostrum a iste labore sequi, rem cumque blanditiis commodi ipsum, enim perspiciatis ipsa molestias beatae nisi laborum rem veniam nostrum doloribus