基于级联多尺度信息融合对抗网络的红外仿真

贾瑞明[1*]，李彤[1]，刘圣杰[1]，崔家礼[1]，袁飞[2]

[1]北方工业大学信息学院，北京 100144；

[2]中国科学院自动化研究所数字内容技术与服务研究中心，北京 100190

摘要 提出了一种应用于红外图像仿真的级联多尺度信息融合生成对抗网络，能由可见光图像估计对应的红外图像。针对可见光与红外图像特征之间的关联与区别，该网络采用级联的对抗网络结构：第一级对抗网络以语义分割图像为辅助任务，使用大感受野的卷积网络结构，重建红外图像的结构信息；第二级对抗网络以可见光的灰度反转图像为辅助任务，采用小感受野的网络结构，补充红外仿真图像的细节纹理信息，并使用多尺度融合模块整合多感受野信息以提升算法精度。在先进算法的通用数据集上进行实验，结果表明，级联多尺度信息融合对抗网络能够实现可见光到红外图像的转换，可得到结构与纹理都较正确的红外仿真图像，在多种客观指标与主观感受上均优于其他类似算法。

关键词 图像处理；红外图像仿真；生成对抗网络；图像域转换；感受野

中图分类号 TP391 **文献标志码** A **doi**：10.3788/AOS202040.1810001

Infrared Simulation Based on Cascade Multi-Scale Information Fusion Adversarial Network

Jia Ruiming[1*], Li Tong[1], Liu Shengjie[1], Cui Jiali[1], Yuan Fei[2]

[1] *School of Information Science and Technology, North China University of Technology, Beijing* 100144, *China*;

[2] *Digital Content Technology and Media Service Research Center, Institute of Automation, Chinese Academy of Sciences, Beijing* 100190, *China*

Abstract In this paper, we propose a cascade multi-scale information fusion generative adversarial network (CMIF-GAN) for infrared image simulation, which can estimate the infrared map from a visible image. Inspired by the connections and differences between visible and infrared features, CMIF-GAN adopts a cascaded structure composed of two levels of adversarial networks. With a large overall receptive field, the first-level adversarial network focuses on reconstructing structural information of the infrared image, and adds a semantic segmentation image task as auxiliary information. To enrich detailed texture information of the infrared image, the second-level adversarial network uses the grayscale inverted visible (GIV) images as auxiliary information and adopts a small overall receptive field network. Otherwise, the second-level adversarial network can integrate the multiple receptive information by a multi-scale fusion module (MFM) to improve algorithm accuracy. Experiments on public dataset demonstrate that CMIF-GAN can efficiently translate visible images to corresponding infrared images, and outperform previous methods in objective metrics and subjective vision.

Key words image processing; infrared image simulation; generative adversarial network; image domain conversion; receptive field

OCIS codes 100.4994; 100.4995; 100.4996

1 引 言

红外图像应用广泛，通过红外摄像机实地采集是最直接最简单的红外图像获取方式。然而红外设备较昂贵，且受环境、人员等限制，这种获取红外图像的方式成本较高。为了提供更为便捷、成本更低

收稿日期：2020-04-07；**修回日期**：2020-05-07；**录用日期**：2020-06-03

基金项目：国家重点研发计划(2017YFB0802300)、国家自然科学基金青年基金(61602480)、北方工业大学学生科技活动项目(110051360019XN140)

* **E-mail**：jiaruiming@ncut.edu.cn

的红外图像，红外图像仿真的相关研究一直在进行中，目前红外图像的仿真获取可以分为两种方式。

第一种，基于三维建模的红外图像仿真。红外目标/场景仿真大多属于这种方式，其中红外目标仿真通常是对车辆、舰船等特定目标进行更精细的建模和计算；而红外场景仿真则是对场景中的多种景物如楼宇、植被等也进行建模和仿真，与前者区别在于建模的数量和精度不同。基于三维建模的红外仿真一般有三个步骤：1）对目标物进行三维模型构建；2）设置目标物的红外辐射特性，进行辐射的计算；3）模拟红外成像系统生成红外仿真图像。文献[1]采用这种算法实现了地面红外场景仿真，文献[2-3]则实现了船舰目标的红外图像仿真。这类方法的优点是不需要真实可见光或红外图像，通过软件直接建模产生场景或特定目标各种视角的红外图像；缺点是建模过程复杂，人力成本和时间成本消耗大，且因其只针对单一场景或单一目标，导致模型泛化性能差。针对以上缺点，文献[4]采用生成对抗网络，从目标类别标签和随机噪声中生成红外目标仿真图像，但其生成红外图像的可控性较低。

第二种，由可见光图像仿真生成对应红外图像。可见光图像资源丰富，且获取方式便捷、低成本。但红外成像系统和可见光成像系统都较复杂且受多变量影响，这种特性使得可见光图像和红外图像的映射关系很难用统一的公式来表示。现有的研究方法一般分为两个阶段：1）对可见光图像进行分割；2）建立不同物体可见光图像和红外的灰度映射关系，再由分割好的可见光图像仿真出红外图像。周强等[5]采用阈值法进行图像分割，结合地物目标的反射率建立灰度映射关系，从而得到仿真图像。李敏等[6]通过脉冲耦合神经元网络实现图像分割，人工标定材质后通过辐射计算得到仿真结果。这类方法也存在模型泛化能力差，仿真结果图精度较低等问题。

从可见光图像仿真生成对应红外图像，如果看作是从可见光图像域到红外图像域的映射，那么属于图像域转换任务的一种。近年来，深度学习方法在图像域转换任务中取得了较好的研究成果，常见如单目深度估计、图像风格转换等。文献[7]最早通过一个端到端的卷积神经网络（CNN）预测深度图，后来提出引入注意力机制、连续条件随机场等概念提升算法性能[8-10]，以及通过优化网络结构优化结果[11-13]，此外还有以多任务[14-17]的学习方式获取辅助信息等。

与图像深度估计任务不同，从可见光转换到红外不仅需要较好的客观评价结果，还需要较佳的视觉效果。受目标函数的约束，图像深度估计任务中的 CNN 方法虽然在客观评价指标获得了较好的结果，但是输出的图像相对比较模糊，丢失了许多纹理细节。而在深度学习领域中，从生成对抗网络（GAN）衍生出的条件生成对抗网络（CGAN）[18]在图像域转换任务[18-23]中表现优异，输出图像的视觉效果较好。

综上可知，由可见光图像仿真红外图像具有重要的研究价值和意义，使用深度学习的方法来实现从可见光到红外图像的仿真具有可行性。传统三维建模的红外仿真算法存在建模复杂、模型泛化能力差等缺点；从可见光图像仿真红外图像的方式，虽然具有时间和人力成本更低、转换效率更高的优点，但实现起来难度较大，并且这一领域里基于深度学习的算法研究较少，因此本文提出了一种生成对抗网络算法来解决这一难题。本文的主要贡献分为以下三个方面。

1）提出了级联多尺度信息融合生成对抗网络（CMIF-GAN），实现了由可见光图像端到端地生成红外仿真图像；并通过实验证明了 GAN 较 CNN 更适用于从可见光到红外图像的转换任务。

2）提出了辅助任务与级联结构相结合的网络框架。首先，CMIF-GAN 采用“由粗到细”的两级网络串联，第一级网络使用大感受野[24]生成网络，重建出红外图像的结构信息，第二级采用小感受野生成网络，补充红外图像的细节信息。其次，在第一级网络中增加语义分割的辅助任务，以得到更准确的宏观结构信息；在第二级网络中增加红外到可见光灰度反转（GIV）图像的辅助任务，以补充红外图像的细节。最后，提出多尺度融合模块（MFM），应用于第二级生成网络，来融合不同感受野下的多尺度信息，提升整体网络性能。

3）在公开数据集 Multispectral Pedestrian Dataset（MPD）[25]上进行了详尽的实验，本文网络模型在多种客观评估指标上具有更好的实验结果。

2　级联多尺度信息融合对抗网络

为了实现由可见光图像仿真生成对应的红外图像，本文提出了一种级联的 GAN 结构，能够由可见光图像端到端地预测红外图像。网络结构如图 1 所示，整体网络结构由两级生成对抗网络串联组成，蓝色部分为第一级，红色部分为第二级。网络输入为可见光图像，辅助信息为语义分割图像和 GIV 图像。

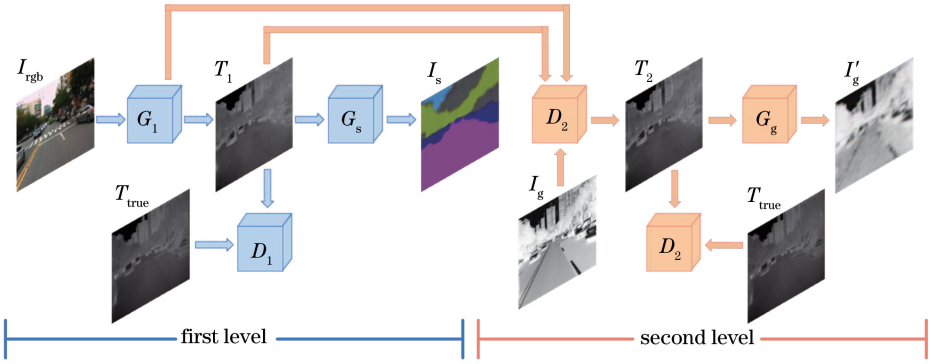

图 1　级联多尺度信息融合对抗网络

Fig. 1　Proposed network of CMIF-GAN

2.1　第一级网络——重建结构

第一级网络以可见光的语义分割图像为参考，从可见光图像中重建出红外图像的基本结构信息。如图 1 所示，第一级网络由一个生成对抗网络和一个辅助任务网络组成，包含两个生成网络 G_1、G_s 和一个判别网络 D_1。可见光图像 I_{rgb} 输入 G_1 中，得到第一级的红外仿真图像 T_1，通过 D_1 判别 T_1 和目标红外图像 T_{true} 的真伪。再将 T_1 作为 G_s 的输入，用来预测语义分割图像。

1）生成网络 G_1、G_s

G_1、G_s 都采用 U-Net[20] 的网络结构，但为了节约计算资源，G_s 减少了卷积的滤波器数。具体网络结构及参数设置如表 1 所示，第一列为网络编码端的网络层和卷积的滤器数，第二列代表经过这一网络层后的输出特征图通道数，三、四列代表解码端。编码端网络层输出跳连拼接到相同行的解码端网络层输出中。网络中所有卷积和反卷积的卷积核都为 4×4，步长都为 2。G_1、G_s 网络层数较深，整体的感受野较大，能够获取图像的整体结构信息。同时，语义分割图像能够体现图像的结构信息，G_s 根据 T_1 预测语义分割图像对 G_1 网络的学习起到了一定引导和限制，使 G_1 更关注结构信息。

表 1　G_1、G_s 网络结构参数配置表

Table 1　Detailed configuration about G_1 and G_s

Encoder (G_1/G_s filters)	Number of channels G_1/G_s	Decoder (G_1/G_s filters)	Number of channels G_1/G_s
		dconv16 (3/3)	3/3
conv1 (64/4)	64/4	dconv15 (64/4)	128/8
conv2 (128/8)	128/8	dconv14 (128/8)	256/16
conv3 (256/16)	256/16	dconv13 (256/16)	512/32
conv4 (512/32)	512/32	dconv12 (512/32)	1024/64
conv5 (512/32)	512/32	dconv11 (512/32)	1024/64
conv6 (512/32)	512/32	dconv10 (512/32)	1024/64
conv7 (512/32)	512/32	dconv9 (512/32)	1024/64
conv8 (512/32)	512/32		

2）判别网络 D_1

第一级的判别网络 D_1 采用文献[20]的判别网络结构，具体网络结构及参数设置如表 2 所示。其中第四列的 Y 代表卷积层后有批量归一化（BN）[26]层，N 代表无。表中 L 代表激活函数泄漏整流线性单元（LReLU），S 代表 Sigmoid 激活函数。

表 2　判别网络结构参数配置表

Table 2　Detailed configuration about discriminator

Network	Input/output channels	Stride	BN	Activation function
conv1	6/64	2	N	L
conv2	64128	2	Y	L
conv3	128/256	2	Y	L
conv4	256/512	2	Y	L
conv5	512/512	1	Y	L
conv6	512/1	1	N	S

2.2　第二级网络——补充细节

第二级网络以GIV图像为辅助信息，在第一级

网络输出红外图像的基础上，丰富红外仿真图像的细节信息。如图 1 所示，第二级网络也由一个对抗网络和一个辅助任务网络组成，与第一级类似。其中，生成网络 G_2 的输入由三个数据拼接组成，包括 G_1 输出的红外图像 T_1、G_1 最后一层的特征图、GIV 图像 I_g。而网络 G_g 是一个辅助任务，通过 G_2 输出的红外图像 T_2，预测 GIV 图像 $I_{g'}$。判别网络 D_2 用来判别 T_2 和 T_{true} 的真伪，结构和 D_1 相同，并且在训练时 D_1、D_2 不共享参数。

在大部分光照良好的条件下，可见光图像的纹理细节信息要多于红外图像。相较于可见光图像，GIV 图像的细节纹理信息和红外图像的细节纹理信息更接近。因此，本文将 GIV 图像作为辅助信息输入到 G_2 中，为网络提供更多的图像纹理细节信息，并通过辅助任务网络 G_g，来引导 G_2 更加关注图像中的细节信息。

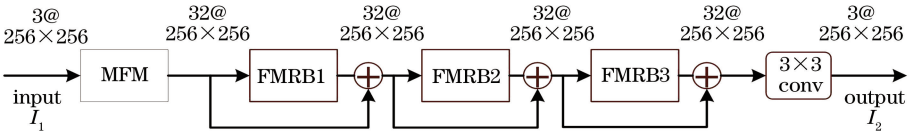

图 2　G_2 网络结构

Fig. 2　Proposed network of G_2

1）生成网络 G_2

生成网络 G_2 的具体结构如图 2 所示。G_2 的输入是由三种数据拼接而成，因此首先使用多尺度融合模块（MFM）将数据整合，得到通道数为 32 的特征图；然后再使用三个快速多尺度残差块（FMRB）[27]将整合数据的纹理细节信息传递到输出端。FMRB 是图像去模糊任务中的模块，具有局部多尺度结构，由两路多重 3×3 卷积拼接组成，能够获得多尺度感受野信息。其在图像去模糊任务中能够较好地学习到细节纹理信息。此外，为了实现不同模块之间的知识共享，G_2 在每个 FMRB 模块之间加入跳跃连接。

MFM 具体结构如图 3 所示。为了获取多感受野的信息，网络输入首先经过四个卷积核大小为 3×3 的空洞卷积，膨胀率分别为 1、2、3、4；然后将不同膨胀率的空洞卷积输出进行相加，从而融合不同感受野的信息，这一步也能够减轻空洞卷积带来的网格效果；接下来将相加后的结果拼接在一起；最后，输入经过一个 1×1 标准卷积与拼接结果相加得到最终输出。

2）轻量小感受野网络 G_g

为了更好的引导 G_2 学习细节纹理信息，得到细节纹理丰富的最终预测红外图像 T_2，G_g 只需要关注 T_2 的细节部分。因此，G_g 的感受野应该较小。G_g 的网络结构如图 4 所示，上半部分是 G_g 的整体网络结构，包含四个模块，中间数字代表通道数；下半部分代表每个模块的具体网络结构，包含三个级联卷积，数字为卷积核大小，卷积步长都为 1，每个卷积后都有一层 LReLu。这种网络结构使得 G_g 的整体网络感受野大小仅为 3×3，且参数量较小。

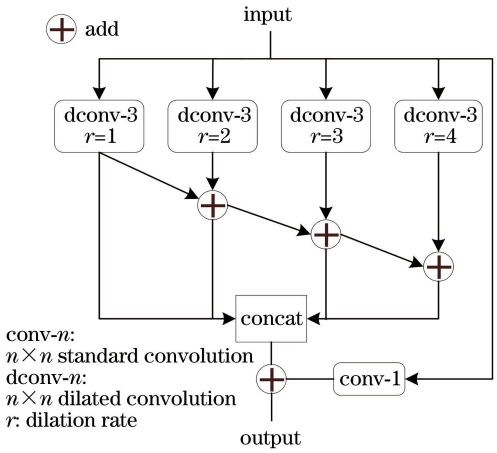

图 3　多尺度融合模块

Fig. 3　Multi-scale fusion module

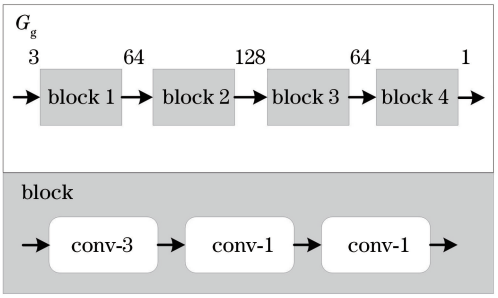

图 4　G_g 网络结构

Fig. 4　Proposed network of G_g

2.3　损失函数

本文的生成网络 G_1、G_s、G_2、G_g 和判别网络 D_1、D_2 以端到端的方式共同训练。判别网络和生成网络的梯度下降交替进行，即 D_1、D_2 的参数先固定，训练 G_1、G_s、G_2、G_g；然后 G_1、G_s、G_2、G_g 的参数固定，训练 D_1、D_2。整体的损失函数 L_{total} 采用最小-最大的训练策略，表达式为

$$\min_{\{G_1,G_s,G_2,G_g\}}\max_{\{D_1,D_2\}}L_{total}=L_{GAN}+L_{pixel},\quad(1)$$

式中：L_{GAN} 为对抗损失函数总和；L_{pixel} 为像素级损失函数总和。L_{GAN} 包含第一级对抗损失 L_{GAN1} 和第二级对抗损失 L_{GAN2}，表达式为

$$L_{GAN}=L_{GAN1}+10\times L_{GAN2}。\quad(2)$$

第一级判别网络 D_1 用于区分合成图像对 $[I_{rgb},T_1]$ 和真实图像对 $[I_{rgb},T_{true}]$，损失函数采用交叉熵的组合形式，表示为

$$L_{GAN1}=E_{I_{rgb},T_{true}}[\ln D(I_{rgb},T_{true})]+E_{I_{rgb},T_1}\{\ln[1-D(I_{rgb},T_1)]\}。\quad(3)$$

第一级判别网络 D_1 用于区分合成图像对 $[I_{rgb},T_2]$ 和真实图像对 $[I_{rgb},T_{true}]$，表示为

$$L_{GAN2}=E_{I_{rgb},T_{true}}[\ln D(I_{rgb},T_{true})]+E_{I_{rgb},T_2}\{\ln[1-D(I_{rgb},T_2)]\}。\quad(4)$$

像素级损失函数总和 L_{pixel} 包含第一级生成网络 G_1、G_s 的 L_1 损失函数 L_{G1}、L_{Gs}，第二级生成网络 G_2、G_g 的 L_1 损失函数 L_{G2}、L_{Gg} 和对纹理更敏感的梯度损失函数 L_{g_G2}，表达式为

$$L_{pixel}=\lambda_1L_{G1}+\lambda_2L_{Gs}+\lambda_3L_{G2}+\lambda_4L_{Gg}+\lambda_5L_{g\_G2},\quad(5)$$

式中：λ 是超参数，代表各个损失函数的权重；G_1、G_2 负责生成红外图像，是目标任务网络，权重最高；网络 G_s、G_g 负责辅助任务，权重较低；梯度损失函数用于增加网络对边缘的感知能力，权重最小。经过多次实验，本文最终将 λ 从 1 到 5 分别设置为 100、5、200、10、0.5。L_1 损失函数代表平均绝对误差，表示为

$$L_1=\frac{1}{N}\sum_{i=1}^{N}|y_i-y_i^*|,\quad(6)$$

式中：i 是像素索引；N 是一幅图像里所有像素总和的数目；y_i、y_i^* 分别代表像素 i 处的真实灰度值和网络预测的灰度值。梯度损失函数 L_g 表达式为

$$L_g=\frac{1}{2N}\sum_{i=1}^{2N}(|\nabla_h y_i-\nabla_h\hat{y}_i|+|\nabla_v y_i-\nabla_v\hat{y}_i|),\quad(7)$$

式中：$\nabla_h\hat{y}_i$、$\nabla_h y_i$ 分别代表目标红外图像 T_{true} 像素 i 处水平方向的梯度值和红外仿真图像像素 i 处水平方向的梯度值；$\nabla_v y_i$、$\nabla_v\hat{y}_i$ 代表垂直方向。

3　实验细节与评估指标

3.1　数据预处理

本文使用公开数据集 MPD 对所提出的网络模型进行训练和测试。MPD 由 Hwang 等[25]提出并制作，内容为配准的可见光图像和对应红外图像的图像对，分辨率为 640×512。其中，训练集和测试集分别含有 50187 和 45141 个图像对。训练集和测试集都包含校园、街道和城郊三个场景，每个场景又分别包含白天和夜晚的拍摄图像。本文选取了三个场景中白天的图像对作为网络的训练集，训练集大小为 33399 个图像对。相应地，本文在 MPD 测试集的白天图像对中随机抽取 565 个图像对作为网络的测试集。送入网络之前，本文通过双线性插值下采样将图像分辨率大小调整至 256×256。

可见光的语义分割图像和灰度反转图像是本文网络的辅助信息。GIV 图像通过将可见光图像由彩色图像转为灰度图像，再进行灰度值反转操作得到。将可见光图像输入到 Refinenet[28] 在 Cityscapes 上训练好的模型中，可以预测得出语义分割图像。Cityscapes 是主要应用于语义分割的大型数据集，主要场景为室外街道，和 MPD 场景类似。

3.2　评估指标

在图像域转换任务以往的工作中，有一些公认的评估指标来评价网络预测图像和真实目标图像的相似度。本文采用平均相对误差(Rel)、对数平均误差(Log10)、均方根误差(RMS)以及准确率($\delta<1.25^i$，$i=1,2,3$)。各指标的计算表达式分别为

$$R_{rel}=\frac{1}{|N|}\sum_{i=1}^{N}|y_i-y_i^*|/y_i^*,\quad(8)$$

$$R_{log10}=\frac{1}{|N|}\sum_{i=1}^{N}|\lg y_i-\lg y_i^*|,\quad(9)$$

$$R_{rms}=\sqrt{\frac{1}{|N|}\sum_{i=1}^{N}\|y_i-y_i^*\|^2},\quad(10)$$

$$\delta=\max\left(\frac{y_i}{y_i^*},\frac{y_i^*}{y_i}\right)<T_{th},\quad(11)$$

式中：i 是像素索引；N 是一幅红外图中像素总和的数目；y_i 和 y_i^* 分别代表像素 i 处的目标图像灰度值和预测图像灰度值。此外，本文也采用峰值信噪比(PSNR)和结构相似性(SSIM)，作为图像去模糊、超分辨等的评估指标，能够较好地反映两幅图像的相似度。

3.3　实验设置细节

本文使用 Pytorch 框架，在内存为 16 GB，GPU 为 NVIDIA Titan XP 的计算机上进行实验。网络采用均值为 0、标准差为 0.2 的高斯分布进行权重初始化，使用 Adam 作为优化器，设置动量为 0.5。设置初始学习率为 0.0001，batch size 为 4。完整训

练过程需要大约 16 h，训练集数据一共迭代了 20 次。

4　实验结果与分析

4.1　与先进算法的对比

本节与图像域转换的先进算法进行了对比实验，客观参数指标如表 3 所示。前两种网络是 CNN，由可见光直接生成红外仿真图像；后三种网络为 GAN，即生成对抗网络结构。先进算法生成的红外仿真图像如图 5 所示，其中第一行是可见光图像，最后一行是目标红外图像。

1）与 CNN 先进算法的对比

从实验结果看，客观指标上 CNN 算法优于 GAN，主观感受上 GAN 算法优于 CNN。表 3 中前两个网络，FCRN[11]、FLEDNet[12] 是实现单目深度估计的端到端 CNN 网络，均不需要辅助信息。在 6 个评价参数中，FLEDNet[12] 有 4 个最优，本文 CMIF-GAN 有 2 个最优，落后于前者。

表 3　算法的客观指标对比

Table 3　Comparison of objective indicators of algorithms

Method	The lower, the better			The higher, the better		
	Rel	Avg log10	RMS	$\delta<1.25$	PSNR	SSIM
FCRN[11]	0.286	0.144	1.060	0.409	21.204	0.962
FLED-Net[12]	**0.238**	**0.100**	**0.853**	0.602	**22.921**	0.987
Pix2pix[20]	0.248	0.107	0.906	0.571	22.431	0.985
Selection-GAN[23]	0.284	0.112	0.958	0.554	21.976	0.982
Proposed	0.257	0.102	0.876	**0.612**	22.657	**0.989**

但从图 5 的仿真图像对比中可以看出，FLEDNet[12] 的红外仿真结果虽然在客观指标上优于本文 CMIF-GAN 结果，但普遍存在图像模糊、纹理丢失的现象。例如图 5(a)中框选的汽车非常模糊，车牌保险杠的形状无法分别；图 5(b)中骑车人退化成一团光晕；图 5(c)中楼宇规则的窗户形状也退化的无法分辨。同时也可以发现，GAN 算法的视觉效果优于 CNN 算法，图像中的边界结构基本正确，并且建筑物、车辆、人物的细节较丰富，更符合人眼视觉感受。

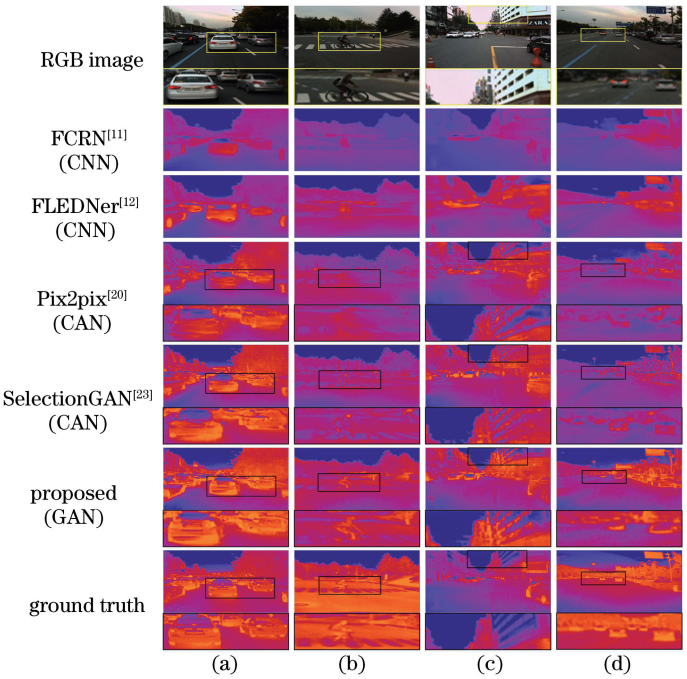

图 5　不同算法生成的红外仿真图。(a)汽车；(b)自行车骑手；(c)楼房；(d)远距离车辆

Fig. 5　Infrared simulation images generated by different algorithms. (a) Car; (b) bicycle rider; (c) building; (d) long-distance vehicle

2）与 GAN 先进算法的对比

客观指标上看，本文 CMIF-GAN 优于其他两种 GAN 算法，6 个指标中 5 个最优；主观效果上，本文的仿真图像中场景与物体的重构更准确，细节也

更丰富，与实际红外图像更接近。表 3 中，Pix2pix[20]、SelectionGAN[23]都是端到端地实现图像域转换的 GAN。Pix2pix[20]是较早提出的一个生成对抗网络，用于图像风格转换任务；2019 年 CVPR 提出的 SelectionGAN[23]包含两个阶段的生成对抗网络，用于不同视角下的图像转换，本文在其级联的架构上进行了改进。

表 3 结果表明，CMIF-GAN 在大部分指标上都达到了最佳性能：在误差 RMS 上优于 SelectionGAN[23] 8.6%；在准确率 $\delta<1.25$ 上提高了 10.5%。主观感受上，如图 5 所示：图 5(a)中框选的汽车部分，Pix2pix[20]图像最差，车体模糊，SelectionGAN[23]图像基本能看清车尾结构，本文方法结果最清晰，能看清楚车牌轮廓、车尾灯；图 5(b)中框选的自行车骑手，Pix2pix[20]结果无法显示车与人的轮廓，SelectionGAN[23]图像中能看清部分的轮廓，而本文方法结果基本能显示出完整人体轮廓，与目标图像最为接近；从图 5(c)中的楼房和图 5(d)中远距离拍摄的车辆能够看出，本文算法的结果与目标红外图像相似度更高，结构纹理更清晰、更丰富、更准确。

综上，虽然 GAN 算法较 CNN 算法在客观指标上略落后，但其图像的边缘更清晰、结构更准确、细节更丰富，更适用于本文任务。在三种 GAN 算法中，本文算法的客观指标与主观感受最优。

4.2　两级网络结构的对比

本文 CMIF-GAN 包含两级生成对抗网络，第一级重建红外仿真图像的结构信息，第二级丰富细节信息。为了验证两级网络结构的必要性，本节对比了一级网络和两级网络的差异，实验结果如表 4 所示。在误差 RMS 上一级网络比两级网络增大了 4.8%；在准确率 $\delta<1.25$ 上，一级网络比两级网络低 3.8%。

表 4　一级网络与 CMIF-GAN 的对比实验

Table 4　Comparison of first level network and CMIF-GAN

Method	The lower, the better			The higher, the better		
	Rel	Avg log10	RMS	$\delta<1.25$	PSNR	SSIM
First level	0.265	0.107	0.918	0.589	22.310	0.987
Proposed	**0.257**	**0.102**	**0.876**	**0.612**	**22.657**	**0.989**

两种结构的仿真图像如图 6 所示，可以看出，一级网络结果比较粗糙，而两级网络结果细节更准确，与目标图像更相似。例如第一幅图像中框选的道路指示牌，一级网络结果中丢失了部分结构，两级网络的轮廓更加完整；第二幅图像中的框选的部分，包括路标杆、树枝，比较两个结果可以看出，两级网络结果的细节纹理信息相对更准确。综上，说明本文“由粗到细”的级联结构是有效的，即第一级网络重建结构信息，第二级网络补充细节纹理信息。

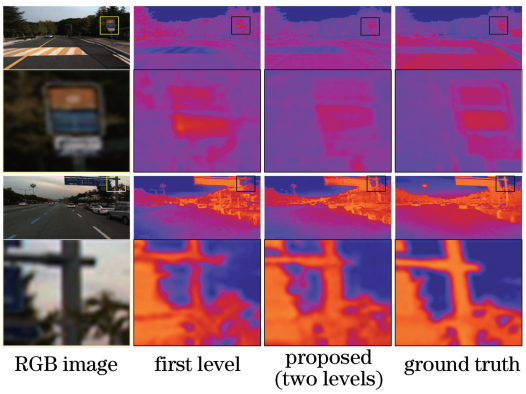

图 6　一级网络与 CMIF-GAN 结果对比图

Fig. 6　Results comparison of first level network and CMIF-GAN

4.3　辅助任务实验对比

本文算法中增加了辅助任务以提升网络性能，本节对辅助任务的作用进行对比。第一级网络的辅助任务为语义分割图像，第二级网络的辅助任务为可见光的灰度反转图像。具体消融实验结果如表 5 所示。结构一，同时去掉语义分割和 GIV 图像这两个辅助任务，即去除 G_s、G_g 网络；结构二，仅去除 GIV 图像，即无 G_g 网络；结构三，仅去除语义分割图像，即无 G_s 网络；第四行代表完整 CMIF-GAN。

从表 5 中可以看出，包含语义分割和 GIV 图像辅助任务的完整 CMIF-GAN 获得最佳实验结果，在准确率 $\delta<1.25$ 上分别高于结构一、结构二、结构三的 4.4%、3.0%、2.3%。结构一无任何辅助任务，指标上性能最差。结构三在各项客观指标上均优于结构二网络，说明 GIV 辅助任务相对语义分割来说，提升作用更大一些。

虽然结构二中语义分割的辅助任务使得网络能够学习到更正确的结构信息，但是客观指标计算过程中并不能为结构信息增加权重。结构三中的 GIV 图像的辅助任务，使网络能够获得更多图像细节信息，即使结构上有些差异，但依然能保证指标更优。这也是客观指标的一种局限性。

表 5　辅助任务实验对比

Table 5　Comparison of auxiliary tasks experiments

Setup	The lower, the better			The higher, the better		
	Rel	Avg log10	RMS	$\delta<1.25$	PSNR	SSIM
$-G_s, G_r$	0.268	0.107	0.912	0.586	22.321	0.988
$-G_g$	0.276	0.106	0.925	0.594	22.280	0.986
$-G_s$	0.264	0.105	0.901	0.598	22.510	0.987
Proposed	**0.257**	**0.102**	**0.876**	**0.612**	**22.657**	**0.989**

4.4　MFM 模块实验分析

本节对 G_2 网络中提出的 MFM 模块的作用进行实验分析。为了提升网络精度，MFM 模块通过不同膨胀率的空洞卷积获取多感受野的信息，并通过相加和拼接操作融合多感受野的信息。MFM 模块的对比实验如表 6 所示：第一行代表 CMIF-GAN 去掉 MFM 模块的实验结果；第二行代表有 MFM 模块，即完整 CMIF-GAN。

从表 6 中可以看出，有 MFM 模块的网络在各个指标上的表现都更好，在误差 Rel 和 RMS 上分别低于无 MFM 模块网络 4.1%和 3.3%，可以验证 MFM 模块有助于提升网络精度，能够较好地学习到有用信息。

表 6　MFM 模块实验对比

Table 6　Comparison of MFM module experiments

Setup	The lower, the better			The higher, the better		
	Rel	Avg log10	RMS	$\delta<1.25$	PSNR	SSIM
−MFM	0.268	0.105	0.905	0.600	22.465	0.987
Proposed	**0.257**	**0.102**	**0.876**	**0.612**	**22.657**	**0.989**

5　结　论

为从可见光图像转换到对应的红外仿真图像，本文提出了包含两级对抗网络的级联多尺度信息融合生成对抗网络，并采用语义分割图像和 GIV 图像作为辅助任务的输入信息。在 MPD 数据集的实验结果表明，相较于其他先进算法，该网络模型在图像域转换任务的多种客观评估指标上均得到较好结果。

参　考　文　献

[1] Mu C P, Peng M S, Dong Q X, et al. Infrared image simulation of ground maneuver target and scene based on OGRE[J]. Applied Mechanics and Materials, 2015, 3752(716): 932-935.

[2] Ma Y, Tian Y. Modeling method of warship radiation model for infrared simulation[J]. Tactical Missile Technology, 2013(3): 67-70,75.
马艳, 田宇. 红外仿真中舰船辐射模型建模方法[J]. 战术导弹技术, 2013(3): 67-70, 75.

[3] Yang M, Li M, Yi Y X, et al. Infrared simulation of ship target on the sea based on OGRE[J]. Laser & Infrared, 2017, 47(1): 53-57.
杨敏, 李敏, 易亚星, 等. 基于 OGRE 的海面舰船目标红外仿真方法[J]. 激光与红外, 2017, 47(1): 53-57.

[4] Xie J R, Li F M, Wei H, et al. Infrared target simulation method based on generative adversarial neural networks[J]. Acta Optica Sinica, 2019, 39(3): 0311002.
谢江荣, 李范鸣, 卫红, 等. 基于生成对抗式神经网络的红外目标仿真方法[J]. 光学学报, 2019, 39(3): 0311002.

[5] Zhou Q, Bai T Z, Liu M Q, et al. Near infrared scene simulation based on visual image[J]. Infrared Technology, 2015, 37(1): 11-15.
周强, 白廷柱, 刘明奇, 等. 基于可见光图像的近红外场景仿真[J]. 红外技术, 2015, 37(1): 11-15.

[6] Li M, Xu Z W, Xie H W, et al. Infrared image generation method and detail modulation based on visible light images[J]. Infrared Technology, 2018, 40(1): 34-38.
李敏, 徐中外, 解鸿文, 等. 基于可见光图像的红外图像生成方法及其细节调制[J]. 红外技术, 2018, 40(1): 34-38.

[7] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[C]// International Conference on Neural Information Processing Systems, 2014: 2366-2374.

[8] Wang P, Shen X H, Lin Z, et al. Towards unified depth and semantic prediction from a single image[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 7-12 June 2015, Boston, MA, USA. New York: IEEE Press, 2015: 2800-2809.

[9] Xu D, Ricci E, Wanli O Y, et al. Multi-scale continuous CRFs as sequential deep networks for

monocular depth estimation [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 21-26 July 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 161-169.

[10] Xu D, Wang W, Tang H, et al. Structured attention guided convolutional neural fields for monocular depth estimation [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 18-23 June 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 3917-3925.

[11] Laina I, Rupprecht C, Belagiannis V, et al. Deeper depth prediction with fully convolutional residual networks[C]//2016 Fourth International Conference on 3D Vision (3DV). 25-28 Oct. 2016, Stanford, CA, USA. New York: IEEE Press, 2016: 239-248.

[12] Jia R M, Liu L Q, Liu S J, et al. Single image depth estimation based on encoder-decoder convolution neural network [J]. Journal of Graphics, 2019, 40(4): 718-724.
贾瑞明, 刘立强, 刘圣杰, 等. 基于编解码卷积神经网络的单张图像深度估计[J]. 图学学报, 2019, 40(4): 718-724.

[13] Jia R M, Li Y, Li T, et al. Monocular image depth estimation network with multi-level feature fusion structure [J/OL]. Computer Engineering [2020-04-01]. https: //doi. org/10. 19678/j. issn. 1000-3428. 0056477.
贾瑞明, 李阳, 李彤, 等. 多层级特征融合结构的单目图像深度估计网络[J/OL]. 计算机工程[2020-04-01]. https: //doi. org/10. 19678/j. issn. 1000-3428. 0056477.

[14] Qi X J, Liao R J, Liu Z Z, et al. GeoNet: geometric neural network for joint depth and surface normal estimation [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 18-23 June 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 283-291.

[15] Yin Z C, Shi J P. GeoNet: unsupervised learning of dense depth, optical flow and camera pose[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 18-23 June 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 1983-1992.

[16] Ranjan A, Jampani V, Balles L, et al. Competitive collaboration: joint unsupervised learning of depth, camera motion, optical flow and motion segmentation [C] // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 15-20 June 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 12232-12241.

[17] Jiao J B, Cao Y, Song Y B, et al. Look deeper into depth: monocular depth estimation with semantic booster and attention-driven loss [J]. Computer Vision-ECCV 2018, 2018: 53-69.

[18] Mirza M, Osindero S. Conditional generative adversarial nets [EB/OL]. (2014-11-06) [2020-04-01]. https: //arxiv. org/abs/1411. 1784.

[19] Hu L M, Zhang Y. Facial image translation in short-wavelength infrared and visible light based on generative adversarial network [J]. Acta Optica Sinica, 2020, 40(5): 0510001.
胡麟苗, 张湧. 基于生成对抗网络的短波红外-可见光人脸图像翻译[J]. 光学学报, 2020, 40(5): 0510001.

[20] Isola P, Zhu J Y, Zhou T H, et al. Image-to-image translation with conditional adversarial networks [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 21-26 July 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 5967-5976.

[21] Ma S, Fu J L, Chen C W, et al. DA-GAN: instance-level image translation by deep attention generative adversarial networks [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 18-23 June 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 5657-5666.

[22] Mejjati Y A, Richardt C, Tompkin J, et al. Unsupervised attention-guided image to image translation [EB/OL]. (2018-11-08) [2020-04-01]. https: //arxiv. org/abs/1806. 02311.

[23] Tang H, Xu D, Sebe N, et al. Multi-channel attention selection GAN with cascaded semantic guidance for cross-view image translation [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 15-20 June 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 2412-2421.

[24] Luo W J, Li Y J, Urtasun R, et al. Understanding the effective receptive field in deep convolutional neural networks [EB/OL]. (2017-01-25) [2020-04-01]. https: //arxiv. org/abs/1701. 04128.

[25] Hwang S, Park J, Kim N, et al. Multispectral pedestrian detection: Benchmark dataset and baseline [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 7-12 June 2015, Boston, MA, USA. New York: IEEE Press, 2015: 1037-1045.

[26] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift[EB/OL]. (2015-03-02) [2020-04-01]. https: // arxiv. org/abs/1502. 03167.

[27] Jia R M, Qiu Z Z, Cui J L, et al. Deep multi-scale

encoder-decoder convolutional network for blind deblurring[J]. Journal of Computer Applications, 2019, 39(9): 2552-2557.
贾瑞明, 邱桢芝, 崔家礼, 等. 盲去模糊的多尺度编解码深度卷积网络[J]. 计算机应用, 2019, 39(9): 2552-2557.

[28] Lin G S, Milan A, Shen C H, et al. RefineNet: multi-path refinement networks for high-resolution semantic segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 21-26 July 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 5168-5177.