Thanks for the reviewers' valuable comments. We appreciate the positive comments on well-motivated approach with promissing performance for fine-grained image recognition. Moreover, we can observe improvements on large scale *ImageNet* recognition task (as shown in the table for Reviewer #2). We address the concerns of reviewers as following.

To Reviewer #1:

Q 1.1 Compared with MA-CNN [9]. The proposed group bilinear requires the intra-group channels to be highly correlated (refer to the definition in Q 3.1), and the proposed semantic grouping can better satisfy such requirements than MA-CNN [9]. Specifically, [9] adopts the idea of k-means, which optimizes each channel to its cluster center. While the proposed grouping method in this paper optimize the correlation of intra-group and inter-group channels in a **pairwise manner** (as shown in Q 1.3 L_g), which has been proved to be able to obtain a tighter cluster (higher correlations), e.g., *mixture modelling by affinity propagation* Brendan Frey et al., nips 2006 and *clustering by passing messages between data points*, Brendan Frey et al., science 2007.
Moreover, we conducted experiments by replacing our grouping loss with [9], and the results also show the effectiveness of our proposed grouping module (i.e., one of the main contributions of this paper):

Grouping Mechanism	Grouping w/o constraints	Constraints in MA-CNN [9]	Constraints in DBTNet (ours)
Accuracy (%)	79.8	83.2	**85.1**

Q 1.2 The concrete loss function. Eq. (4) is exact the criterion to optimize the parameter $\mathbf{A}$, and it can be formulated as: $L_g = L_{intra} + L_{inter}$, where $L_{intra} = \sum_{\substack{0\le i,j<N \\ \lfloor i/G\rfloor = \lfloor j/G\rfloor}} -d_{ij}^2$ and $L_{inter} = \sum_{\substack{0\le i,j<N \\ \lfloor i/G\rfloor \neq \lfloor j/G\rfloor}} d_{ij}^2$ are designed to maximize/minimize the intra/inter-group correlations, respectively. Note that the notations above are the same with Eqn. (3), and the pairwise correlation is $d_{ij} = \frac{\tilde{\mathbf{m}}_i^T \tilde{\mathbf{m}}_j}{\|\tilde{\mathbf{m}}_i\|_2 \cdot \|\tilde{\mathbf{m}}_j\|_2}$. The overall loss L is shown as: $L = L_c + \lambda \sum_b^B L_g^{(b)}$, where L_c is softmax cross entropy loss for classification, $L_g^{(b)}$ is semantic grouping loss over the b^{th} block, B is the number of residual blocks, and λ is the weight of semantic grouping loss. We will add these equations in the method section.

Q 1.3 Constrains of the index mapping matrix. Thanks for your comments. $\mathbf{A}$ is an approximate index mapping matrix, whose rows are constrained to be (approximate) one-hot vectors via a $softmax$ with small "temperature". For example, a vector $\mathbf{x}$ can be approximately transformed into a one-hot vector by: $softmax(\mathbf{x}/T)$, where T is the temperature and is set to 0.0001 in our experiments. We will add this missing detail in the method section.

Q 1.4 Experiment settings for Table 7. As described in Page 6, Line 214 and Line 219, we conduct ablation studies with 224×224 input images for fast training and use 448×448 input images in Table 7 for fair comparison.

To Reviewer #2:

Q 2.1 Loss function. Thanks for your advice, and the concrete loss function can be found in Q 1.1 for Reviewer #1.

Q 2.2 Inconsistent notations. Thanks for your comments, and we will correct the notation "stage 3,4" into "Stage IV,V" respectively. "Last layer" indicates conducting group bilinear over the last layer of the backbone, which is added by default in Table 5. Thus "stage 3+4" in Table 5 is exact "last layer+Stage IV+Stage V" in Table 6.

Q 2.3 Results on ImageNet. The proposed models are pre-trained on ImageNet-1K. It can be observed that DBTNet-50 outperforms Resnet-50 and iSQRT-COV-8k with an obvious margin (1.6% and 0.5% absolute improvements respectively), and it achieves comparable results with iSQRT-COV-32k, whose feature dimension is 16 times larger:

Approach	ResNet-50 [17]	iSQRT-COV [13]	iSQRT-COV [13]	DBTNet-50 (ours)
Dimension	2k	8k	32k	2k
Top-1 err. (the lower, the better)	23.9	22.8	22.1	**22.3**

We use standard data augmentation methods provided by MXNet, i.e., random resized crop and random mirror.

Q 2.4 Missing references. Thanks for your advice, and we will add discussions for the missing references.

To Reviewer #7:

Q 3.1 Definition of semantic groups. A semantic group indicates a series of channels which represent the same semantic pattern, that is, the channels within a semantic group have responses in the same positions for a given image. Specifically, we obtain semantic groups by equally dividing 512 arranged channels (Eqn. (3)) into 16 groups and optimizing the responses of intra/inter-group channels to share larger/smaller spacial overlaps by Eqn. (4).

Q 3.2 Clarification for contributions. The proposed group bilinear makes deep bilinear transformation **doable** and the proposed semantic grouping ensures **competitive performance**. Designing suitable grouping methods plays a key role. As shown in the table for Reviewer #1, different grouping mechanisms achieve different results with large variances (79.8, 83.2, and 85.1). Specifically, the proposed semantic groups can enhance intra-group correlation, thus rich pairwise interactions can be obtained by the intra-group bilinear; inter-group correlation is suppressed, which makes the aggregation among groups free from information merging.
Moreover, such a design can also achieve promissing performance on ImageNet task (see the table for Reviewer #2).