Improving Style Transfer with Calibrated Metrics

Anonymous WACV submission

Paper ID 1034

1. Quick Overview

Notice that in Fig 5 all Gatys related methods except *Gatys with mean and covariance control* have quite low E compared to the E for cross-layer methods in Fig 6. But *Gatys with mean and covariance control* has different symmetries to Gatys (because one is controlling both mean and covariance, rather than just the Gram matrix; the symmetries are like those of the cross-layer method). This suggests it is likely that the symmetry is at least part of the reason why some methods outperform others.

There are two possible reasons. First, the symmetry results in poor solutions being easy to find. Second, the symmetry causes optimization problems. Both issues appear to be in play. Figures 5 and 6 together suggest that methods have considerable variance in performance, which is consistent with poor solutions being easy to find. But the good performance of GAL (see Fig. 4) suggests that optimization is an issue, too.

Symmetries can create problems for optimization methods, because symmetries must be associated with strong gradient curvature at least some points. GAL uses a standard optimization trick to simplify the optimization problem; the success of this trick suggests that optimization of Gatys' loss is hard.

1.1. GAL

Gatys' loss is a function of feature values at each layer. One usually assumes that the feature values taken at layer l are a known function of the feature values at layer $l-1$. Here the function is given by the appropriate convolutional layer, etc. However, we could "cut" the network between layers, then introduce a constraint requiring that variables on either side of the cut be equal. We solve this constrained problem using the augmented lagrangian method (see [4] for this strategy applied to MRFs).

Write $f^l_{k,p}$ for the response of the k'th channel at the p'th location in the l'th convolutional layer; drop subscripts as required, and write $f^l = \phi^l(f^{l-1}_{.,.})$ for the function mapping layer to layer. GAL cuts the layers only at R41. We have not tried other cuts. It would be interesting to see what happened with more cuts, but the optimization problem gets big quickly. We introduce dummy variables $V_{k,p}$, and the constraint $V = \phi^4(f^3_{.,.})$. Write λ for lagrange multipliers corresponding to the constraint, I for the image, and $\lambda^{(i)}$ for the i'th estimate of those lagrange multipliers, etc.

The augmented lagrangian is now

$$\begin{aligned}\mathcal{L}(I,V,\lambda) \quad &= \textstyle\sum_{l\neq 4} w_l L^l_{style}(I, I_{style}) \\ &+ w_4 L^4_{style}(V, I_{style}) \\ &+ L_{content}(V, I_{content}) \\ &+ L_{aug}(I, V, \lambda)\end{aligned}$$

where w_l is the style weight of each layer, L^l_{style} is the style loss for layer l, and $L_{content}$ is the content loss at R41, and

$$\begin{aligned}L_{aug}(I,V,\lambda) \quad &= \tfrac{1}{KP}\textstyle\sum_{k,p}\Big(\lambda_l * (V_l - \phi^4(f^3_{.,.}(I))) \\ &+ \rho(V_l - \phi^4(f^3_{.,.}(I)))^2\Big)\end{aligned}$$

In the primal step, we first optimize the lagrangian with respect to I, using fixed V, λ using LBFGS. We then fix I, and optimize with respect to V (notice this involves solving a relatively straightforward linear system). The dual step then re-estimates the lagrange multipliers as usual:

$$\lambda_4^{(i+1)} = \lambda_4^{(i)} + \rho^{(i)}(V_4^{(i)} - f^4(I_n^{(i)})).$$

Finally, we update ρ by $\rho^{(i+1)} = 1.4\rho^{(i)}$.

Figure 1 and Figure 2 display our 50 style images. Except the Universal style transfer, all other methods synthesize image from Gaussian noise with LBFGS optimizer. The content images and style images are resized to same width of 512 as the input for style transfers.

1.2. Cross-layer with control of mean and covariance (XLCM)

We observe that feature mean difference between I_s and I_c is directly related to the optimization performance of style transfer, e.g. when the content image have similar feature mean as style image the transfer image has better style quality. Therefore we introduce the L2 loss between

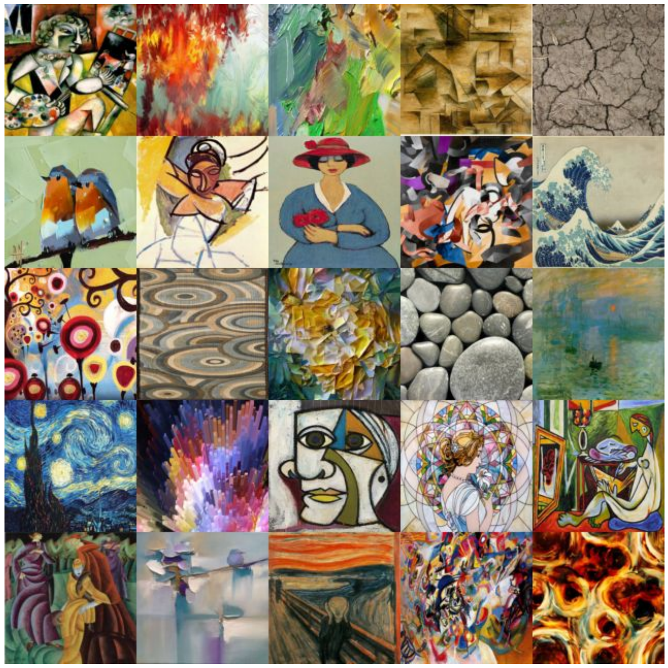

Figure 1: *The first group of 50 styles.*

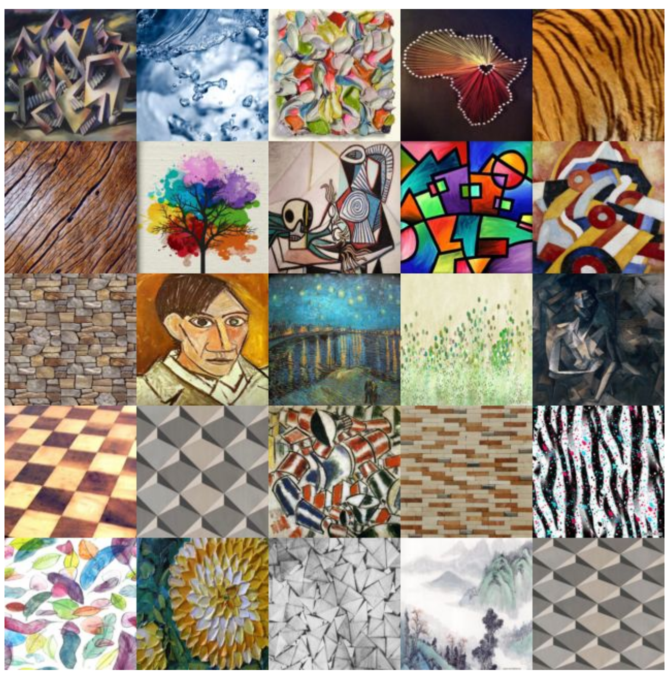

Figure 2: *The second group of 50 styles.*

each feature channel's mean of I_n and each feature channel's mean of I_s to enforce the transfer image has close feature mean to style image. Here is the loss for mean control.

$$L_{mean} = \sum_k \left(\sum_p \frac{f^l(I_n)}{P} - \sum_p \frac{f^l(I_s)}{P} \right)^2$$

On the other hand, the covariant control is to replace cross-layer gram matrix by corresponding cross-layer gram matrix with each feature subtracted by by its mean. Here is the new cross-layer loss with covariant control.

$$Cov_{ij}^{l,m}(I) = \sum_p \left[f^l_{i,p}(I) - \bar{f}^l_{i,p}(I) \right] \left[\uparrow f^m_{j,p}(I) - \uparrow \bar{f}^m_{j,p}(I) \right]^T .$$

Here $\bar{f}^l_{i,p}(I)$ is the tensor duplicated in p dimension with the mean of $f^l_{i,p}(I)$ over p.

2. Quantization of transferred images under user study regression models

Recall in Section 4 of original text we regress base E and C statistic to user preference. We obtain one best E-model from E-test user preference, and one best C-model from that of C-test. These two models assign E and C scores for each transferred image (Sec. 4.1 of original text). Thus, we gather a scatter plot of all transferred images, and we quantize this scatter plot into a 3-by-3 grid, each cell has roughly same number of images. From this grid we generate a visualization of EC space (Fig.1 in original text).

This quantization shows similar trends with Figure 4-6 in the original text. Table 1 shows the Top 5 methods ranking for all quantiles. In quantile of high C-score, high E-score, GAL is the top method. XM dominates both (middle C, middle E) and (high C, middle E), and Universal dominates both (middle C, low E) and (high C, low E). Other high E quantiles are dominated by cross-layer related methods. The worst quantile(low C-score,Low E-score) has Gatys aggressive as the most popular.

This difference in symmetry groups is important. Risser argues that the symmetries of gram matrices in Gatys' method could lead to unstable reconstructions; they control this effect using feature histograms. What causes the effect is that the symmetry rescales features while shifting the mean. For the cross-layer loss, the symmetry cannot rescale, and cannot shift the mean. In turn, the instability identified in that paper does not apply to the cross-layer gram matrix and our results could not be improved by adopting a histogram loss.

Write $\mathbf{x}_i$, (resp $\mathbf{y}_i$ for the feature vector at the i'th location (of N in total) in the first (resp second) layer. Write $\mathcal{X}^T = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, etc.

Symmetries of the first layer: Now assume that the first layer has been normalized to zero mean and unit covariance. There is no loss of generality, because the whitening transform can be written into the expression for the group. Write $\mathcal{G}(\mathcal{W}) = (1/N)\mathcal{W}^T\mathcal{W}$ for the operator that forms the within layer gram matrix. We have $\mathcal{G}(\mathcal{X}) = \mathcal{I}$. Now consider an affine action on layer 1, mapping $\mathcal{X}_1$ to $\mathcal{X}_1^* = \mathcal{X}_1\mathcal{A} + \mathbf{1}\mathbf{b}^T$; then for this to be a symmetry, we must have $G(\mathcal{X}_1^*) = \mathcal{A}\mathcal{A}^T + \mathbf{b}\mathbf{b}^T = \mathcal{I}$. In turn, the symmetry group can be constructed by: choose $\mathbf{b}$ which does not have unit length; factor $N(\mathcal{I} - \mathbf{b}\mathbf{b}^T)$ to obtain $\mathcal{A}(\mathbf{b})$ (for example, by using a cholesky transformation); then any element of the group is a pair $\big(\mathbf{b}, \mathcal{A}(\mathbf{b})\mathcal{U}\big)$ where $\mathcal{U}$ is orthonormal. Note that factoring will fail for $\mathbf{b}$ a unit vector, whence the restriction.

The second layer: We will assume that the map between layers of features is linear. This assumption is not true in practice, but major differences between symmetries observed under these conditions likely result in differences when the map is linear. We can analyze for two cases: first, all units in the map observe only one input feature vector (i.e. 1x1 convolutions; the *point sample* case); second, spatial homogeneity in the layers.

The point sample case: Assume that every unit in the map observes only one input feature from the previous layer (1x1 convolutions). We have $\mathcal{Y} = \mathcal{X}\mathcal{M} + \mathbf{1}\mathbf{n}^T$, because the map between layers is linear. Now consider the effect on the second layer. We have $\mathcal{G}(\mathcal{Y}) = \mathcal{M}\mathcal{M}^T + \mathbf{n}\mathbf{n}^T$. Choose some symmetry group element for the first layer, $(\mathbf{b}, \mathcal{A})$. The gram matrix for the second layer becomes $\mathcal{G}(\mathcal{Y}^*)$, where $\mathcal{Y}^* = (\mathcal{X}\mathcal{A} + \mathbf{1}\mathbf{b}^T)\mathcal{M}^T + \mathbf{1}\mathbf{n}^T$. Recalling that $\mathcal{A}\mathcal{A}^T + \mathbf{b}\mathbf{b}^T = \mathcal{I}$ and $\mathcal{X}^T\mathbf{1} = 0$, we have

$$\mathcal{G}(\mathcal{Y}^*) = \mathcal{M}\mathcal{M}^T + \mathbf{n}\mathbf{n}^T + \mathbf{n}\mathbf{b}^T\mathcal{M}^T + \mathcal{M}\mathbf{b}\mathbf{n}^T$$

so that $\mathcal{G}(\mathcal{X}_2^*) = \mathcal{G}(\mathcal{X}_2)$ if $\mathcal{M}\mathbf{b} = 0$. This is relatively easy to achieve with $\mathbf{b} \neq 0$.

Spatial homogeneity: Now assume the map between layers has convolutions with maximum support $r \times r$. Write u for an index that runs over the whole feature map, and $\psi(\mathbf{x}_u)$ for a stacking operator that scans the convolutional support in fixed order and stacks the resulting features. For example, given a 3x3 convolution and indexing in 2D, we might have

$$\psi(\mathbf{x}_{22}) = \begin{pmatrix} \mathbf{x}_{11} \\ \mathbf{x}_{12} \\ \ldots \\ \mathbf{x}_{33} \end{pmatrix}$$

In this case, there is some $\mathcal{M}$, $\mathbf{n}$ so that $\mathbf{y}_u = \mathcal{M}\psi(\mathbf{x}_u) + \mathbf{n}$. We ignore the effects of edges to simplify notation (though this argument may go through if edges are taken into account). Then there is some $\mathcal{M}$, $\mathbf{n}$ so we can write

$$\mathcal{G}(\mathcal{Y}) = (1/N)\sum_u \mathcal{M}\psi(\mathbf{x}_u)\psi(\mathbf{x}_u)^T\mathcal{M}^T + \mathbf{n}\mathbf{n}^T$$

Now assume further that layer 1 has the following (quite restrictive) spatial homogeneity property: for pairs of feature vectors within the layer $\mathbf{x}_{i,j}$, $\mathbf{x}_{i+\delta,j+\delta}$ with $\mid \delta \mid \leq r$ (ie within a convolution window of one another), we have $\mathbb{E}\big[\mathbf{x}_{i,j}\mathbf{x}_{i+\delta,j+\delta}\big] = \mathcal{I}$. This assumption is consistent with image autocorrelation functions (which fall off fairly slowly), but is still strong. Write ϕ for an operator that stacks $r \times r$ copies of its argument as appropriate, so

$$\phi(\mathcal{I}) = \begin{pmatrix} \mathcal{I} & \ldots & \mathcal{I} \\ \ldots & \ldots\ldots & \\ \mathcal{I} & \ldots & \mathcal{I} \end{pmatrix}.$$

Then $G(\mathcal{Y}) = \mathcal{M}\phi(\mathcal{I})\mathcal{M}^T + \mathbf{n}\mathbf{n}^T$. If there is some affine action on layer 1, we have $G(\mathcal{Y}^*) =$

(low C-score, high E-score) **Cross-layer,aggressive:24.06%**, XLCM:20.92%, XLC:11.92%, XL:11.30%, GatysCM:9.21%	(middle C-score, high E-score) **XLC:14.56%**, Cross-layer,aggressive:13.60%, XLCM:13.41%, XL:13.22%, GAL:10.15%	(high C-score, high E-score) **GAL:25.56%**, XM:15.04%, XL:10.53%, GatysL:8.52%, GatysCM:6.77%
(low C-score, middle E-score) **GatysCM:15.29%**, GatysC:12.86%, Cross-layer, aggressive:11.65%, GatysL:11.65%, XLCM:8.50%	(middle C-score, middle E-score) **XM:11.69%**, GatysM:11.49%, GatysL:10.69%, GatysH:10.08%, GatysC:8.87%	(high C-score, middle E-score) **XM:15.45%,** GatysH:14.02%, Gatys:13.41%, GAL:13.01%, GatysM:11.18%
(low C-score, low E-score) **Gatys aggressive:23.97%**, GatysC:12.57%, XLC:10.02%, GatysCM:8.84%, GatysM:7.47%	(middle C-score, low E-score) **Universal:12.83%**, GatysH:10.73%, Gatys aggressive:10.47%, GatysM:10.21%, Gatys:9.69%	(high C-score, low E-score) **Universal:45.28%**, Gatys:15.75%, GatysH:7.87%, GatysM:6.69%, GatysL:4.53%

GatysH – Gatys, with histogram loss
GatysL – Gatys, with layerwise style weights
GatysM – Gatys, with mean control
GatysC – Gatys, with covariance control
GatysCM – Gatys, with mean and covariance control
XL – Cross-layer
XM – Cross-layer, multiplicative
XLC – Cross-layer, with control of covariance
XLCM – Cross-layer, with control of mean and covariance
GAL – Gatys, augmented Lagrangian method
Universal – Universal Style Transfer

Table 1: Top 5 methods ranking for each quantile under regression scores coordinate generated by selected E-model and C-model. Each transferred image has five E-statistic and one C-statistic, they are used to regress user preference in E-test and C-test (Sec. 4.1 in original text). Selected E and C models regress scores (higher is better) for each transferred image. We divide the scatter into 3-by-3 quantiles, and show method distribution for each quantile.

$\mathcal{M}\left(\psi(\mathcal{A})\phi(\mathcal{I})\psi(\mathcal{A}^T)+\psi(\mathbf{b})\psi(\mathbf{b}^T)\right)\mathcal{M}^T+\mathbf{n}\mathbf{n}^T$, where we have overloaded ψ in the natural way. Now if $\mathcal{M}\psi(\mathbf{b}) = 0$ and $\mathcal{A}\mathcal{A}^T+\mathbf{b}\mathbf{b}^T=\mathcal{I}$, $\mathcal{G}(\mathcal{Y}^*)=\mathcal{G}(\mathcal{Y})$.

The cross-layer gram matrix: Symmetries of the cross-layer gram matrix are very different. Write $\mathcal{G}(\mathcal{X},\mathcal{Y}) = (1/N)\mathcal{X}^T\mathcal{Y}$ for the cross layer gram matrix.

Cross-layer, point sample case: Here (recalling $\mathcal{X}^T\mathbf{1}=0$)we have $\mathcal{G}(\mathcal{X},\mathcal{Y})=\mathcal{M}^T$. Now choose some symmetry group element for the first layer, $(\mathcal{A},\mathbf{b})$. The cross-layer gram matrix becomes

$$\begin{aligned}\mathcal{G}(\mathcal{X}^*,\mathcal{Y}^*) &= (1/N)(\mathcal{A}\mathcal{X}^T+\mathbf{b}\mathbf{1}^T)\left[(\mathcal{X}\mathcal{A}^T+\mathbf{1}\mathbf{b}^T)\mathcal{M}^T+\mathbf{1}\mathbf{n}^T\right]\\ &= \mathcal{M}^T+\mathbf{b}\mathbf{n}^T\end{aligned}$$

(recalling that $\mathcal{A}\mathcal{A}^T+\mathbf{b}\mathbf{b}^T=\mathcal{I}$ and $\mathcal{X}^T\mathbf{1}=0$). But this means that the symmetry requires $\mathbf{b}=\mathbf{0}$; in turn, we must have $\mathcal{A}\mathcal{A}^T=\mathcal{I}$.

Cross-layer, homogeneous case: We have

$$\mathcal{G}(\mathcal{X},\mathcal{Y}) = (1/N)\sum_u \mathbf{x}_u\left[\psi(\mathbf{x}_u)^T\mathcal{M}^T+\mathbf{n}^T\right] = \mathcal{M}^T.$$

Now choose some symmetry group element for the first layer, $(\mathcal{A},\mathbf{b})$. The cross-layer gram matrix becomes

$$\begin{aligned}\mathcal{G}(\mathcal{X}^*,\mathcal{Y}^*) &= (1/N)\sum_u\Big\{(\mathcal{A}\mathbf{x}_u+\mathbf{b})\\ &+ \left[\left(\psi(\mathbf{x}_u)^T\psi(\mathcal{A}^T)+\psi(\mathbf{b})\right)\mathcal{M}^T+\mathbf{n}^T\right]\Big\}\\ &= \mathcal{M}^T+\mathbf{b}\mathbf{n}^T\end{aligned}$$

(recalling the spatial homogeneity assumption, that $\mathcal{A}\mathcal{A}^T + \mathbf{b}\mathbf{b}^T = \mathcal{I}$ and $\mathcal{X}_1^T\mathbf{1} = 0$). But this means that the symmetry requires $\mathbf{b} = \mathbf{0}$; in turn, we must have $\mathcal{A}\mathcal{A}^T = \mathcal{I}$.

3. Construction of Affine Maps for Symmetry Groups

This difference in symmetry groups is important. Risser argues that the symmetries of gram matrices in Gatys' method could lead to unstable reconstructions; they control this effect using feature histograms. What causes the effect is that the symmetry rescales features while shifting the mean. For the cross-layer loss, the symmetry cannot rescale, and cannot shift the mean. In turn, the instability identified in that paper does not apply to the cross-layer gram matrix and our results could not be improved by adopting a histogram loss.

Write $\mathbf{x}_i$, (resp $\mathbf{y}_i$ for the feature vector at the i'th location (of N in total) in the first (resp second) layer. Write $\mathcal{X}^T = [\mathbf{x}_1, \dots, \mathbf{x}_N]$, etc.

Symmetries of the first layer: Now assume that the first layer has been normalized to zero mean and unit covariance. There is no loss of generality, because the whitening transform can be written into the expression for the group. Write $\mathcal{G}(\mathcal{W}) = (1/N)\mathcal{W}^T\mathcal{W}$ for the operator that forms the within layer gram matrix. We have $\mathcal{G}(\mathcal{X}) = \mathcal{I}$. Now consider an affine action on layer 1, mapping $\mathcal{X}_1$ to $\mathcal{X}_1^* = \mathcal{X}_1\mathcal{A} + \mathbf{1}\mathbf{b}^T$; then for this to be a symmetry, we must have $G(\mathcal{X}_1^*) = \mathcal{A}\mathcal{A}^T + \mathbf{b}\mathbf{b}^T = \mathcal{I}$. In turn, the symmetry group can be constructed by: choose $\mathbf{b}$ which does not have unit length; factor $N(\mathcal{I} - \mathbf{b}\mathbf{b}^T)$ to obtain $\mathcal{A}(\mathbf{b})$ (for example, by using a cholesky transformation); then any element of the group is a pair $\big(\mathbf{b}, \mathcal{A}(\mathbf{b})\mathcal{U}\big)$ where $\mathcal{U}$ is orthonormal. Note that factoring will fail for $\mathbf{b}$ a unit vector, whence the restriction.

The second layer: We will assume that the map between layers of features is linear. This assumption is not true in practice, but major differences between symmetries observed under these conditions likely result in differences when the map is linear. We can analyze for two cases: first, all units in the map observe only one input feature vector (i.e. 1x1 convolutions; the *point sample* case); second, spatial homogeneity in the layers.

The point sample case: Assume that every unit in the map observes only one input feature from the previous layer (1x1 convolutions). We have $\mathcal{Y} = \mathcal{X}\mathcal{M} + \mathbf{1}\mathbf{n}^T$, because the map between layers is linear. Now consider the effect on the second layer. We have $\mathcal{G}(\mathcal{Y}) = \mathcal{M}\mathcal{M}^T + \mathbf{n}\mathbf{n}^T$. Choose some symmetry group element for the first layer, $(\mathbf{b}, \mathcal{A})$. The gram matrix for the second layer becomes $\mathcal{G}(\mathcal{Y}^*)$, where $\mathcal{Y}^* = (\mathcal{X}\mathcal{A} + \mathbf{1}\mathbf{b}^T)\mathcal{M}^T + \mathbf{1}\mathbf{n}^T$. Recalling that $\mathcal{A}\mathcal{A}^T + \mathbf{b}\mathbf{b}^T = \mathcal{I}$ and $\mathcal{X}^T\mathbf{1} = 0$, we have

$$\mathcal{G}(\mathcal{Y}^*) = \mathcal{M}\mathcal{M}^T + \mathbf{n}\mathbf{n}^T + \mathbf{n}\mathbf{b}^T\mathcal{M}^T + \mathcal{M}\mathbf{b}\mathbf{n}^T$$

so that $\mathcal{G}(\mathcal{X}_2^*) = \mathcal{G}(\mathcal{X}_2)$ if $\mathcal{M}\mathbf{b} = 0$. This is relatively easy to achieve with $\mathbf{b} \neq 0$.

Spatial homogeneity: Now assume the map between layers has convolutions with maximum support $r \times r$. Write u for an index that runs over the whole feature map, and $\psi(\mathbf{x}_u)$ for a stacking operator that scans the convolutional support in fixed order and stacks the resulting features. For example, given a 3x3 convolution and indexing in 2D, we might have

$$\psi(\mathbf{x}_{22}) = \begin{pmatrix} \mathbf{x}_{11} \\ \mathbf{x}_{12} \\ \dots \\ \mathbf{x}_{33} \end{pmatrix}$$

In this case, there is some $\mathcal{M}$, $\mathbf{n}$ so that $\mathbf{y}_u = \mathcal{M}\psi(\mathbf{x}_u) + \mathbf{n}$. We ignore the effects of edges to simplify notation (though this argument may go through if edges are taken into account). Then there is some $\mathcal{M}$, $\mathbf{n}$ so we can write

$$\mathcal{G}(\mathcal{Y}) = (1/N)\sum_u \mathcal{M}\psi(\mathbf{x}_u)\psi(\mathbf{x}_u)^T\mathcal{M}^T + \mathbf{n}\mathbf{n}^T$$

Now assume further that layer 1 has the following (quite restrictive) spatial homogeneity property: for pairs of feature vectors within the layer $\mathbf{x}_{i,j}$, $\mathbf{x}_{i+\delta,j+\delta}$ with $\mid \delta \mid \leq r$ (ie within a convolution window of one another), we have $\mathbb{E}\big[\mathbf{x}_{i,j}\mathbf{x}_{i+\delta,j+\delta}\big] = \mathcal{I}$. This assumption is consistent with image autocorrelation functions (which fall off fairly slowly), but is still strong. Write ϕ for an operator that stacks $r \times r$ copies of its argument as appropriate, so

$$\phi(\mathcal{I}) = \begin{pmatrix} \mathcal{I} & \dots & \mathcal{I} \\ \dots & \dots\dots & \\ \mathcal{I} & \dots & \mathcal{I} \end{pmatrix}.$$

Then $G(\mathcal{Y}) = \mathcal{M}\phi(\mathcal{I})\mathcal{M}^T + \mathbf{n}\mathbf{n}^T$. If there is some affine action on layer 1, we have $G(\mathcal{Y}^*) = \mathcal{M}\left(\psi(\mathcal{A})\phi(\mathcal{I})\psi(\mathcal{A}^T) + \psi(\mathbf{b})\psi(\mathbf{b}^T)\right)\mathcal{M}^T + \mathbf{n}\mathbf{n}^T$, where we have overloaded ψ in the natural way. Now if $\mathcal{M}\psi(\mathbf{b}) = 0$ and $\mathcal{A}\mathcal{A}^T + \mathbf{b}\mathbf{b}^T = \mathcal{I}$, $\mathcal{G}(\mathcal{Y}^*) = \mathcal{G}(\mathcal{Y})$.

The cross-layer gram matrix: Symmetries of the cross-layer gram matrix are very different. Write $\mathcal{G}(\mathcal{X}, \mathcal{Y}) = (1/N)\mathcal{X}^T\mathcal{Y}$ for the cross layer gram matrix.

Cross-layer, point sample case: Here (recalling $\mathcal{X}^T\mathbf{1} = 0$)we have $\mathcal{G}(\mathcal{X}, \mathcal{Y}) = \mathcal{M}^T$. Now choose some symmetry group element for the first layer, $(\mathcal{A}, \mathbf{b})$. The cross-layer gram matrix becomes

$$\begin{aligned} \mathcal{G}(\mathcal{X}^*, \mathcal{Y}^*) &= (1/N)(\mathcal{A}\mathcal{X}^T + \mathbf{b}\mathbf{1}^T)\left[(\mathcal{X}\mathcal{A}^T + \mathbf{1}\mathbf{b}^T)\mathcal{M}^T + \mathbf{1}\mathbf{n}^T\right] \\ &= \mathcal{M}^T + \mathbf{b}\mathbf{n}^T \end{aligned}$$

(recalling that $\mathcal{A}\mathcal{A}^T + \mathbf{b}\mathbf{b}^T = \mathcal{I}$ and $\mathcal{X}^T\mathbf{1} = 0$). But this means that the symmetry requires $\mathbf{b} = \mathbf{0}$; in turn, we must have $\mathcal{A}\mathcal{A}^T = \mathcal{I}$.

Cross-layer, homogeneous case: We have

$$\mathcal{G}(\mathcal{X},\mathcal{Y}) = (1/N)\sum_u \mathbf{x}_u \left[\psi(\mathbf{x}_u)^T\mathcal{M}^T + \mathbf{n}^T\right] = \mathcal{M}^T.$$

Now choose some symmetry group element for the first layer, $(\mathcal{A},\mathbf{b})$. The cross-layer gram matrix becomes

$$\begin{aligned}\mathcal{G}(\mathcal{X}^*,\mathcal{Y}^*) &= (1/N)\sum_u \Big\{ (\mathcal{A}\mathbf{x}_u + \mathbf{b}) \\ &+ \left[\left(\psi(\mathbf{x}_u)^T\psi(\mathcal{A}^T) + \psi(\mathbf{b})\right)\mathcal{M}^T + \mathbf{n}^T\right]\Big\} \\ &= \mathcal{M}^T + \mathbf{b}\mathbf{n}^T\end{aligned}$$

(recalling the spatial homogeneity assumption, that $\mathcal{A}\mathcal{A}^T + \mathbf{b}\mathbf{b}^T = \mathcal{I}$ and $\mathcal{X}_1^T\mathbf{1} = 0$). But this means that the symmetry requires $\mathbf{b} = \mathbf{0}$; in turn, we must have $\mathcal{A}\mathcal{A}^T = \mathcal{I}$.