Meta-Reinforced Synthetic Data for One-Shot Fine-Grained Visual Recognition

Satoshi Tsutsui (Indiana University), Yanwei Fu (Fudan University), David Crandall (Indiana University)

Introduction

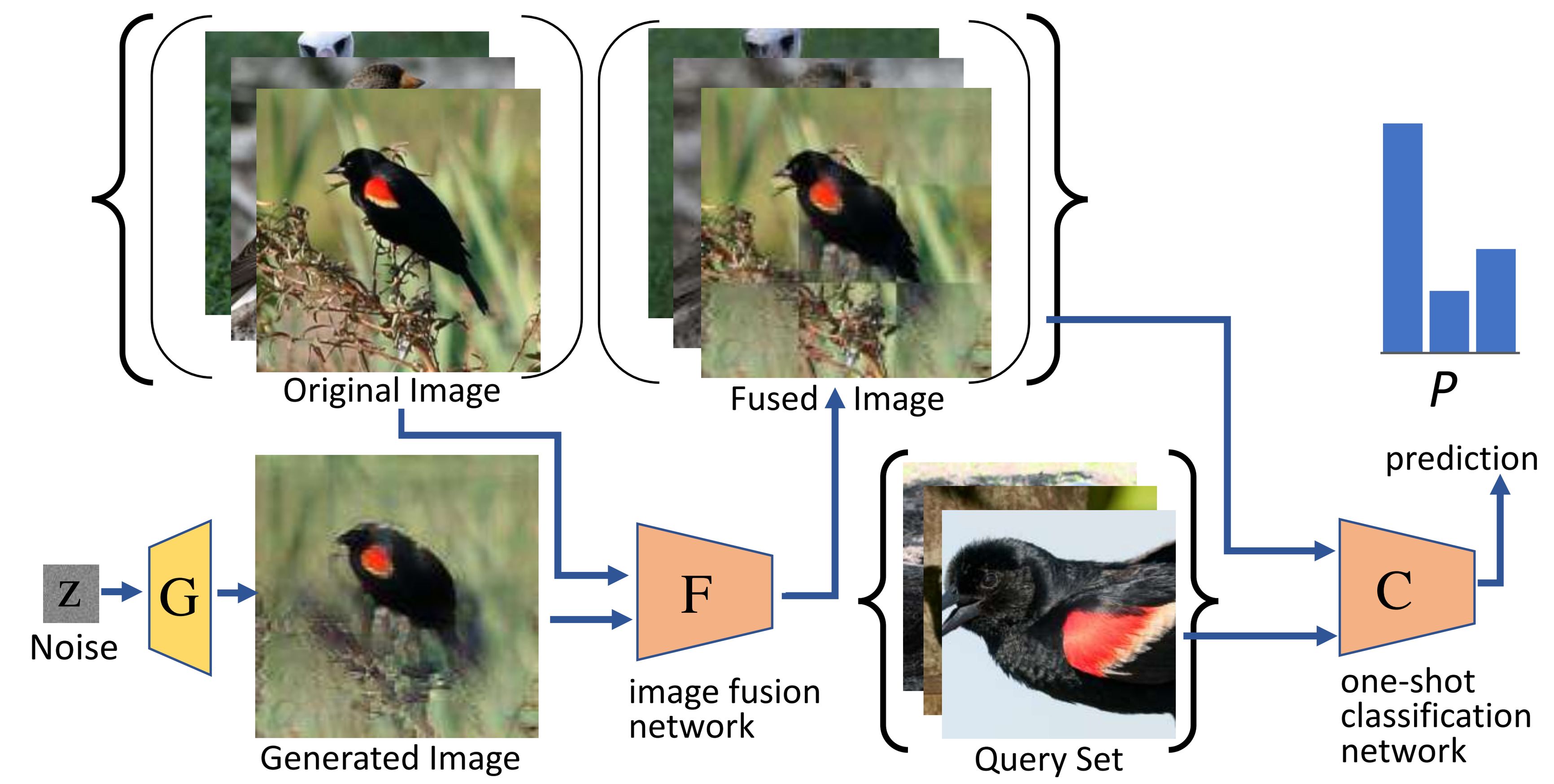
Quiz: Hawk or Falcon?

Fine-grained visual recognition
- Harder than normal classification.
- Difficult to collect data.
- Need **one-shot learning**.

Our Ideas
- Want to use Generative Adversarial Networks (GANs).
  - Challenge: GAN training itself needs a lot of data.
- Fine-tune GANs trained on ImageNet.
  - Challenge: Generated images decreased accuracy.
- Learn to reinforce generated images with original images.
- Use meta-learning to learn best mixing strategy.

Answer to the quiz: Hawk is left, and Falcon is right.

Framework



Experiments

- Code: http://vision.soic.indiana.edu/metairnet/
- Base one-shot classifier is Prototypical Networks (ProtoNet [3]), backbone is ImageNet-pretrained Resnet18 or Conv4.
- Datasets:
  - Caltech UCSD Birds (CUB).
    - train:val:test = 5,885 (100 classes):2,950 (50 classes):2,953 (50 classes)
  - North American Birds (NAB).
    - train:val:test = 24,557 (278 classes):11,960 (138 classes):12,010 (139 classes)

Results:

Table 2: 5-way-1-shot accuracy (%) on CUB/NAB dataset with ImageNet pre-trained ResNet18

Method	Data Augmentation	CUB Acc.	NAB Acc.
Nearest Neighbor	-	79.00 ± 0.62	80.58 ± 0.59
Logistic Regression	-	81.17 ± 0.60	82.70 ± 0.57
Softmax Regression	-	80.77 ± 0.60	82.38 ± 0.57
ProtoNet	-	81.73 ± 0.63	87.91 ± 0.52
ProtoNet	FinetuneGAN	79.40 ± 0.69	85.40 ± 0.59
ProtoNet	Flip	82.66 ± 0.61	88.55 ± 0.50
ProtoNet	Gaussian	81.75 ± 0.63	87.90 ± 0.52
MetaIRNet (Ours)	FinetuneGAN	84.13 ± 0.58	89.19 ± 0.51
MetaIRNet (Ours)	FinetuneGAN, Flip	$\mathbf{84.80 \pm 0.56}$	$\mathbf{89.57 \pm 0.49}$

Table 3: 5-way-1-shot accuracy (%) on CUB dataset with Conv4 without ImageNet pre-training

MetaIRNet	ProtoNet [28]	MatchingNet [31]	MAML [10]	RelationNet [29]
$\mathbf{65.86 \pm 0.72}$	63.50 ± 0.70	61.16 ± 0.89 [4]	55.92 ± 0.95 [4]	62.45 ± 0.98 [4]

Key Idea 1: Fine-tune BigGAN generator with a single image

- Transfer generative knowledge from one million generic images in ImageNet to a domain specific image [2].
- Instead of unstable adversarial training, we minimize both the noise and the difference between the input and output.

$$\mathcal{L}_1\left(G(z), \mathbf{I}_z\right) + \lambda_p \mathcal{L}_{perc}\left(G(z), \mathbf{I}_z\right) + \lambda_z \mathcal{L}_{EM}\left(z, r\right), \tag{1}$$

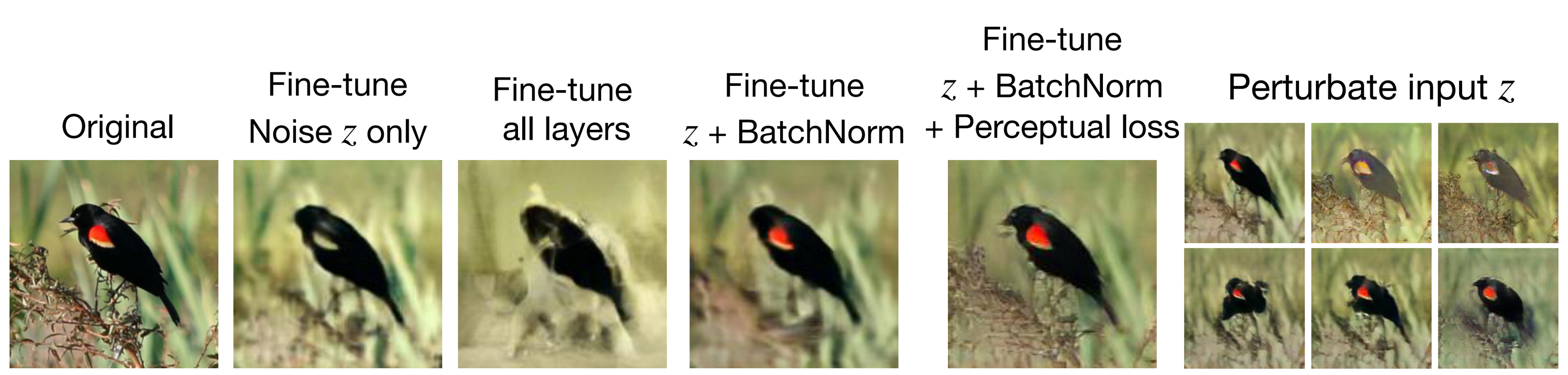
where z is a noise, G is the generator, $\mathbf{I}_z$ is an image, $G(z)$ is a generated image, $\mathcal{L}_1$ is L1 loss, $\mathcal{L}_{perc}$ is perceptual loss, $\mathcal{L}_{EM}$ is an earth mover distance between z and random noise $r \sim \mathcal{N}(0, 1)$ to regularize z to be sampled from a Gaussian, and λ_p and λ_z are coefficients of each term.

- To avoid overfitting, we update batch normalization layers only.

Specifically, only the γ and β of each batch normalization layer are updated in each layer,

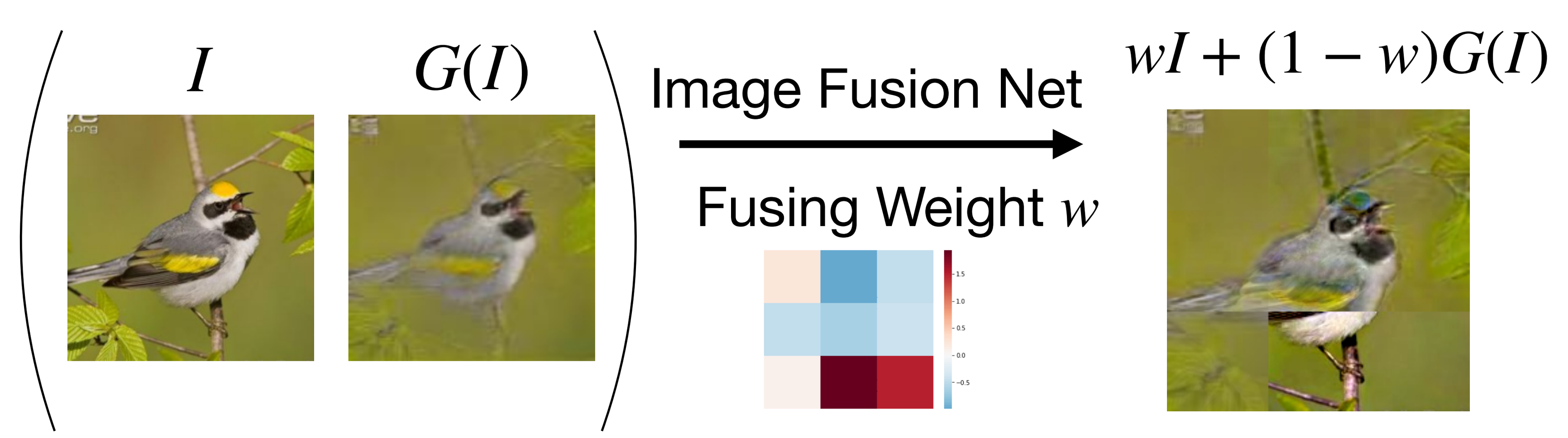
$$\hat{x} = \frac{x - \mathbb{E}(x)}{\sqrt{\mathrm{Var}(x) + \epsilon}} \qquad h = \gamma\hat{x} + \beta, \tag{2}$$

where x is the input feature from the previous layer, and $\mathbb{E}$ and Var indicate the mean and variance functions, respectively. Intuitively and in principle, updating γ and β only is equivalent to adjusting the activation of each neuron in a layer.

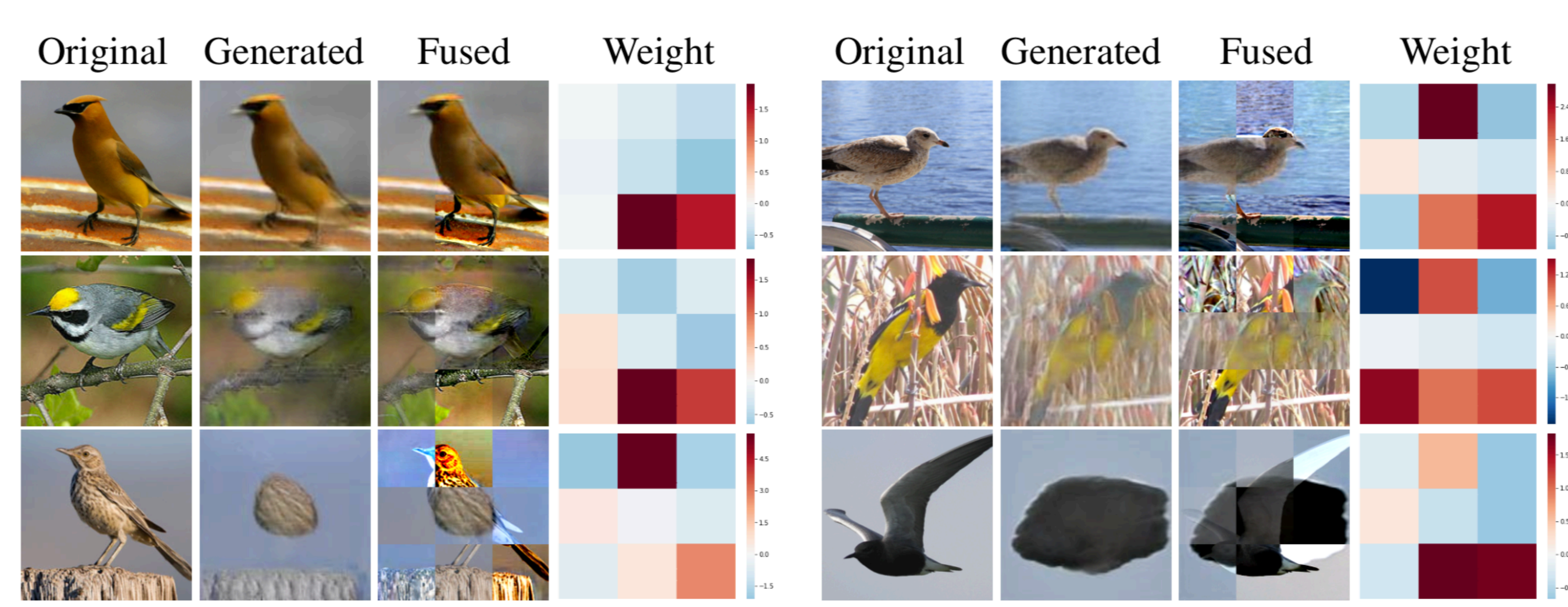


Key Idea 2: Reinforce generated image with the original.

- Linearly combine with a 3 x 3 block [1].
- Weights are learned by meta-learning.

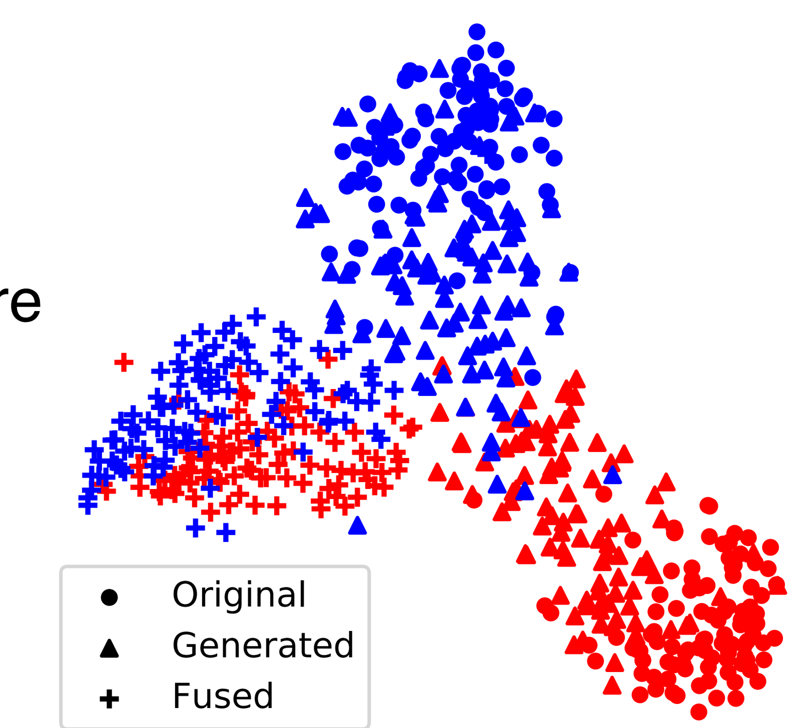


Some examples:



Visualization:

- Plot t-SNE of two classes, blue and red.
  - Generated images are closer to real ones.
  - Reinforced images are distinctive from others.



Conclusions:

- Composites of real and synthetic training images improve fine-grained one-shot recognition.
- Future work should explore other mixing strategies, and theoretical results on why it works.

[1] Zitian Chen, Yanwei Fu, Yu-Xiong Wang, Lin Ma, Wei Liu, and Martial Hebert. Image deformation meta-networks for one-shot learning. In CVPR 2019.
[2] Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *ICCV* 2019.
[3] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. In NIPS, 2017.

We thank Yi Li for helping draw figures and Minjun Li and Atsuhiro Noguchi for helpful discussions. Part of this work was done while Satoshi Tsutsui was an intern at Fudan University. Yanwei Fu was supported in part by the NSFC project (#61572138), and Science and Technology Commission of Shanghai Municipality Project (#19511120700). David Crandall was supported in part by the National Science Foundation (CAREER IIS-1253549), and the Indiana University Office of the Vice Provost for Research, the College of Arts and Sciences, and the Luddy School of Informatics, Computing, and Engineering through the Emerging Areas of Research Project "Learning: Brains, Machines, and Children."